

2020-08

Evidence on social and financial performance: mapping the empirical garden of forking paths

Luca Berchicci, Andrew King. 2020. "Evidence on Social and Financial Performance: Mapping the Empirical Garden of Forking Paths." *Academy of Management Proceedings*, Volume 2020, Issue 1, pp. 17546 - 17546. <https://doi.org/10.5465/ambpp.2020.17546abstract>

<https://hdl.handle.net/2144/42419>

"Downloaded from OpenBU. Boston University's institutional repository."

EVIDENCE ON SOCIAL AND FINANCIAL PERFORMANCE: MAPPING THE EMPIRICAL GARDEN OF FORKING PATHS

Abstract: Worldwide, almost 30% of professionally managed assets are invested in funds that use social performance as a screening or selection criteria. Scholars have encouraged such investment by contending that social and financial gain are linked, but reviews of empirical research on the connection between social and financial performance have been inconclusive. In fact, six of the most influential articles on the subject reach conflicting conclusions despite using the same sources of data and appearing in the same peer-reviewed journal. Some scholars opine that no synthesis of these disparate findings is feasible, but we use new ideas from epistemology and statistics to show how it can be done. We conclude that the interpretation of the evidence depends on empirical assumptions, particularly about the location of meaningful variance: differences between firms imply a positive relationship; but year-to-year differences within firm histories imply a negative association between social and financial performance.

EVIDENCE ON SOCIAL AND FINANCIAL PERFORMANCE: MAPPING THE EMPIRICAL GARDEN OF FORKING PATHS

World-wide, more than \$30 trillion, or almost 30% of professionally managed assets, are invested in funds that use sustainability, responsibility, or social impact as a screening or selection criteria (GSIA, 2019). Some of these investors are surely motivated by altruism, but others hope that firms that do good will also do well for their portfolios. Scholars and professional managers have encouraged such beliefs by contending that social and financial gain are linked, but reviews of empirical research on the connection between social and financial performance have been inconclusive. Margolis and Walsh (2001) claim that most studies find a small positive relationship between social and financial performance, but they acknowledge that there is also a wide dispersion in results. Orlitzky, Schmidt, and Rynes (2003) also infer a generally positive connection, but contend that it may be contingent on the use of particular measures of financial performance. Most recently, Michael Porter, Mark Kramer, and George Serafeim (2019) have argued that they believe no empirical study has found credible evidence of a relationship between social performance and “alpha” – a common measure of abnormal stock return. Given that Porter, Kramer, and Serafeim have usually been enthusiastic champions of the potential for such a relationship, their skepticism carries particular weight.

Many studies of the relationship between social performance (SP) and financial performance (FP) have used the same sources of data: social measures from research firm Kinder, Lydenberg and Domini (KLD) and financial data from Standards and Poor’s. The most influential article of this group, Waddock & Graves (1997), was among the first to report results from a systematic quantitative study of the SP-FP relationship. Using KLD and S&P data from 1989-1991, they estimate a positive and significant relationship between social and financial performance. After their analysis, the *Strategic Management Journal* published five closely

related studies, but each began with different assumptions and ended with different conclusions (see Appendix 1). In 2000, McWilliams and Siegel re-estimated the SP-FP relationship using a different specification, this time including Research and Development (R&D) expenditures as a control variable. They found no association between social and financial performance. In 2001, Hillman and Keim argued that KLD measures of social performance should be separated into two scales, one measuring social management and the other issue participation. They found a positive association for the former, but a negative one for the latter. Hull & Rothenberg (2008) extended previous research by arguing that the SP-FP relationship should be assumed to be moderated, not just mediated, by R&D expenditures. Based on their new model specification, they conclude that social performance is linked to financial performance only for low R&D firms (Hull and Rothenberg, 2008). Barnett & Salomon (2012) argue that previous analyses had mistakenly assumed a linear association, when the real relationship might be curvilinear. Based on their empirical model, they conclude that social performance is only financially beneficial for low and high performers. Finally, Zhao and Murrell (2016) return to the original model of Waddock & Graves (1997) and retest it with updated and more expansive data. They find an association between social performance and accounting measures of financial performance, but not with those based on market value.

Together, these six articles have been cited more than ten thousand times, and their impact continues to grow. Some scholars interpret them as providing definitive evidence with respect to a particular relationship, while others conclude that their mixed findings mean that no general interpretation is warranted (Aguinis and Glavas, 2012). Some even despair of the possibility of ever reaching consensus on the proposed SP-FP link. Writing more broadly about the empirical literature on social and financial performance, Rowley and Berman (2000: 401) argue that “CSP-FP research represents an attempt to legitimize the researcher and the business

and society field, rather than build understanding.” They contend that it is “difficult to produce worthwhile comparisons across studies or generaliz[e] beyond the boundaries of a specific study (p. 397)”; they propose jettisoning the whole enterprise of large-scale quantitative analysis; and they instead encourage researchers to conduct detailed studies of possible specific connections (Rowley and Berman, 2000). They do not indicate, however, how readers should form aggregated knowledge from such fragmented studies.

In this article, we explore a new way to evaluate and combine research on the social-financial performance link. We draw inspiration for our analysis from recent research in economics, political science and management (Simonsohn, Simmons et al., 2015, Durlauf, Navarro et al., 2016, King, Goldfarb et al., 2019). This research highlights the importance of empirical assumptions in determining estimated effects, and it proposes various ways that the influence of these assumptions can be made more transparent to readers. We demonstrate the usefulness of this approach by mapping the space of assumptions used in six influential articles in the *Strategic Management Journal*.

EPISTEMOLOGY OF AGGREGATING EVIDENCE

Most of what we think we know about business management derives from what we hear or read about the research of other scholars. But how do we aggregate knowledge across different studies? In the case of the link between social and financial performance, most previous research has relied on principles of meta-analysis. Studies on a similar topic are collected and their reported coefficient estimates are combined to create an overall score. This approach has a long and reputable history, but it also involves inherent weaknesses. Broad analyses, such as Margolis & Walsh (2001), must lump together studies using different measures, scales, functional forms, control variables, samples, and so on. The method is also vulnerable to

publication bias, because only certain findings may be selected by authors (or reviewers) for publication (Hunter and Schmidt, 2004). Finally, meta-analyses may aggregate together studies that have been done correctly with those containing a critical flaw (Hunter and Schmidt, 2004). All of these points are particularly relevant for research on the social-financial performance link, where Waddock (2004: 5) admits that “parallel and sometimes confusing universes exist.”

The six articles we consider in this study do not come from different universes. They use the same data, similar measures, and are published in the same peer-reviewed journal. Yet, they still result in “findings” that are hard to combine into a useful whole. To learn from them, a reader must know that the estimates can be trusted and are sincere (Fricker, 2002; Wilholt, 2013). “Trust” means that the reader must think the author made the same empirical assumptions that the reader would have done had she conducted the study (Fricker, 2002; Wilholt, 2013). “Sincere” means that the reported frequentist estimates¹ would occur as predicted in repeated samples if the reported assumptions were used to guide the empirical analysis (Fricker, 2002; Hacking, 1965).

Unfortunately, readers can rarely observe the assumptions that authors have made, and thus do not know how to use reported analysis (Longino, 1990). Consider for example, a case where an author chooses to eliminate an extreme outlier from a data set. The reader is unlikely even to know of this exclusion and less likely to be able to assess its justification. Thus, the reader must trust that the author made the right choice. Empirical research requires dozens or hundreds of such choices. So many, in fact, that statisticians Andrew Gelman and Erick Loken liken the empirical process to a walk thorough a “garden of forking paths” (Gelman and Loken, 2013). Depending on the assumption/choice made at each fork, the researcher will exit the

¹ These are the most common estimates used in social science. Developed by Fisher and Neyman & Pearson, they allow a prediction of how frequently an estimate will appear in identically created samples taken from the same population (Schneider, 2015). Traditional “significance tests” using p-values or T-tests are common examples.

“garden” at a different spot. To know the implications of the final result, the reader must observe and agree with each choice, or trust presumptively that the author made the same choices the reader would have done (Fricker, 2002; Wilholt, 2013).

The reader’s difficulty in assessing the validity of empirical findings is made more problematic by the need also to observe the timing of empirical choices. To properly conduct frequentist analysis, researchers must specify in advance their hypotheses, sampling plan, test procedure, inference rules and so on (Spanos, 2010). If a hypothesis is formed after evidence is found, a process known as HARKing, then the test statistics do not accurately predict what estimates will be found in repeated samples (Simmons, Nelson et al., 2011). This problem of the timing of selected assumptions is further complicated by the peer review process. If journal reviewers select findings based on their outcomes, they may tend to select those results that are more counter-intuitive, and thereby implicitly endorse the use of inappropriate empirical assumptions (De Long and Lang, 1992).²

Given uncertainty about the accuracy and robustness of coefficient estimates, how then can we readers make sense of the empirical evidence published in academic journals? In recent years, many researchers have proposed that extensive replication of empirical findings could help clarify which studies provide “informative value” and which do not (Simonsohn, Nelson et al., 2014, Bettis, Helfat et al., 2016). Some have even argued that readers should defer making inferences from reported estimates until they have been confirmed by a large body of related analysis (Simmons, Nelson et al., 2011). Unfortunately, this may entail a long wait. Building up sufficient studies to allow a p-curve analysis would mean the publication of several exact

² It is possible that this phenomenon explains why the publication of Waddock & Graves’ (1997) positive relationship between SP & FP was quickly followed by McWilliams & Siegel’s (2000) calculation of a neutral one.

replications from identical samples from the same population. Under current publication norms, such replications are rarely conducted or published (Bettis, Ethiraj et al., 2016).³

This problem of knowledge validation and aggregation is common to many areas of science, and scholars from several disciplines have begun to converge on a common proposal. They eschew individual replications or meta-analyses of existing results. They instead suggest that scholars determine the set of assumptions used by a collection of studies and then map the connection between each element of this set and its associated estimate. In economics, statistician Edward Leamer has long advocated creating large maps of possible relationships between assumptions and findings – a process he calls “extreme bounds analysis” (Leamer, 1985). In finance, Sal-i-Martin (1997) follows Leamer’s advice to investigate conflicting findings in macro-economics. In political science, Durlauf, Navarro, and Rivers (2016) evaluate the literature on the effect of concealed-weapon permits on crime. They determine the full set of assumptions used from previous studies and then estimate outcomes for all possible combinations of these assumptions. In psychology, Simonsohn, Nelson, and Simmons (2015) suggest that scholars should estimate and report descriptive and inferential statistics on all reasonable specifications. In management, King, Goldfarb and Simcoe (2019) argue that scholars should report maps connecting feasible assumptions and estimation outcomes. All of these scholars argue that maps of the connections between assumptions and estimates will aid readers in verifying and aggregating reported analyses.

In this article, we bring these new methods to the question of a link between social and financial performance. We know of no precedence for our study with respect to this connection.

³ A notable exception is one of the papers in our set of six: Zhao and Murrell (2016).

The closest precedent for the method of our analysis is Durlauf, Navarro, and Rivers (2016). We augment their approach with mapping suggestions from King, Goldfarb and Simcoe (2019).

Following precedent in the literature, we begin our analysis by identifying the space of assumptions implied by previous research. We do not prioritize any particular set of assumptions or model specifications, but instead try to develop a feasible set of alternatives for evaluation.

We then estimate all models based on these assumptions and report the results in two main ways.

First, to aid inference by readers with strong priors⁴ about the correct models, we report “epistemic maps” connecting the full range of assumptions to their related estimates (see Figure 2). We also report maps where we hold constant a critical empirical assumption while allowing others to vary (see Figures 2-5). Such maps allow readers to index from their assumptions to conditional estimations and to observe the reliability of these estimates. For example, if a particular scholar believes that firm-level differences more accurately reflect the relationship between social and financial performance, she can constrain her consideration to the collection of models that include firm-level fixed effects. She can then observe whether estimates from her preferred class of models vary with changes in other assumptions (i.e. controls, interactions, etc.).

Second, to allow inference by readers with diffuse priors⁵ across different assumptions, we also use Bayesian analysis to calculate posterior probabilities⁶ for the entire set of models. This allows us to construct a probability-weighted average effect based on groups of models. It also allows us to select, conditional on the assumption of diffuse priors, “best” models for inference.

The set of assumptions used in previous studies

⁴ Strong Priors: one’s defined beliefs on the ‘right’ set of assumptions prior to observing evidence.

⁵ Diffuse Priors: equal beliefs prior to observing evidence. In our case, this means that a person thinks all empirical assumptions in our set (and the resulting empirical models) are equally likely to be the “right” ones before they observe the evidence.

⁶ The conditional probability of something after observing the evidence.

To create our map of the connection between empirical assumptions and estimations, we reviewed the six highly-influential studies identified in Appendix 1. We determine the range of assumptions used or implied. Doing so allows us to consider the main areas of debate in the literature. For example, should the relationship between social and financial performance be specified as linear or quadratic? Is the effect moderated or mediated by other variables? Should firms be compared to each other, or should only within-firm variance be used in estimates? What control variables should be included? In total, we identified six main classes of assumptions used in these influential articles. These influence measurement of outcome variables, measurement of predictor variables, functional relationships, level of informative variance, appropriate controls, and sample time-period (Figure 1).

Assumptions about measurement of outcome variables

Scholars differ on how financial performance should be measured. Waddock and Graves (1997) choose to use accounting estimates of financial performance (e.g. return on assets), while other scholars chose to employ measures based on market value (Market Value Added, Market to Book, or Tobin's Q). This choice is based on assumptions about the nature of the relationship between social and financial performance. For example, Hillman and Keim (2001) choose to use a measure of "market value added" because they expect social performance to influence long-term firm value and intangible assets. Other scholars, such as Barnett and Salomon (2012) choose to use accounting measures because they believed these are more informative. Several scholars choose to evaluate both market and accounting measures.

Assumptions about construction of predictor variables

Scholars also differ on how independent variables should be measured. All use data from Kinder, Lydenberg, and Domini, but they aggregate these scores into scales in different ways. Waddock and Graves (1997) argue that some KLD measures are more important than others and

weight them accordingly. Hull & Rothenberg (2008) assume that social issues with more subcategories are more important and weight these higher. Hillman and Keim (2001) argue that a subset of measures should be combined to form an estimate of “social management” or “social issue participation”. Barnett & Salomon (2012) choose to use unitary weights for all of the KLD measures. McWilliams and Siegel (2000) use a dummy variable capturing those firms with scores sufficiently high to allow them to appear, based on their KLD scores, in the Domini Social Index.

Assumptions about functional relationships

Scholars also differ in their assumptions about the relationship between social and financial performance. Most assume that the relationship is linear, but Barnett and Salomon (2012) disagree. Based on theories of stakeholder influence capacity, they hypothesize that the association will be curvilinear, and they specify this in a model with a quadratic form. Hull and Rothenberg (2008) assume that the relationship is moderated by firm and industry attributes, which they assume are exogenously given. Thus, they specify a model that includes interaction terms for both the Firm’s R&D intensity and the Industry’s Advertising Intensity.

Assumptions about the level of informative variance

Scholars also disagree about whether the relationship between social and financial performance can best be estimated by comparison across firms or by comparison of performance within a firm over time. If firms are actively changing over the panel, one might expect that “within-firm variance” would allow accurate estimates. However, if firms have largely reached equilibrium states, variance between firms may be more informative. At issue is whether each firm should be allowed a separate and independent base level of performance, or a “fixed effect”.

Assumptions about appropriate control variables

The authors also use different assumptions about the need for control variables in their analysis. Since Waddock & Graves (1997), most scholars have assumed that some firm attributes

might influence both predictor and outcome variables. As a result, they have specified models that include a set of control variables that consist of firm size (sales or employees) and risk (debt/assets). McWilliams and Siegel (2000) make an impassioned argument for including both R&D and Advertising Intensity as control variables. They contend that these attributes are known to influence financial performance and also may be associated with social performance. If so, model specifications where they are left out might result in biased estimates of the relationship between social and financial performance.

Sample time period

All of the authors use different samples for their analysis. These reflect assumptions about viable links between KLD and Compustat data as well as data availability at the time when the analysis was conducted. These assumptions are not well documented in the six papers we evaluated, so we chose to substitute our own assumption. We assume that the relationship should be evident in the entire panel of currently available data. In future work, we plan to consider different time periods.

The combined assumptions and implied model space

Figure 1 shows the different assumptions that determined the model space we estimate. It consists of possible combinations of the model elements described above. For simplicity, we choose to remove some closely related choices. Even so, the implied space of plausible models is quite large. It is comprised of two outcome variables, four alternative ways to measure the predictor variable, three functional assumptions, two types of informative variance (between or within firm), and 96 alternative configurations of control variables. A simple product of these options would imply 4,608 models, but some are not unique because models with moderating interactions should include certain other variables, and some variables, such as industry advertising intensity, are redundant in FE models. When these non-unique models are removed,

we have 3,200 distinct models of the relationship between social and financial performance (see Appendix 2 for the computation of the final model set).

Insert Figure 1 about here

ANALYTICAL PROCESS

Data Sources and Sample Creation

KLD began collecting data on firm social and environmental dimensions in 1991, starting with an initial sample of 650 firms (mainly S&P 500 firms within the KLD 400 Social Index). In 1988, it increased the sample to the 1000 largest firms in the United States; and in 2013, KLD expanded their analysis to non-US companies. Across years, KLD dimensions expanded as well. It started with eight attributes, as investigated by Waddock and Graves (2007). For five of which (employee relations, product, community relations, environment and diversity), firms receive scores based on both their strengths and weaknesses (the score ranges from +1 to -1, where -1 identifies an area of weakness, +1 an area of strength, and 0 captures a neutral score). For three attributes, firms receive scores indicating weaknesses only (i.e. South Africa, military and nuclear power). Later, some attributes were removed from the KLD analysis (e.g., South Africa in 1994); others moved and merged into new dimensions (e.g. in 2004, KLD moved the “Indigenous Peoples Relations Strength” from “Community” to “Human Rights”), while others were renamed and expanded (for instance, the “Other category” was renamed “Corporate Governance” in 2002). Currently, there are 13 categories: firms receive scores for both strengths and weaknesses for seven (e.g., corporate governance and human rights) and weaknesses only for six of them (e.g., alcohol, gambling, tobacco, firearms).

To measure each firm's financial performance (FP), we obtained financial data from Compustat and CRSP data at Wharton's WRDS data center. These data provide us with all of our firm-level financial measures, along with the number of employees 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 identically across the two databases. Thus, to match the two databases, we constructed a program that employed various alternatives for name and IDs. Matches were ranked and then checked for accuracy by the researchers. In total, we were able to match 5,986 firms for 42,736 firm-year observations, between 1991 and 2015. As we explain below, this is not our final sample.

Restricted and unrestricted samples

As in every dataset, ours has a number of issues that require researcher judgement. First, many of the observations in the sample are missing values. Usually, this missing information relates to R&D expenditures or advertising. To conduct our Bayesian analysis of model selection, we need to use a consistent sample across all of the models, but to do this, we must remove from analysis many firms. To allow analysis while assuring robustness, we chose to create a "restricted" sample of observations that would be consistent across all specifications. We also tested the reliability of our results by graphing estimates from the full, unrestricted, sample.

In a very few cases, estimated measures were many standard deviations outside the norm. For example, our measure of return on assets contains a few observations with values below -10,000. Given the median of ROA is 3.48 in the full sample, we decided to exclude observations above 500% and below -500% ROA. Likewise, we decided not to include observations where the reported R&D spending exceeded sales. These observations tended to be for biotech innovation firms that had no products.

To create our restricted sample, we first estimated all of the models and marked the sample with the least number of observations. We then used these observations as our restricted sample of 15,066 observations. This sample includes 2,024 firms, the average of which is observed over 8.6 years.

Variable Construction

Dependent variables. Prior research has used a range of accounting and market-based firm performance measurements. We select the common form of each type to analyze. We calculate Return on Assets (ROA) as the net income divided by the total asset of the firm in a given year. We multiplied 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. Were we able to measure the replacement value of assets our measure would be the same as Tobin's Q.

Independent variables. We built four different scales of social performance using KLD data.

The first one, *W&G*, uses the method proposed by Waddock and Graves (1997). Their scale is a weighted average score 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 weightings of these measures are based on expert opinion (see (Waddock and Graves, 1997) for details).

H&R is constructed following the method proposed by Hull and Rothenberg (2008), which is based on Waddock and Graves (1997). They, however, used a different weight system in which dimensions with sub-categories (e.g., environment) received “a disproportionately greater weight than those (e.g., tobacco) with only one subcategory” (Hull and Rothenberg, 2008: 784). Unfortunately, they do not specify the weighting system. Given this limitation, we use the five dimensions with sub-categories used by Waddock and Graves (1997) and added one more, corporate governance as “potential social concern” (Hull and Rothenberg, 2008: 784).

H&K is based on the scale proposed by Hillman and Keim (2001) for stakeholder management. 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).⁷

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

To allow comparison of effect sizes, we normalize each of these scales so that they have 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 *R&D*, we divide the firm’s annual R&D spending by its total sales in each year.

⁷ For the sake of parsimony and ease of comparison, we do not include their measure of social issue participation.

Control variables. Following prior research, we include additional firm controls such as *size*, calculated by firm sales (\$1M), *risk*, calculated as firm debt/asset ratio, industry advertising expenditure (*adv*). We also calculate two sets of dummy variables to capture *industry* effects at SIC two-digit level and *year* effects. Because of multicollinearity, our models cannot include both *industry* dummies and industry advertising intensity (*adv*). Finally, we included lagged DVs in the control variable set.

Descriptive Statistics

Table 2 shows the descriptive statistics of the restricted sample of 15,066 observations. ROA and MTB have similar means, but quite different standard deviations variance. It is worth noticing that although all the SP independent variables have different means, they are highly correlated with each other, ranging from 0.95 to 0.99.

 Insert Table 1 about here

Statistical Analysis

Following prior research, we estimate simple linear models, quadratic models, and models with moderating variables (R&D). All three types of models are estimated with and without firm-level “fixed effects”. The former models, which we will refer to as “OLS” models, estimate the relationship between SP and FP using both within firm and between firm variability. The latter models, which we will refer to as “FE” models, estimate the relationship using only the inter-year variance of firms. For both estimations, we obtain robust standard errors. In total, the possible combinations of empirical assumptions result in 3,200 model specifications (see the computational map to establish the number of models in Appendix 2). We estimated

coefficients for all 3,200 models and stored the marginal effect 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 and the resulting Bayesian Information Criterion (BIC) value.⁸

Graphing

After running all of the models and storing their estimates, we used the STATA graphic interface to display the marginal effects of the SP coefficients. To ease the comparison across models and their interpretation, we divided the marginal effects of the independent variables at their mean by their standard errors (b/se, or the t-statistic). We also calculate the lower and upper bound of the 95% confident interval of the b/se for each model. To build the Figures 2-4, we apply different sorting orders for the b/se, DV, SP variables and the functional models. For instance, Figure 2 includes all the models and it simply sorted by the b/se of each model – from the largest value (on the left) to the smallest value (on the right). The continuous black line depicts the b/se, the gray area its 95% confident interval. Figure 2 is then divided by the DV and sorted by b/se as illustrated in figures 2a (DV=ROA) and 2b (DV=MTB). In addition, we mark six models estimated by the six papers in Appendix 1 by their b/se and confident interval with red dash lines (see the references of these models in the text box next to Figure 2b).

Bayesian Model Selection and Averaging

If researchers and readers have diffuse priors about the correct model to use as inference, they may choose to prioritize estimates from models that better fit the observed evidence.

⁸ We used a routine designed by Brent Goldfarb called *speccall* for Stata that runs and stores all the estimates in stata file.

By Bayes Rule, the posterior probability of any model i , conditional on the observed evidence D , can be calculated as:

$$P(M_i | D) = \frac{P(D|M_i)*P(M_i)}{\sum_1^n (P(D|M_i)*P(M_i))} \quad \text{Eq. 1}$$

where there are 1 to n possible models of data. Since we cannot observe the priors of readers with respect to a group of possible models, and because we are interested in providing information to those readers with equal priors, we assume that $P(M_i) = c$, and $\sum_i^n c = 1$. This means that equation 1 can be simplified to:

$$P(M_i | D) = \frac{P(D|M_i)}{\sum_1^n (P(D|M_i))} \quad \text{Eq. 2}$$

We take advantages of the fact that

$$\frac{P(M_i | D)}{P(M_j | D)} = \frac{e^{-BIC_i/2}}{e^{-BIC_j/2}} \quad \text{Eq. 3}$$

and compute probabilities relative to a reference model r with the highest BIC. If we compare all models to this reference model, then

$$\sum_i^n \frac{P(M_i | D)}{P(M_r | D)} = 1 \quad \text{Eq. 4}$$

Substituting into Equation 2, we can calculate the posterior probability of each model i .

$$P(M_i | D) = \frac{e^{-BIC_i/2}}{e^{-BIC_r/2}} \quad \text{Eq. 5}$$

The probability-weighted coefficient estimate \bar{B} can then be calculated using

$$E(\bar{B} | D) = \sum_i^n B_i \frac{e^{-BIC_i/2}}{e^{-BIC_r/2}}$$

Assuming that the estimate variance s^2 of each model is independent⁹, we can calculate the variance of this pooled coefficient estimate as:

$$E(\bar{s} | D) = \sqrt{\left\{ \sum_i^n s_i^2 \frac{e^{-BIC_i/2}}{e^{-BIC_r/2}} \right\}}$$

RESULTS

Figure 2 shows our estimates for all 3,200 models. We divide these by the dependent variable, with Figure 2a showing estimates of ROA and Fig 2b showing estimates for MTB. In both figures, we graph the t-statistic (B/SE), and the 95% confidence interval for the estimate. Any estimate where the shaded region does not include zero can be thought of as indicating a “significant” coefficient B were the estimate to come from a single, prespecified test.

Both graphs show that if authors were to randomly select and prespecify one of 3,200 models and then conduct a frequentist analysis of B , they would have the highest chance of concluding that they could reject $H: B < 0$, a reasonable chance of concluding that they could not reject the NULL, $H: B = 0$, and a smaller chance of concluding that they should reject $H: B > 0$. For estimates with respect to MTB, the chances would be 47, 41, and 12%. For ROA, the chances would be 50, 47, and 3%. Thus, it appears that if one were to have priors that all models were equally valid, one would be very likely to “find” a positive relationship between social and financial performance. It seems unsurprising then, that previous summaries of research have concluded that most studies find evidence of a positive and significant result.

 Insert Figures 2, 2a, 2b about here

⁹ This is a strong assumption. In future work, we plan to use a more accurate Monte Carlo analysis to construct the posterior probabilities, weighted estimates, and weighted sample variance.

Figure 2 also indicates which estimates come from models used by the six previous studies in Appendix 1. These estimates do not represent exact replications of these previous papers, because our sample differs from previous studies. They represent what those authors would have found had they had used our sample in their analysis. Nevertheless, they seem to corroborate several previously reported estimates.

Using the models proposed by two authors (W&G and H&R), we find a positive and “significant” relationship between their measure of social performance and ROA. Consistent with the results reported in Barnett & Salomon (2012), we find, when using their scale for social performance, a negative, but weak, average marginal relationship between SP and FP. With respect to predictions of the Market to Book ratio, we corroborate Hillman & Keim’s estimate of a positive and significant relationship between social performance and MTB (see Figure 2b).

Interestingly, the model proposed by McWilliams & Siegel (2000) results, in our sample, in a very different outcome than the one they report in their more limited sample. Adding R&D to the model does not eliminate the significance of the SP-FP link. Indeed, the M&S model delivers one of the strongest effects of all. It is important to note, however, that we do not replicate exactly their measure of social performance. They chose to use a binary variable indicating firms that had performed well-enough to be selected for the Domini Social Index. We use a continuous measure which ranks all firms.

What explains different results

The effect of different scales for Social Performance

The analysis reported in Figure 2a & b largely corroborates previously reported findings. It does not clarify why estimates differ so greatly across the previously published results. To explore this, we sort the estimates both by the dependent variable (ROA or MTB) and the various

different implementations of SP measures. For simplicity, we also consider only simple linear models. As shown in Figures 3a and 3b, the result is insensitive to the specific scales employed by the authors. For all measures, a simple OLS model will result in a beta estimate that is positive and significant.

The inclusion of firm attributes, particularly firm size, reduces the size of the coefficient estimate. Inclusion of a lagged dependent variable also reduces the scale of the estimated effect. Nevertheless, even models with a full complement of controls and lagged variables still results in an estimate of a significant and positive coefficient for ALL four different implementations of SP scales.

 Insert Figures 3a and 3b about here

The effect of moderating and mediating effects on estimates for Social Performance

As discussed earlier, several authors have proposed that R&D could influence OLS estimates of the SP-FP relationship. Figure 4a and 4b show estimates for models mediated (including R&D as a control) and moderated by R&D (e.g including SP*R&D). As before, the models are sorted by specification, SP scale, and the b/se or t-statistic (pos to neg). As shown, mediation models result in an estimated of a positive and “significant” association. All moderation models result in an estimate of $B > 0$, but for models including a lagged dv, the 95% confidence interval of the t-statistic includes zero. This is true whether the dependent variable is MTB or ROA.

 Insert Figures 4a and 4b about here

Estimates based on within-firm variance (fixed effects models)

Several authors have proposed that analysis of the SP-FP relationship should be done using only the variance within a firm's history. To do this, they specify a fixed effect (a dummy variable for each firm). This means that all of the estimates for that firm are demeaned, so that any relationship is calculated only from year to year changes in social or financial performance. Barnett and Salomon (2012) further opine that the SP-FP relationship is not linear but curvilinear, and they specify a quadratic model. As shown in Figures 5a and 5b, including fixed-effects has a dramatic impact on the estimated relationship. Now, only 8 models (<1%) result in an estimate of $B > 0$, and none of these estimates pass traditional significance tests. For the rest of these models, the estimate of B is negative, and for 60 (6% of the total) this estimate would traditionally be deemed negative and significant. That is, authors conducting a single test using one of these models would conclude they should reject $H: B > 0$.

 Insert Figures 5a and 5b about here

Finding the “Best” model

Readers of this article will bring to it their own priors about which models are most appropriate. Those who think estimates should be obtained by comparing market valuation (MTB) across firms may conclude that our results reveal that SP & FP are likely to be associated in future samples from the same population. It should not matter whether or not these readers have strong priors about the correct way to create SP scales from KLD data or the specific controls that should be included. What may change a reader's inference, is a belief that models should include firm fixed effects. Readers with this belief might infer that the association between SP & ROA is weak or missing. Those that think SP should be estimated using ROA

may even conclude that they should reject the hypothesis that there is a positive association between SP & FP.

 Insert Table 2 about here

For those readers who are unsure which models are most appropriate (those with “diffuse priors”), they might prefer to observe the preponderance of evidence across all of the models, or to weight certain models with better “fit” as particularly useful for inference. To aid these readers, we conducted further analysis by calculating how a reader with diffuse priors should assess the models based on their fit alone.

Table 2 shows the models chosen to be “best” within a given group. Models are grouped by different dependent variables, measures of social performance, and the inclusion of firm fixed effects (FE). For each of these groups of models, the table reports the specification of the model with the highest posterior probability (i.e. the “best” model) and the coefficient estimates for $B \cdot SP$ for that model. It also reports an estimate for B & s for all models in that group based on an average weighted by each model’s posterior probability $P(M|E)$.

As shown in Table 2, for each group of models (DV, IV, FE[0,1]), the same specifications had the highest posterior probability, regardless of the specific scale of SP chosen. In all cases, these “best” models included a lagged DV, R&D ratio, and year fixed effects. In some cases, best models included advertising, sales, or interaction terms.

For OLS models predicting MTB, all of the “best” models resulted in coefficient estimates of a significant and positive relationship between SP & FP. However, the results for all other “best” models were much more mixed. Fixed effects models of MTB resulted in estimates of small B coefficients, and traditional significance tests would imply that the NULL hypothesis could not be rejected with confidence. Also, OLS estimates of ROA resulted in positive but

uncertain estimates for which the NULL hypothesis could not be rejected¹⁰. Fixed effect predictions of ROA suggest the exact opposite. Now, coefficient estimates were mostly negative.

Weighted estimates of coefficients and standard errors closely matched those of the best models. For every group, the best model made up more than 50% of the total probability weight of the average estimates.¹¹ Thus, the results from probability weighted estimate closely match those of the “best” models.

For those readers who are indifferent to the use of a particular scale, we also calculate the best model across all scales for the set of models (DV[MTB,ROA], FE([0,1])). We find that the W&G scale or the H&K scale is usually selected as “best”. We also find that the set of models point in opposite directions with respect to inference. Models of MTB using OLS result in a *B* estimate are suggestive of a positive association between SP & FP. Models of ROA using firm fixed effect result a *B* estimate suggestive of a negative association between SP & FP.

DISCUSSION

Our analysis results in coefficients of conflicting sign, scale, and significance. How do we think it should be interpreted? We know that it is not amenable to frequentist interpretations because we did not stipulate in advance a specific hypothesis, specification, and sampling plan. Nor can it be interpreted in a Bayesian manner because we have not attempted to estimate the probability of all other explanations of the observed data. Thus, we can only engage in a kind of abduction, or conjecture making, about the patterns we see. These “guesses” should then be subject to test.

¹⁰ Earlier, we confirmed that Waddock & Graves’ (1997) model results in a positive and significant coefficient estimate. Our Bayesian analysis suggest that their model delivers a stronger result than the “best” of the OLS models.

¹¹ This is consistent with results from previous research using Bayesian model averaging (Durlauf et al, 2016).

Given our uncertainty about the data-generation process, no single model is highly informative.

Our first observation is that specification choice determines the estimates obtained. This means that depending on author priors and assumptions about the right specification, different results will be obtained. It also means that less scrupulous scholars could search through a range of specifications to find any result they desired, positive, negative, or inconsequential. Thus, readers should be cautious in extrapolating from any particular study.

The effect of R&D remains unknown

Based on prior work, we had expected that the addition of R&D as a control would reduce the size of the coefficient for SP. We found a strong and robust effect in the opposite direction. Even specifications used by previous authors resulted in estimates of a positive coefficient for both R&D and for the interaction of R&D and SP. One might conjecture that innovation is a complement to social performance. It is also possible that the effect depends on some elements of the sample. Older studies used a smaller group of firms and could observe them only over a short time period. At best, our analysis suggests that the effect of R&D remains unknown.

Among all firms, those with higher social performance tend to do well financially

We observe that positive estimates for the SP-FP relationship arise from specifications where firm observations are pooled, while inconsequential or negative estimates arise from analysis that includes firm fixed effects. In other words, across firms, there is an association between social performance and financial performance. This association holds even when current financial performance is taken into account by adding a lagged term. Thus, it appears that across all firms, there is a positive association between higher social performance than others and an increase in financial performance. One should, of course, not assume this is a durable or

predictive statement. It is tempting, however, to conjecture that all of the money flowing into SP-screened funds has raised the stock prices of the best rated firms.

Year to year changes in performance are associated with negative or negligible impact.

Models that include firm-level fixed effects result in coefficient estimates that are small or negative. Why might this be? One available explanation is that an increase in social performance involves cost, and this directly effects a firm's ROA or market valuation. But if improvement in Social Performance is costly, then why do some high SP firms seem to gain financially (as suggested by the pooled analysis)? There are many possible explanations. One might be that all of our specifications use the wrong lag structure and financial benefits appear only a year or more in the future. Another might be that first movers gained capabilities that now allow them to reap financial gains. Yet another explanation is that some attribute of social performance is highly correlated with unmeasured sectoral differences and it is these differences that are actually associated with financial performance. In future work, we intend to explore these conjectures.

Limitations

Our study has many limitations, and should be extended in several ways. First, we do not consider assumptions about identification, and thus we can say nothing about claims about causal connection. Second, we do not consider a variety of other methods for conveying the mapped estimates. We could, for example, simply post a look-up table so that any reader can connect her assumptions to the implied estimates. Finally, we consider only a small subset of studies in creating our map of feasible assumptions. Much remains to be done, and our research is just a small step.

CONCLUSION

At the highest level, our analysis shows how we can build shared understanding of important topics such as the link between social and financial performance. We show that individual research estimates are sensitive to empirical assumptions, and thus any use of these estimates as a basis of understanding must be contingent on the reader's acceptance of these assumptions. To allow competent use of empirical analysis, researchers must be transparent about the choices they made as they walked through the empirical "garden of forking paths". We also show how collections of assumptions can be mapped to a model space and associated with coefficient estimates: maps can be constructed to allow readers with strong priors to find a connection between favored assumptions and their linked estimates; Bayesian analysis can be used to allow readers with uncertain priors to pick a "best", or "average" estimate for inference. In total, our analysis shows how even conflicting results can be aggregated into a synthesized understanding of what may be known and what surely remains unknown.

With respect to the SP-FP link, we show that it is in fact possible to make "worthwhile comparisons" across studies, and that such research can be more than just a way to "legitimize the researcher and the business & society field" (Rowley and Berman, 2000). We show how empirical studies, even those with opposing findings, can be integrated to improve our understanding.

We hope that other researchers will adopt the methods used in this analysis. Personally, we found it liberating to be free of the need to select and defend a particular set of assumptions and model specifications. Rather than worry about selecting and defending a particular a particular path through the empirical "garden", we could attempt to synthesize and aggregate previous studies. We could enjoy the process of understanding assumptions made by prior research. We felt ourselves free to seek understanding, free of the usual pressure to find and

defend a single “significant” coefficient estimate. We had more fun, learned more, and held out some hope that we helped others to reach their own understanding as well – regardless of the assumptions from which they started.

REFERENCES

- Aguinis, H. and A. Glavas (2012). What We Know and Don't Know About Corporate Social Responsibility: A Review and Research Agenda. *Journal of Management* **38**(4): 932-968.
- GSIA (2019). Global sustainable investment review. *Global Sustainable Investment Alliance*: Washington, DC, USA.
- Barnett, M. L. and R. M. Salomon (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., S. Ethiraj, A. Gambardella, C. Helfat and W. Mitchell (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., C. E. Helfat and J. M. Shaver (2016). The necessity, logic, and forms of replication. *Strategic Management Journal* **37**(11): 2193-2203.
- De Long, J. B. and K. Lang (1992). Are all economic hypotheses false? *Journal of Political Economy* **100**(6): 1257-1272.
- Durlauf, S. N., S. Navarro and D. A. Rivers (2016). Model uncertainty and the effect of shall-issue right-to-carry laws on crime. *European Economic Review* **81**: 32-67.
- Fisher, R. (1960). *The design of experiments* (1935, 1st). Edinburgh: Oliver and Boyde.
- Fricker, E. (2002). Trusting others in the sciences: a priori or empirical warrant? *Studies in History 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 and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University. Retrieved from www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
- Hacking, I. (1965). *The Logic of Statistical Inference*. Cambridge, UK, Cambridge University Press.
- Hillman, A. J. and G. D. Keim (2001). Shareholder value, stakeholder management, and social issues: What's the bottom line? *Strategic Management Journal* **22**(2): 125-139.
- Hull, C. E. and S. Rothenberg (2008). Firm performance: The interactions of corporate social performance with innovation and industry differentiation. *Strategic Management Journal* **29**(7): 781-789.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies* (Vol. 4). Thousand Oaks, CA: Sage Publications, Inc.
- King, A., B. Goldfarb and T. Simcoe (2019). Learning from testimony on quantitative research in management, Working Paper.
- Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review* **75**(3): 308-313.
- Longino, H. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*, Princeton University Press.

- Margolis, J. D. and J. P. Walsh (2001). *People and profits? : the search for a link between a company's social and financial performance*. Mahwah, N.J., Lawrence Erlbaum Associates.
- McWilliams, A. and D. Siegel (2000). Corporate social responsibility and financial performance: Correlation or misspecification? *Strategic Management Journal* **21**(5): 603-609.
- Orlitzky, M., F. L. Schmidt and S. L. Rynes (2003). Corporate social and financial performance: A meta-analysis. *Organization Studies* **24**(3): 403-441.
- Porter, G. S., G. Serafeim and M. Kramer (2019). Where ESG Fails. *Institutional Investor*, October
- Rowley, T. and S. Berman (2000). A brand new brand of corporate social performance. *Business & society* **39**(4): 397-418.
- Sala-i-Martin, X. X. (1997). I just ran four million regressions, National Bureau of Economic Research.
- Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics* **102**(1): 411-432.
- Simmons, J. P., L. D. Nelson and U. Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* **22**(11): 1359-1366.
- Simonsohn, U., L. D. Nelson and J. P. Simmons (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General* **143**(2): 534.
- Simonsohn, U., J. P. Simmons and L. D. Nelson (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications., Working Paper.
- Spanos, A. (2010). Is frequentist testing vulnerable to the base-rate fallacy? *Philosophy of Science* **77**(4): 565-583.
- Waddock, S. (2004). Parallel universes: Companies, academics, and the progress of corporate citizenship. *Business and Society Review* **109**(1): 5-42.
- Waddock, S. A. and S. B. Graves (1997). The corporate social performance - Financial performance link. *Strategic Management Journal* **18**(4): 303-319.
- Wilholt, T. (2013). Epistemic trust in science. *The British Journal for the Philosophy of Science* **64**(2): 233-253.
- Zhao, X. P. and A. J. Murrell (2016). Revisiting The Corporate Social Performance-Financial Performance Link: A Replication Of Waddock And Graves. *Strategic Management Journal* **37**(11): 2378-2388.

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

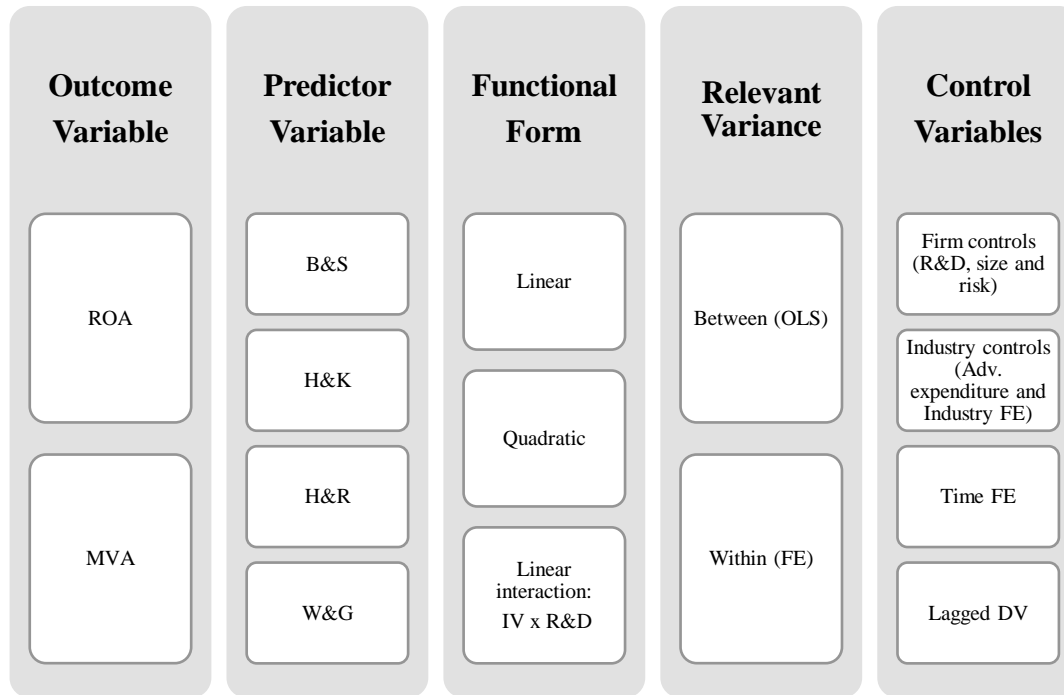


Figure 2. All 3,200 models sorted by b/se.

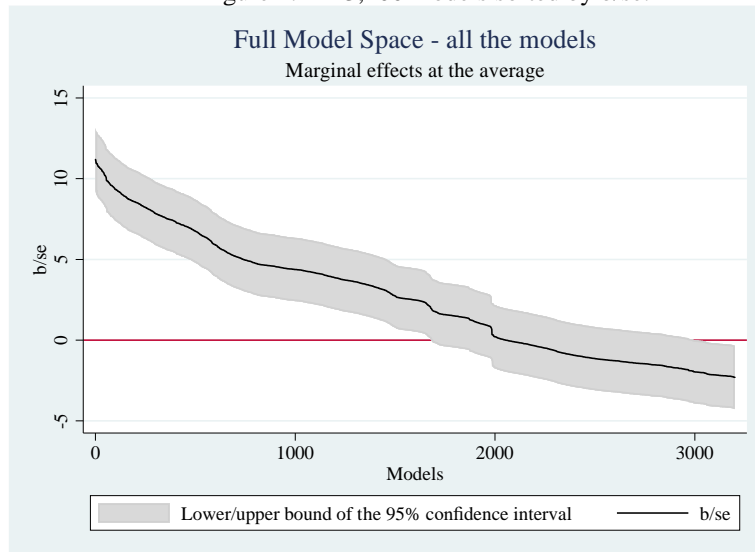
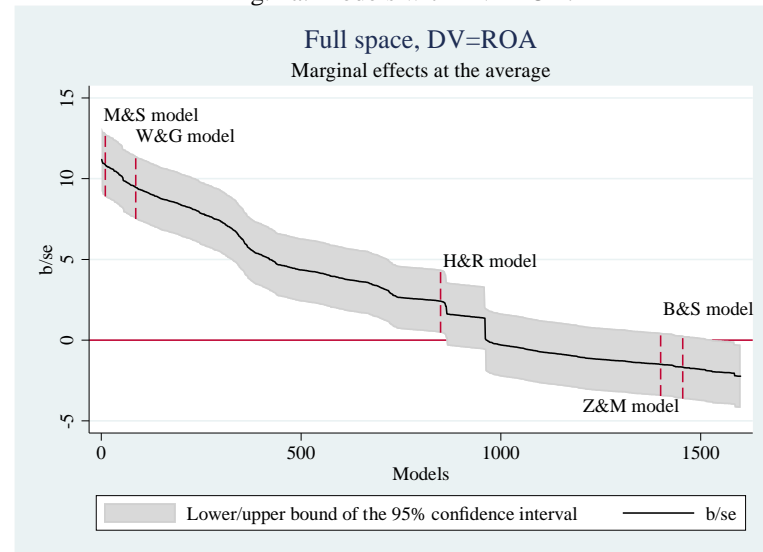


Fig. 2a. Models with DV=ROA.



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&K model” to Table 2 in Hillman & Klein (2001);
- “H&R model” to Model 3, Table 2 in Hull & Rothenberg (2008);
- “B&S model” to Model 6, Table 2 in Barnett & Salomon (2012); and
- “Z&M model” to Table 4 with DV=ROA in Zhao and Murrell (2016).

Fig. 2b. Models with DV=MTB

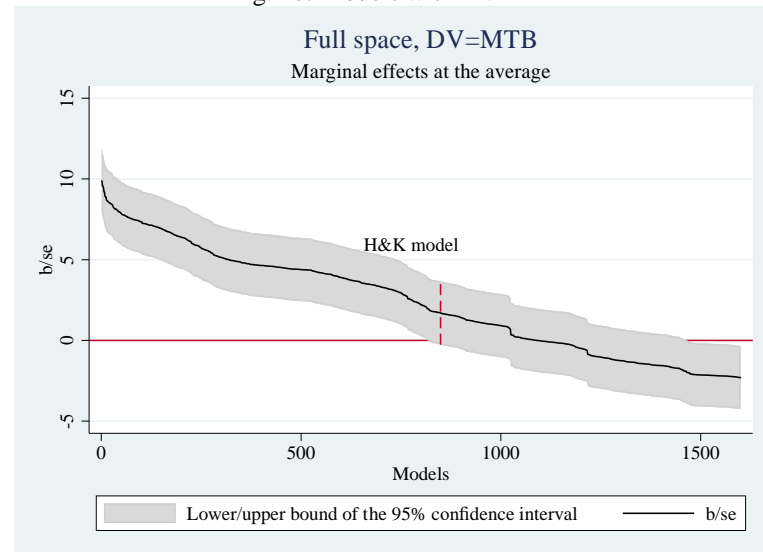


Figure 3a: All linear OLS models with DV=ROA.

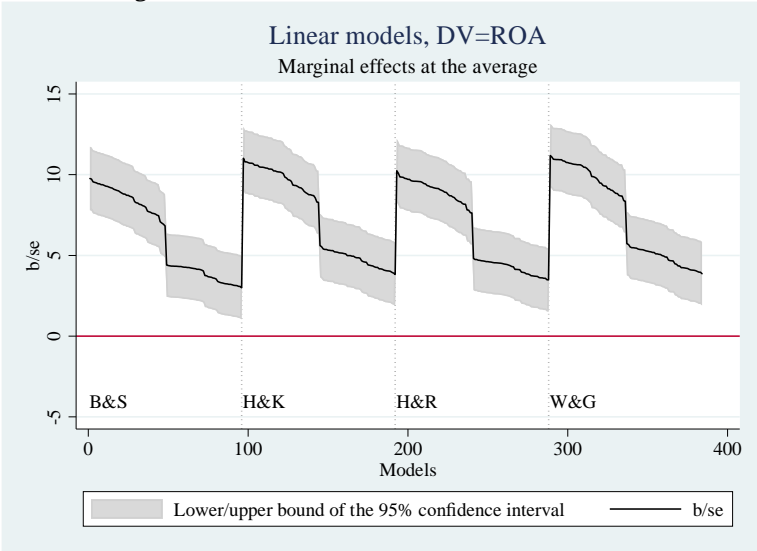


Figure 3b: All linear OLS models with DV=MTB.

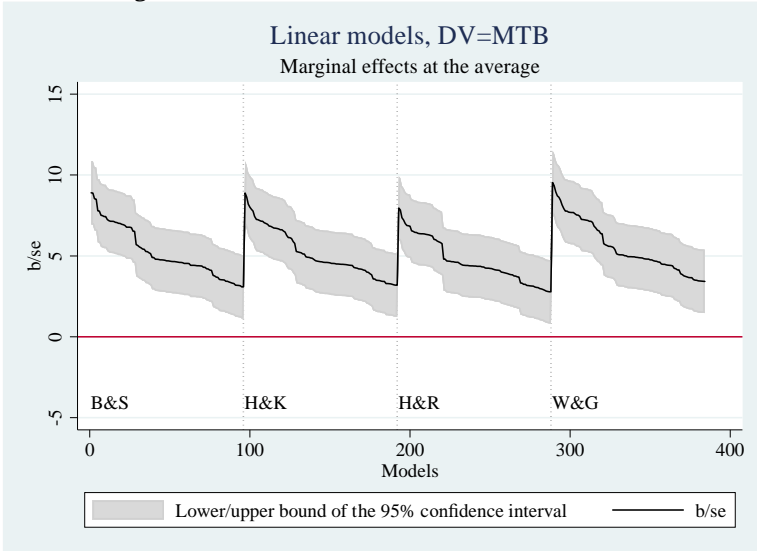


Fig. 4a. Linear OLS models with RD as control and as moderator, DV=ROA

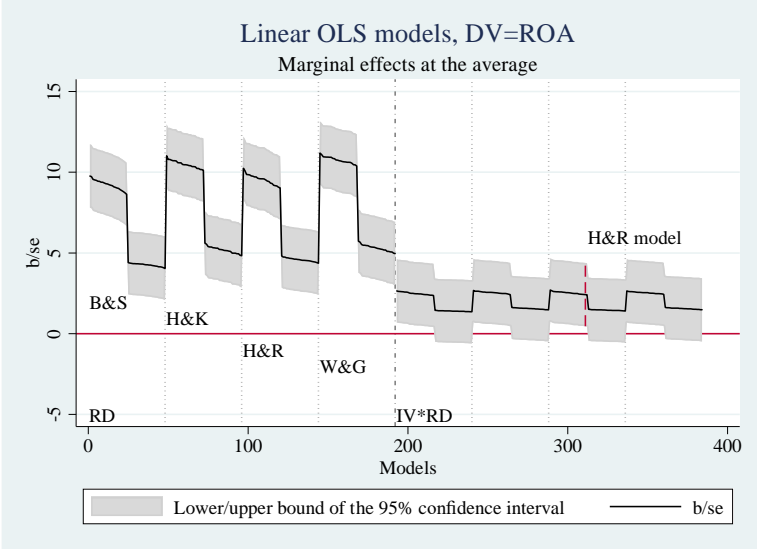


Fig. 4b. Linear OLS models with RD as control and as moderator, DV=MTB

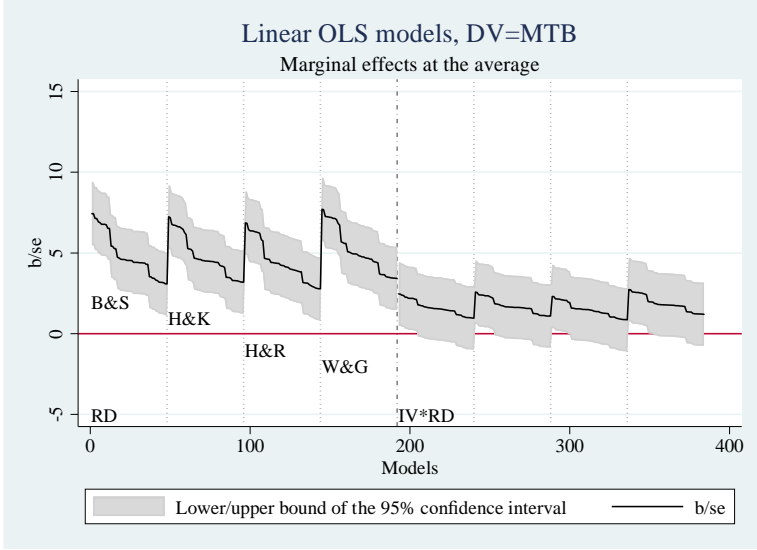


Fig. 5a. All linear FE models separated from all quadratic FE models. DV=ROA.

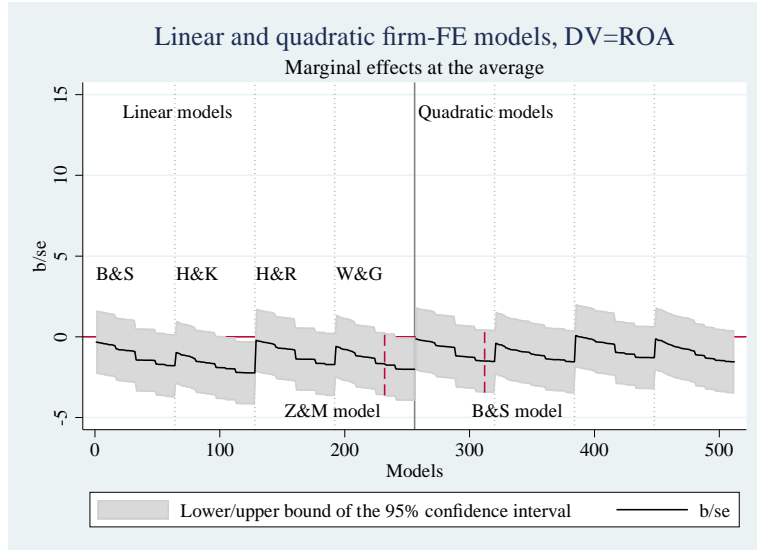


Fig. 5b. All linear FE models separated from all quadratic FE models. DV= MTB.

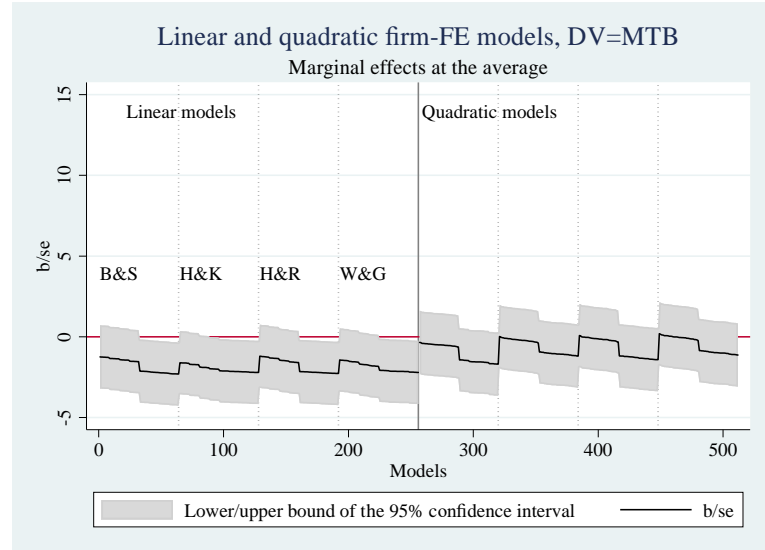


Table 1. Descriptive statistics.

| Variable | Mean | Std. Dev. | Min | Max | | | | | | | | | |
|----------|----------|-----------|----------|---------|--------|--------|--------|--------|--------|--------|--------|-------|--------|
| ROA | 4.237 | 12.128 | -317.310 | 69.501 | 1 | | | | | | | | |
| MTB | 2.486 | 4.275 | -247.956 | 139.612 | 0.112 | 1 | | | | | | | |
| B&S | -0.126 | 1.102 | -5.220 | 7.228 | 0.072 | 0.054 | 1 | | | | | | |
| H&K | 0.055 | 1.123 | -4.414 | 7.504 | 0.084 | 0.052 | 0.950 | 1 | | | | | |
| H&R | -0.057 | 1.110 | -4.553 | 7.036 | 0.080 | 0.048 | 0.985 | 0.968 | 1 | | | | |
| W&G | 0.025 | 1.119 | -4.509 | 7.294 | 0.083 | 0.056 | 0.958 | 0.995 | 0.965 | 1 | | | |
| R&D | 0.068 | 0.106 | 0 | 0.990 | -0.323 | 0.177 | 0.053 | 0.050 | 0.037 | 0.058 | 1 | | |
| Size | 6087.398 | 21730.540 | -475.42 | 483521 | 0.058 | -0.017 | -0.007 | 0.078 | 0.040 | 0.059 | -0.091 | 1 | |
| Risk | 0.187 | 0.197 | 0 | 2.095 | -0.150 | -0.162 | -0.054 | -0.043 | -0.048 | -0.045 | -0.190 | 0.007 | 1 |
| Adv | 4.363 | 1.344 | 0.188 | 7.660 | 0.053 | 0.059 | 0.031 | 0.066 | 0.055 | 0.054 | 0.035 | 0.115 | -0.062 |
| Obs. | 15,066 | | | | | | | | | | | | |

Table 2: "Best" and probability-weighted estimates of SP/FP association.

| Model Group | | | Specification of "best" model All Independent Variables | Estimates: "best" model | | | | Weighted Estimates | | |
|----------------------|-----------------------|---------|--|-------------------------|-------|--------------|--------|--------------------|----------------|--------------------|
| DV | SP _{measure} | Firm FE | | B | s | T | P(M E) | B _w | S _w | T _{Bw/Sw} |
| MTB _(t+1) | B&S | No | SP MTB(t) R&D risk adv year(fe) | 0.071 | 0.022 | 3.19 | 66% | 0.082 | 0.025 | 3.29 |
| MTB _(t+1) | H&K | No | SP MTB(t) R&D risk adv year(fe) | 0.070 | 0.022 | 3.20 | 58% | 0.081 | 0.035 | 2.33 |
| MTB _(t+1) | H&R | No | SP MTB(t) R&D risk adv year(fe) | 0.062 | 0.022 | 2.79 | 76% | 0.069 | 0.025 | 2.72 |
| MTB _(t+1) | W&G | No | SP MTB(t) R&D risk adv year(fe) | 0.076 | 0.022 | 3.44 | 63% | 0.087 | 0.028 | 3.14 |
| MTB _(t+1) | B&S | Yes | SP SPxRD MTB _(t) R&D sales year(fe) | -0.006 | 0.051 | -0.11 | 98% | -0.006 | 0.051 | -0.12 |
| MTB _(t+1) | H&K | Yes | SP SPxRD MTB _(t) R&D sales year(fe) | 0.016 | 0.051 | 0.31 | 99% | 0.016 | 0.051 | 0.31 |
| MTB _(t+1) | H&R | Yes | SP SPxRD MTB _(t) R&D sales year(fe) | 0.004 | 0.050 | 0.09 | 99% | 0.004 | 0.050 | 0.09 |
| MTB _(t+1) | W&G | Yes | SP SPxRD MTB _(t) R&D sales year(fe) | 0.009 | 0.050 | 0.18 | 98% | 0.009 | 0.050 | 0.17 |
| ROA _(t+1) | B&S | No | SP SPxRD ROA(t) R&D risk adv year(fe) | 0.344 | 0.246 | 1.40 | 95% | 0.343 | 0.246 | 1.40 |
| ROA _(t+1) | H&K | No | SP SPxRD ROA(t) R&D risk adv year(fe) | 0.398 | 0.251 | 1.59 | 95% | 0.398 | 0.251 | 1.58 |
| ROA _(t+1) | H&R | No | SP SPxRD ROA(t) R&D risk adv year(fe) | 0.353 | 0.244 | 1.44 | 95% | 0.353 | 0.245 | 1.44 |
| ROA _(t+1) | W&G | No | SP SPxRD ROA(t) R&D risk adv year(fe) | 0.405 | 0.255 | 1.59 | 95% | 0.404 | 0.255 | 1.58 |
| ROA _(t+1) | B&S | Yes | SP ROA _(t) R&D year(fe) | -0.215 | 0.127 | -1.69 | 90% | -0.222 | 0.130 | -1.71 |
| ROA _(t+1) | H&K | Yes | SP ROA _(t) R&D year(fe) | -0.267 | 0.121 | -2.21 | 93% | -0.269 | 0.123 | -2.19 |
| ROA _(t+1) | H&R | Yes | SP ROA _(t) R&D year(fe) | -0.211 | 0.128 | -1.65 | 90% | -0.218 | 0.131 | -1.66 |
| ROA _(t+1) | W&G | Yes | SP ROA _(t) R&D year(fe) | -0.236 | 0.118 | -1.99 | 91% | -0.241 | 0.122 | -1.98 |
| MTB _(t+1) | All | No | SP(W&G) MTB(t) R&D risk adv year(fe) | 0.076 | 0.022 | 3.44 | 34% | 0.084 | 0.029 | 2.89 |
| MTB _(t+1) | All | Yes | SP(H&K) SPxRD MTB _(t) R&D sales year(fe) | 0.016 | 0.051 | 0.31 | 90% | 0.015 | 0.051 | 0.29 |
| ROA _(t+1) | All | No | SP(W&G) SPxRD ROA(t) R&D risk adv year(fe) | 0.405 | 0.255 | 1.59 | 50% | 0.401 | 0.253 | 1.58 |
| ROA _(t+1) | All | Yes | SP(H&K) ROA _(t) R&D year(fe) | -0.267 | 0.121 | -2.21 | 44% | -0.248 | 0.125 | -1.98 |

Appendix 1: Attributes of Six Papers Used to Establish the Model Space

| Functional form | Specification | Seminal Paper | DV | IV | Controls | Main Findings |
|---------------------------------|---|--|--|--|---|--|
| Linear form | $R_{it} = \alpha + \beta X_{it-1} + \beta \theta_{it-1} + \varepsilon_{it}$ | <p>Waddock and Graves (1997)</p> <p>McWilliams and Siegel (2000)</p> <p>Hull and Rothenberg (2008)</p> | <p>ROA, ROE, ROS</p> <p>Accounting and market variables (not specified)</p> <p>ROA</p> | <p>Weighted SP attributes</p> <p>SP dummy variable for being part of the KLD data in 1991-1996</p> <p>Weighted SP attributes as in Waddock and Graves (1997)</p> <p>Moderators: Industry Differentiation (ind. Adv. intensity) and innovation (R&D exp.)</p> | <p>Debt/total assets, Total sales, Total assets, Number of employees</p> <p>Size, R&D intensity, industry advertising intensity</p> <p>Assets, sales, risk,</p> | <p>SP positively and significantly associated with ROA & ROS.</p> <p>No effect of SP on FP when including R&D intensity</p> <p>SP affects FP in low-innovation firms and in industries with little differentiation</p> |
| Linear with FE at firm | $R_{it} = \beta X_{it-1} + \beta \theta_{it-1} + \varepsilon_{it} + u_i$ | Zhao and Murrell (2016) replication study of Waddock and Graves (1997) | ROA, ROE, ROS, Tobin's q, MTB, MVA | Weighted SP attributes as in Waddock and Graves (1997) | Debt/total assets, Total sales, Total assets, Number of employees | Positive and significant effect of SP on FP only when using ROA as DV. |
| Linear with lagged DV | $R_{it} = \alpha + R_{it-1} + \beta X_{it-1} + \beta \theta_{it-1} + \varepsilon_{it}$ | Barnett and Salomon (2012) Model 1 and 2 | ROA, Net Income | KLD score (the sum of strengths and weaknesses of 7 social performance criteria, minus the weaknesses in controversial business activities) | Size, R&D intensity, Advertising intensity | Positive and significant relationship |
| Quadratic with FE and lagged DV | $R_{it} = R_{it-1} + \beta X_{it-1} + \beta X_{it-1}^2 + \beta \theta_{it-1} + \varepsilon_{it} + u_i$ | Barnett and Salomon (2012) | | | | Statistically significant U-shaped relationship between SP and FP |
| Linear with first difference | $R_{it} - R_{it-1} = \beta X_{it} + \beta \theta_{it-1} + (\varepsilon_{it})$ <p>Also implied, but not tested, is a fully differenced model:</p> $R_{it} - R_{it-1} = \beta (X_{it} - X_{it-1}) + \beta (\theta_{it} - \theta_{it-1}) + \varepsilon_{it}$ | Hillman and Keim (2001) | MVA | SM (empl. relations, diversity issues, product issues, community relations, and env issue) and SIP (non-U.S. issues, Other, Exclusionary screens) | Sales, net income, risk, industry | SM positively and significantly associated with MVA SP negatively and significantly associated with MVA |

Appendix 2. Computational map to determine the model space to be mapped and the models to estimate (based on Figure 1). In total 3,200 models.

| Outcome | Predictor | Function | Estimate | Controls | X | total | Outcome | Predictor | Function | Estimate | Controls | X | total |
|---------|-----------|-----------|---------------|---|----|-------|---------|-----------|--------------------|---------------|---|----|-------|
| ROA | B&S | Linear | Between (OLS) | Size (0,1) | 2X | | ROA | B&S | moderation IV x RD | Between (OLS) | Size (0,1) | 2X | |
| MTB | H&K | Quadratic | | Adv (0,1), FE industry (0,1) [not together] | 3X | | MTB | H&K | | | Adv (0,1), FE industry (0,1) [not together] | 3X | |
| | H&R | | | FE Time (0,1) | 2X | | | H&R | | | FE Time (0,1) | 2X | |
| | W&G | | | Lagged DV (0,1) | 2X | | | W&G | | | Lagged DV (0,1) | 2X | |
| | | | | R&D (0, 1) | 2X | | | | | | R&D (1) | 1X | |
| | | | | Risk (0, 1) | 2X | | | | | | Risk (0, 1) | 2X | |
| 2X | 4X | 2X | 1X | | 96 | =1536 | 2X | 4X | 1X | 1X | | 48 | =384 |
| Outcome | Predictor | Function | Estimate | Controls | X | total | Outcome | Predictor | Function | Estimate | Controls | X | total |
| ROA | B&S | Linear | Within (FE) | Size (0,1) | 2X | | ROA | B&S | moderation IV x RD | Within (FE) | Size (0,1) | 2X | |
| MTB | H&K | Quadratic | | Adv (0,1), FE industry (0,1) [not together] | 2X | | MTB | H&K | | | Adv (0,1), FE industry (0,1) [not together] | 2X | |
| | H&R | | | FE Time (0,1) | 2X | | | H&R | | | FE Time (0,1) | 2X | |
| | W&G | | | Lagged DV (0,1) | 2X | | | W&G | | | Lagged DV (0,1) | 2X | |
| | | | | R&D (0, 1) | 2X | | | | | | R&D (1) | 1X | |
| | | | | Risk (0, 1) | 2X | | | | | | Risk (0, 1) | 2X | |
| 2X | 4X | 2X | 1X | | 64 | =1024 | 2X | 4X | 1X | 1X | | 32 | =256 |

The upper left quadrant illustrates the model space for linear and quadratic OLS models. The product of the control variables (96 combinations) with the DV (2), SP variables (4), functions (2) and OLS estimate (1) produces 1,536 models. The lower left quadrant depicts the model space for linear and quadratic FE models. The combinations of control variables are 64 because industry-FE are excluded. The upper right quadrant includes models with the moderation effect of RD only for linear OLS models. There are 48 models since R&D needs to be present in every model. The right lower quadrant includes FE models with the moderation effect of R&D.