Corporate Sustainability:
A Model Uncertainty Analysis of Materiality

Luca Berchicci
Rotterdam School of Management (RSM)
Erasmus University Rotterdam
lberchicci@rsm.nl

Andrew A. King
Questrom School of Business
Boston University
aaking@bu.edu

Draft: March 23, 2021

Abstract: For more than thirty years, scholars have investigated the connection between corporate sustainability and financial performance. In 2016, Khan, Serafeim, and Yoon published what appeared to be a major breakthrough in this quest. They reported that materiality guidance from the Sustainability Accounting Standards Board (SASB) enabled the formation of weighted scales of sustainability measures that robustly predict stock returns. Their publication has influenced greatly both practice and scholarship, but it remains an initial assessment. In this article, we extend the analysis of SASB materiality-weighting by conducting a model uncertainty analysis. We replicate the 2016 estimate, but show that it is unrepresentative of the pattern of results from other reasonable models, and may be a statistical artifact. Finally, we turn to machine learning to explore the prospects for useful materiality guidance, and show that for one popular source of data on corporate sustainability, predictive guidance may be difficult to achieve.

149 words

Keywords: materiality, social and financial performance, research methods, epistemology, model uncertainty, replication.

JEL Classifications: Q51, D22, L25, C11, C18
1. Introduction

For more than thirty years, scholars and investors have searched in vain for a robust association between a corporation’s present performance with respect to “sustainability”\(^1\) and the future returns of its stock (Orlitzky et al. 2003, Orlitzky 2013). From this frustrating history, some scholars infer that no reliable connection exists, but others conclude only that present-day measures of corporate sustainability are unsatisfactory (Porter et al. 2019). Most scales of corporate sustainability aggregate together different types of actions, and whether these actions are material to investors is seldom considered. Moreover, as Eccles and Serafeim (2013) point out, the materiality of different actions may depend on the sector in which the firm operates: “carbon emissions are more material for a coal-fired utility than for a bank” (2013:5). A valuable measure of corporate sustainability, Eccles and Serafeim (2013) opine, must account for the conditional effect of different actions.

In 2016, Khan, Serafeim, and Yoon (KSY) published a first empirical test of the value of materiality information in the creation of a predictive measure of corporate sustainability. Using new guidance on materiality from the Sustainability Accounting Standards Board (SASB), they filtered existing sustainability data from Kinder, Lydenberg and Domini (KLD) to form industry-contingent scales of corporate performance with respect to sustainability issues. They estimate that had investors possessed their improved scales in the years 1991-2013, they would have been able to select stock portfolios with strikingly higher returns – 3 to 6% per year (Khan et al. 2016).

The impact of Khan, Serafeim, & Yoon (2016) is hard to overstate. It has influenced scholars, advocacy organizations, corporate managers, and investors. It is widely interpreted as demonstrating both the value of SASB’s materiality measures and the existence of a real connection between corporate sustainability and financial performance (SASB 2017; CERES 2020; Chasan 2019). In two years after the release of KSY’s working paper, funds using SASB data more than doubled their assets under

---

\(^1\) Following precedent, we will measure corporate sustainability as a combination of a corporation’s performance with respect to the natural environment, its social stakeholders, and the governance it uses in managing its operations.
management, to $50 trillion (SASB 2017). Advocacy organizations now encourage corporations to “focus reporting on the most material issues”, and hundreds of organizations now use SASB materiality guidance in creating their sustainability reports (CERES 2020).

Yet, there are also reasons to be cautious in drawing general conclusions from KSY’s results. The accuracy of sustainability measures, including those used by KSY, have been called into question (Berg et al. 2020). Moreover, promising associations between sustainability scales and stock returns have been reported before, but eventually proven to be fragile or spurious (McWilliams and Siegel, 2000; Porter, Serafeim, and Kramer, 2019). Most importantly, Khan, Serafeim, and Yoon (2016) remains, in its author’s words, only “first evidence”, and thus it inevitably provides a limited basis for inference. Its methods are sophisticated, and its analysis seems unimpeachable, but its results are contingent on the assumptions and choices its authors made in conducting their research.

The limitations on inference presented by a single study is a problem that is not unique to KSY. Indeed, statisticians Andrew Gelman and Eric Loken argue that almost all empirical research is akin to a walk through a “garden with forking paths”: as researchers work through their analysis, they are forced to make choices that send them down one path or another, and these choices influence where they eventually exit the garden (Gelman and Loken 2013). A single coefficient estimate, or even a connected set of estimates from robustness tests, Gelman and Loken argue, may not accurately reflect where other earnest researchers will come out. To make informed inferences, scholars must observe estimates made from other reasonable paths through the garden (Leamer 1985). Such “epistemic” or “model” uncertainty analysis is now being used in a number of fields to improve understanding of important research results (e.g., Durlauf et al. 2016).

In the research reported here, we conduct a model uncertainty analysis of the relationship between materiality-weighted sustainability and stock return. We use Khan, Serafeim, and Yoon (2016) as the basis for our analysis, but also consider other empirical assumptions and model specifications. We confirm the reproducibility of KSY’s reported estimate, but show that it is not representative of estimates made using other valid assumptions. We then evaluate the full set of estimates and use
Bayesian analysis to determine both “best” and aggregate estimates. In contrast to KSY, we find little
evidence that SASB materiality guidance allows the creation of a scale of corporate sustainability that
predicts stock returns. We then evaluate the construction of KSY’s particular measure of material-
sustainability and deduce that its correlation with stock return is probably a statistical artifact. Finally,
we use machine learning to explore whether an alternative materiality-weighting scheme might allow
KLD measures to be used to predict stock returns. We find that sustainability measures associated with
stock return in a training sample are not predictive of returns in a holdout sample, suggesting that
historical associations may not provide a sound basis for materiality guidance.

2. Model Uncertainty Analysis

Model uncertainty analysis\(^2\) is well-suited to the evaluation of the relationship between corporate
sustainability and corporate financial performance, because researchers of the subject have great leeway
in how they may choose to measure or model the connection between sustainability and stock return, so
epistemic uncertainty limits the inferences that can be made from any single report. Yet, practitioners
seek accounting procedures whose application will allow superior returns, and scholars seek reliable
evidence for use in theory building. Consequently, a better understanding of the degree to which
published estimates provide robust evidence has both practical and academic importance.\(^3\)

At its core, model uncertainty analysis represents an alternative approach to empirical research.
In the conventional mode, researchers try their best to make empirical choices that will allow them to
estimate accurately the relationship of interest. They then carefully consider and report the aleatoric
uncertainty of these estimates. For example, they may report how sample variability effects the
probability that a particular interval contains the “true” coefficient estimate, or they may calculate the
frequency with which a random process would generate an estimate larger than the one they have
observed. The fact that these estimates are conditional on particular empirical choices is addressed

\(^2\) Over time, the nomenclature for these approaches has changed, from extreme bounds analysis, to epistemic
uncertainty analysis, and to model uncertainty analysis. We will adopt the latter term.

\(^3\) As far as we can tell, the practice has seldom, if ever, been used by scholars in accounting.
through robustness tests, and if these result in estimates with similar magnitudes and significance, researchers usually assume that epistemic uncertainty is minimal, and propose that general inferences can be made.

In contrast, model uncertainty analysis prioritizes consideration of the uncertainty created by the model selection process. It acknowledges that some epistemic choices are strongly guided by theory or evidence, but others are little more than guesses. For example, there may be good reasons to prioritize a particular statistical analysis (e.g., linear regression), but few reasons to select a particular lag structure, or set of control variables, and so on. As Leamer (1983) points out, such epistemic uncertainty can create a greater variance in coefficient estimates than the more commonly considered aleatoric uncertainty. As a result, he contends, researchers should not “behave as if a given data set admitted a unique inference” (1985: 308). Instead, researchers should advise the reader on how the evidence might be interpreted: “a menu of inferences should be presented and as clear as possible a statement should be made about the assumptions that are necessary to make one inference or another” (Leamer 1985: 312).

How a such a menu of estimates should be evaluated and presented has been an active area of research. Leamer himself initially argued for a rather binary approach: since fragile inferences were “not worth taking seriously” (1985: 308), researchers should determine the boundaries within which their estimates were robust to known epistemic uncertainty. His approach, “extreme bounds analysis”, was slow to be adopted, a problem he blamed on distaste for Bayesian analysis, but which may have had a more pragmatic explanation: extreme bounds analysis often led to pessimistic and unhelpful conclusions. For example, in an early attempt to implement extreme bounds analysis, Levine & Renelt (1992) evaluated various models of economic growth for robust relationships. To their dismay, they discovered that all estimates were “fragile” to reasonable changes in assumptions. Given the difficulties presented by extreme bounds analysis, scholars began to move away from its binary judgements (fragile or robust). For example, after running four million growth models, Sala-i-Martin (1997) claimed that “by looking at the entire distribution” of estimates, he could discern patterns of variables that were connected to
growth. He did not provide, however, guidance to future scholars for how such patterns should be discerned, interpreted, or presented.

Eduardo Ley and his coauthors played a critical role in the development of a formal approach to evaluating a menu of estimates by defining methods for selecting a “best model” or for forming an aggregate estimate from a set of models (Fernández et al. 2001a; Fernández et al. 2001b). But their method requires strong assumptions about the prior probabilities of the models being analyzed, and in practice this usually means that all models are assumed to be (a priori) equally probable. To allow inference by researchers with strong priors, several authors have suggested various graphical approaches. In the spirit of Leamer (1985), each is intended to allow a reader to find and interpret a set of estimates that match their epistemic priors. A variety of scholars have worked on the practical application of these approaches. The closest analog to the methods used in the current study is Durlauf, Navarro, and Rivers (2016).

Below, we follow precedent by first identifying the space of assumptions that will bound our model uncertainty analysis. For each of the implied models within our window, we then calculate coefficient estimates. To aid in the interpretation of these multiple estimates, we follow precedent by displaying them graphically and by analyzing them using Bayesian methods.

2. The Space of Model Uncertainty

All analyses of model uncertainty must bound the set of models to be considered. Leamer (1985) proposes a method based on subjective assessment of the prior probabilities of model specifications, while Madigan and Raftery (1994) suggest an approach based on posterior model probabilities. We follow Leamer’s approach in setting our initial model space, and then follow the spirit of Madigan and Raftery (1994) when conducting our Bayesian analysis.

The goal in selecting a model space is the choice of a set of assumptions that is broad enough to provide a good view of important elements of epistemic uncertainty, but not so broad that it becomes unwieldy or difficult to convey. Practically, a good model space allows uncertain model elements to
vary, but holds fixed those elements that are warranted by theory or evidence. For example, the use of certain statistical estimators is well supported by theory and evidence, so researchers may judge the epistemic uncertainty to be narrow. In contrast, the proper way to measure sustainability performance is relatively unguided by theory or evidence, so scholars may judge the epistemic uncertainty to be wide, and therefore choose to incorporate measures based on alternative assumptions.

We used KSY’s study to set the center of our analysis and defined a model space around it. Practically, we tried to replicate KSY’s study exactly, noting where our empirical choices were constrained by theory/evidence and where they relied on guesswork. Thus, KSY’s analysis is in the middle of what scholars sometimes call “Occam’s Window” (Madigan and Raftery 1994). Using model uncertainty analysis, we can look through this window to get a more complete impression of the relationship between material-sustainability and stock return.

The Center of our Window. KSY’s method involves two broad stages. First, they create a “signal” of each firm’s sustainability by combining and processing materiality guidance from the Sustainability Accounting Standards Board (SASB) and sustainability data from Kinder, Lydenberg and Domini (KLD). As we discuss below, this requires matching SASB and KLD data, placing firms in SASB industries, selecting certain measures for creating scores, processing these scores, and creating a portfolio of firms with top scores. Second, KSY evaluate portfolios by estimating how they would have performed had investors possessed the SASB data necessary to create them.

In our judgment, the most uncertain stages of the process include 1) mapping SASB to KLD data, 2) mapping firms to SASB industries, 3) score processing, 4) sample selection, and 5) specification of the statistical model. For each of these stages, we add into our uncertainty analysis alternative assumptions that we judged, a priori, to be equally probable.

Mapping SASB Materiality to Sustainability Scores. SASB provides guidance about which sustainability topics are likely to be material in a particular industry, but they do not evaluate firms with respect to any of these topics. To create SASB-weighted sustainability scores, researchers must connect SASB topics with sustainability measures from a rating organization such as Kinder, Lydenberg, and
Domini (KLD). KSY report that they found the matching of SASB topics to KLD measures to be a straightforward process that resulted in “minimal” disagreement among evaluators. In contrast, the authors of this study found the matching process to be confusing and ambiguous. Thus, we concluded that our model space should include alternative assumptions with respect to the SASB-KLD connection.

Mapping Firms to SASB Industries. Using SASB materiality data also requires connecting firms to SASB’s sector and industry definitions. KSY do not report how they accomplished this step, and once again we found it to be a difficult and subjective process, leading us to conclude that we should add alternative industry mappings to our model uncertainty space.

Score Processing. KLD create their “signal” of a firm’s sustainability by 1) differencing their raw scores, 2) “orthogonalizing” the differences, and 3) selecting a top quintile of firms. We believe there is little need to consider uncertainty related to the first choice, because it is well excepted that measures based on first-differences help reduce bias from unobserved firm attributes (Angrist and Pischke 2008). We thus do not add variance on this choice (differencing) to our uncertainty space.

KSY’s use of orthogonalization is less well established, and they do not provide any justification, so we include in our model space both orthogonalized and unorthogonalized (raw) measures. Similarly, KSY do not provide a justification of their use of a binary predictor variable, and indeed report conducting a robustness analysis using continuous forms. Thus, we include in our model space both continuous and binary forms.

Sample. The proper sample for analysis is another area of uncertainty, in part because KLD itself changed its process over time. Between 1991 and 2001, KLD created “environmental, social, & governance” (ESG) scores for about 650 companies. They increased their sample, in 2002, to over 1,000, and then to 3,000 companies in 2003. We agree with KSY that ratings in the smaller sample may differ from those in the larger one. We further note that KLD was sold to MSCI in early 2010 and their scoring system adjusted. We think it reasonable to assume that scores in these later years may differ as

---

4 KSY argue that “ESG” and “sustainability” are understood to be synonyms and are used interchangeably in the literature. We agree, and follow their lead.
well. Thus, we contend that these three time periods should be evaluated both separately and as part of a full panel.

SASB’s growth also influences sample choices. At the time of KSY’s publication, SASB had developed materiality guidance for six business sectors, but the coverage has since grown to eleven. We have no strong priors about whether these five new sectors will act similarly to the previous six, so we add this sample variability to our model uncertainty space.

Final specification. KSY estimate the relationship between their materiality measure and stock returns in the 12 months following the release of KLD scores. The true lag structure is unknown, of course, but we think that KSY’s choice makes intuitive sense. After 12 months, new KLD data are available, so it seems reasonable to expect that the effect of a focal data release would be most evident during the following 12-month period. Thus, we chose not to add into our model space variability on the assumed lag structure.

KSY estimate a number of model specifications, but the most comprehensive model includes firm-level attributes, sector fixed effects, and time effects (see KSY Table 6 Panel A). They do not provide a justification for the inclusion of firm-level covariates, and we know of no reason to include these firm attributes – particularly given the use of a differenced predictor variable. KSY also do not provide a justification for the inclusion of sector and time fixed effects, but there is considerable evidence that stock returns do vary by sector and time. We contend, however, that “sectorXtime” fixed effects (i.e., dummy variables obtained for each sector X year) are also justified because stock prices in different sectors may experience different temporal patterns. Thus, we include in our model space specifications including or excluding firm-level attributes and incorporating different types of fixed effects.\(^5\)

---

\(^5\) That is, the inclusion of “sector & time” fixed effects or “sectorXtime” fixed effects.
ANALYTICAL PROCESS

Data Sources and Sample Creation

Our analysis required the combination of several databases. Following KSY, we obtained annual firm-level sustainability data from KLD, and we also obtained monthly and annual financial data from Compustat and CRSP. We then linked these data using a combination of firm identifiers and corporate names. SASB provided us with their materiality scores and some links between firms and industry sectors. Finally, we also requested and obtained from KSY their materiality signal and portfolios.

Mapping SASB materiality to KLD scores. Anyone wishing to conduct research on materiality-weighted ESG measures must link SASB’s topics to KLD measures. To allow variability in assumptions about proper matches, we compiled links from three different research groups.

1) The authors of this paper separately evaluated SASB topics and KLD measures to form links between the two databases. Following KSY’s recommended procedure, we then compared our choices, discussed disagreements, and selected a final set of connections (hereafter AUTHmatch).

2) We tried to replicate KSY’s SASB-KLD mapping using the information they provided in Appendix III of KSY’s 2016 paper. Unfortunately, since KSY reported these links at the sector, rather than the industry level, and SASB has updated its sectors and topics, some judgement was still required. Hereafter, we term these links KSYmatch.

3) A group of faculty scholars at a top research university agreed to share their links with us (hereafter TECHmatch). Following precedent, they had used researcher judgment to form links.

---

6 KLD and Compustat data are so commonly used by researchers that we anticipated that it would be a simple task to link them. Thus, we were surprised to discover that firm-identifiers are not maintained in a consistent way, so that the identifiers cannot always be trusted. KLD keeps firm-identifiers constant, while Compustat updates them to the most recent corporate identifier. This means that if a company (such as Dow) acquires another company (Du Pont), Compustat retroactively replaces records for Du Pont with Dow identifiers! We overcame this problem by using Compustat Snapshot, which does not backdate identifiers, and by checking all matches for the correspondence of company names.
Mapping Firms to SASB Sectors. To use SASB materiality data, SASB sectors and industries must be connected to firms or SIC codes. We were able to form two independent sets of links.

1) The authors of this manuscript formed connections, discussed differences, and converged to a final match.

2) Since the publication of KSY, SASB has linked its industries to some firms. We requested and received these connections and used them to infer how experts at SASB connect their industry classifications to SIC. When SASB had determined an industry link for a firm, we used that link. When they had not, we used our imputed rule about how SASB links SASB industries to SIC.\(^7\)

With our three SASB-KLD links and our two SASB-SIC connections, we formed six alternative SASB-weighted KLD scales of material sustainability.

Score Processing. KSY use materiality-weighted ESG scores to create a “signal” of material sustainability. Their process involves differencing, orthogonalization, and dichotomization. We accept differencing as well-supported by statistical theory and thus did not consider alternatives, but we believe there is much more uncertainty about whether orthogonalization or dichotomization should be used. To capture alternative approaches, we constructed raw and orthogonalized scores in both dichotomized and continuous forms. For simplicity, we refer to all four variables as measuring “material sustainability”, but qualify the details of their construction when discussing them.

Differencing:

\[
S_{\text{ASB}}_{i,t}^m = 1 \text{ if } E_{\text{SG}}_{i,t}^m \text{ is material, else } = 0 \quad \text{Eq. 1.1}
\]

\[
S_{\text{wESG}}_{i,t}^m = S_{\text{ASB}}_{i,t}^m \times E_{\text{SG}}_{i,t}^m \quad \text{Eq. 1.2}
\]

\[
\Delta S_{\text{wESG}}_{i,t} = \sum_{1}^{M} S_{\text{wESG}}_{i,t}^m - \sum_{1}^{M} S_{\text{wESG}}_{i,t-1}^m \quad \text{Eq. 1.3}
\]

\(^7\) To ensure that our imputation did not affect the pattern of our results, we conducted robustness tests using just the direct company to industry mappings obtained by SASB. Doing so reduces the sample, but does not change the pattern of results or our proposed interpretations.
where there are $i$ firms in $t$ years measured with respect to $m$ topics. We followed this approach for each of our different mappings between SASB and ESG. We then formed a set of “raw”, continuous measures of “Material Sustainability” directly from these differenced scores:

$$MSU_{it}^{raw} = \Delta SwESG_{it} \tag{2.1}$$

Orthogonalization:

Following KSY, we also formed scores using an orthogonalization process

$$\forall t: \Delta SwESG_i = \beta\{\Delta [Firm\ Attributes_i]\} + S_t + e_i \tag{3.1}$$

$$\forall t: MSU_{it}^{orth} = e_i \tag{3.2}$$

Following KSY, we specify Firm Attributes as a vector including size, market-to-book ratio, etc. (see KSY (2016), Table 6, Panel A, Models 1 and 2, on page 1717). We estimated the regressions separately for each year $t$. $MSU_{it}^{orth}$ is the orthogonalized and continuous measure of “material sustainability” in ESG performance for firm $i$ in year $t$.

Dichotomization:

KSY dichotomize their continuous measure, separating out those firms rated in the top 20%, relative to $MSU_{it}^{orth}$, to form a portfolio of firms that are “high on material sustainability issues”. Because KSY do not stipulate their process, we assume that this dichotomous variable is made relative to sector $s$ and year $t$.

$$\forall s: MSU_{it}^{top} = 1 \text{ if } MSU_{it}^{orth} \text{ is in the top 20% of } MSU_{it}^{orthog}, \text{ else } MSU_{i}^{top} = 0 \tag{4.1}$$

We sought to extend this dichotomization process to our raw measures ($MSU_{it}^{raw}$), but discovered that for it was not possible to do so, because fewer than 20% of the firms in the sample experienced improve raw scores. Thus, we formed a dichotomous variable separating out firms whose scores did improve from those that did not.

$$\forall s: MSU_{it}^{imp} = 1 \text{ if } \Delta SwESG_{it} > 0, \text{ else: } MSU_{it}^{imp} = 0 \tag{4.2}$$

---

8 Note, we assign a negative sign to KLD concerns, so that the elimination of a concern results in an improved score.

9 Below, we discuss that we revisited this assumption when interpreting our results.
In total, for each of our mappings, we calculate four different measures of material sustainability.

<table>
<thead>
<tr>
<th>Equation</th>
<th>Term</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>$MSUS_{it}^{raw}$</td>
<td>MatSust(raw, continuous)</td>
</tr>
<tr>
<td>4.2</td>
<td>$MSUS_{it}^{imp}$</td>
<td>MatSust(raw, improved)</td>
</tr>
<tr>
<td>3.2</td>
<td>$MSUS_{it}^{orth}$</td>
<td>MatSust(orth, continuous)</td>
</tr>
<tr>
<td>4.2</td>
<td>$MSUS_{it}^{top}$</td>
<td>MatSust(orth, topquint)</td>
</tr>
</tbody>
</table>

**Sample.** KSY conducted their analysis when KLD data was available from 1991-2013 and when SASB had evaluated six sectors. As far as we could determine, it is unknown whether materiality relationships were consistent across this entire period. We do know that KLD’s method or sample changed in 2002 and 2010. Thus, we decided to evaluate samples from three partial periods (1991-2001, 2002-2009, 2010-2013), as well as for the full KSY time period (1991-2013). Since KSY’s publication, additional data (to 2016) have become available, but we decided to use these latter years as a holdout sample for testing the implications of a machine learning analysis.

Each period was considered for samples including six SASB sectors considered by KSY as well as for the eleven sectors now available.

**Final specification.** KSY analyze the impact of material ESG measures as predictors of stock returns over the 12 months following disclosure of annual KLD data. They theorize that an investor could use the reported data to create a portfolio of investments, and they seek to see how this portfolio would perform. Their most unconstrained analysis uses their full panel and the following specification:

$$RET_{i,y,m} = \beta \cdot MSUS_{i,y-1} + \theta \{ \Delta [Firm Attributes_{i,y-1}] \} + S_{iy} + D_{iy} + e_{i,y,m} \quad \text{Eq. 5.1}$$

where there are $i$ firms, in $y$ years, and $m$ months. The firm-level variables on the right-hand side of the equation are lagged one year. Following KSY, we estimate returns for the 12 months after KLD reveals its annual data. The vector of firm attributes includes the Lagged Annual Return, Size (the logarithm of the market capitalization of the firm at the end of the previous month), Book-to-Market, Turnover, Return on Equity, Analyst Coverage, Leverage, R&D spending, Advertising Intensity, SG&A and Capital Expenditures. We followed KSY’s procedure in creating these variables (see KSY’s Model 2 in
Table 6 Panel A. For simplicity, we did not include a dummy for immateriality. The firm’s sector is captured by a dummy variable $S$, and fixed effects for year $D$.

As mentioned earlier, we believe a more robust specification would allow the time dummies to vary within each sector, so we also specify:

$$\text{RET}_{i,y,m} = \beta \cdot MSUS_{i,y-1} + \theta \{\Delta [\text{Firm Attributes}_{i,y-1}]\} + \theta_{i,s,y} + e_{i,y,m}$$  

Eq. 5.2

$\theta_{i,s,y}$ are unique dummy variables for each sector $s$ in each year $y$.

In addition, the justification for including the vector of control variables in all the models is not obvious. Thus, we also run the equations 5.1 and 5.2 without control variables:

$$\text{RET}_{i,y,m} = \beta \cdot MSUS_{i,y-1} + S_{i,y} + D_{i,y} + e_{i,y,m}$$  

Eq. 5.3

$$\text{RET}_{i,y,m} = \beta \cdot MSUS_{i,y-1} + \theta_{i,s,y} + e_{i,y,m}$$  

Eq. 5.4

All specifications are analyzed using all four forms of the sustainability variable.

**KSY’s actual measure.** KSY generously offered to assist us in our research by providing both their estimated “signal” of materiality and the portfolio of firms with top performance that they created from it. We term these measures $MSUS_{it}^{KSY}$ and $MSUS_{it}^{KSY, top}$ respectively. They did not provide us with their raw data, but we were able to estimate it by rounding their continuous “signal” to the nearest integer\textsuperscript{10}. Once we obtained our estimated raw score ($MSUS_{it}^{k SY, raw}$), we followed equations 4.1 and 4.2 to create $MSUS_{it}^{k SY, top}$ and $MSUS_{it}^{k SY, imp}$

**Total Model Space.** Figure 1 shows the space of models using the various combinations described above. In total, the space includes 768 possible models using variables constructed with our measures of “material sustainability”, and 32 models using KSY’s actual data or close variants. We contend that these 800 models capture some of the reasonable choices that researchers might make were they to conduct an independent analysis of the relationship or use these data in making and tracking investment

---

\textsuperscript{10} We checked the fidelity of this process by performing the same routine on our orthogonalized scores and then comparing our estimates to the true non-orthogonalized scores ($MSUS_{it}^{raw}$). We found a strong correlation (R=0.95) between the estimated raw score and the true raw score, and thus inferred that this process could be used to estimate KSY’s raw data.
portfolios. Thus, we assert, the informative value of KSY’s estimate for scholars and practitioners can be better assessed when evaluated within this larger set of results.

Descriptive Stats. Table 1 presents the descriptive statistics of the firms used in our analysis, both for the six SASB sectors in KSY’s analysis and for all eleven sectors now available. We followed KSY in imputing R&D when it was missing. We also removed from the sample those firms, mostly in biotech, where R&D intensity exceeded 100%.

Table 2 shows the descriptive statistics and the correlations among our various measures of the orthogonalized form of our variables. The measure KSY used in their analysis (provided to us by KSY) is reported in the first row (since it is only defined for the original six sectors, the sample is reduced to 11,610). The real KSY measure and our replications using alternative mappings are correlated within a range between 0.57 and 0.34. Most of the variance appears to come from differences in SASB to KLD mappings. Not surprisingly, the links we formed based on KSY’s description of their SASB-KLD match (in Appendix I of KSY) are most strongly associated with the measure that KSY shared with us (see rows 2 & 3 in Table 2).

Graphing

The best way to present estimates from a model space remains an area of debate. Simonsohn, Simons & Nelson (2019) propose that estimates can be connected to modeling assumptions by binary codes. King, Goldfarb, and Simcoe (2020) suggest that the data be presented in a manner that is useful to readers with differing priors about which assumptions are best, and thus they propose segmenting graphs by various assumptions so that users can assess estimates contingent on a particular assumption. Consistently, Durlauf et al. (2016) have used graphs of estimates sorted by marginal coefficient estimates. Because our model space is large, we found the method proposed by Simonsohn et al. (2019) to be very
unwieldy. Thus, we decided to follow the approach of sorting and segmenting estimates prior to graphing.

In all graphs, we show the marginal effect of the predictor, estimated at the mean (labeled as “b”), as well as the 95% confidence interval for this estimate. In addition, we mark our replication of KSY’s actual model with a red line.

**Bayesian Model Selection and Averaging**

A formal synthesis of the total collection of results can be created using Bayesian analysis. This approach allows a more objective and transparent analysis, but requires the use of strong assumptions about the prior probabilities of particular models: if we assume that each model has an equal probability of matching the true data generation process, we can use a model’s “posterior probability” to select a best estimate or to create a probability-weighted average estimate (Claeskens and Hjort 2008). By Bayes Rule, the posterior probability of any Model $i$, conditional on the observed evidence $D$, can be calculated as:

$$P(M_i | D) = \frac{P(D|M_i)P(M_i)}{\sum_{i=1}^{n} (P(D|M_i)P(M_i))} \quad \text{Eq. 6.1}$$

where there are 1 to $n$ possible models of data. $P(D|M_i)$ is the probability of getting the evidence $D$ given any Model $i$. And $P(M_i)$ is the prior probability for each model. However, if we assume that all of the models have equal prior probability then $P(M_i) = c$, and $\sum_{i=1}^{n} c = 1$, we can then simplify the equation and eliminate the priors. This means that equation 6.1 becomes:

$$P(M_i | D) = \frac{P(D|M_i)}{\sum_{i=1}^{n} P(D|M_i)} \quad \text{Eq. 6.2}$$

We still need to calculate $P(D|M_i)$. Following Chib (1995), we can take advantages of the fact that the well-known Bayesian Information Criteria (BIC) can be used to calculate the relative posterior probability of two models.
\[
P(\mathbb{M}_i \mid \mathbb{D}) = \frac{P(\mathbb{D} \mid \mathbb{M}_i)}{P(\mathbb{D} \mid \mathbb{M}_r)} = \frac{e^{-\text{BIC}_i/2}}{e^{-\text{BIC}_r/2}} \quad \text{Eq. 6.3}
\]

This allows all models to be compared relative to a reference model \( r \). If we compare all models in set \( A \) to this reference model, then

\[
\sum_{i=1}^{n} P(\mathbb{M}_i \mid \mathbb{D}) = 1 \quad \text{Eq. 6.4}
\]

Substituting into Equation 6.3, we can calculate the relative probability weight for each model \( i \).

\[
W_i = P(\mathbb{M}_i \mid \mathbb{D}, A) = \frac{e^{-\text{BIC}_i/2}}{e^{-\text{BIC}_r/2}} \quad \text{Eq. 6.5}
\]

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

\[
\mathbb{E}(\bar{B} \mid \mathbb{D}) = \sum_{i=1}^{n} B_i W_i \quad \text{Eq. 6.6}
\]

Assuming that the estimate variance \( s^2 \) of each model is independent\(^{11}\), we can calculate the variance of this pooled coefficient estimate as:

\[
\mathbb{E}(\bar{s} \mid \mathbb{D}) = \sqrt{\{\sum_{i=1}^{n} s_i^2 W_i\}} \quad \text{Eq. 6.7}
\]

**RESULTS**

Figure 2 shows estimates from all of the models, ordered by the marginal effect of a unit change in the independent variable. For binary variables, this is the marginal effect of the positive case, and for continuous variables it is the effect of a unit increase in score. The grey regions show the 95% confidence interval for the estimate. Over the full collection of models, 54% of the coefficient estimates are positive and 46% negative; 80% of the estimates imply annualized returns less than +/-2%/year and 60% of the results imply annualized returns less than +/-1% per year. For 4.0% of the models, we estimate a positive coefficient with a 95% confidence interval not inclusive of zero; 4.5% of models

\(^{11}\) A more accurate analysis would employ Monte Carlo techniques to construct the weighted sample variance.
result in coefficient estimates that are negative and have confidence intervals that do not include zero.\textsuperscript{12} We suggest that if one approaches the analysis with weak priors about the best model specification, then the full pattern of results shown in Figure 2 should be interpreted as providing little evidence of a reliably positive or negative relationship between material-sustainability and stock return.

Yet, as shown in Figure 2 (see red line), we also confirm KSY’s reported estimate of a strong and positive relationship when we use their top quintile measure and their specified model \((b = 0.51\% \text{ per month})\). This differs only slightly from KSY’s report of \(b = 0.52\% \text{ per month} \) (Khan et al. 2016). Thus, we show that their result is reproducible, and confirm that their selected portfolio has a high return. We note, however, that their estimate is far from the median of the distribution; indeed, it is larger than the estimates from 98\% of the models.

\textsuperscript{12} Determination of statistical significance using frequentist methodologies requires that measures and models be fixed prior to estimation. Clearly this is not the case here, and estimates selected ex-post will have biased means and standard errors.
(AUTmatch); the largest of which suggest a 10% loss per year. Estimates using TECH’s links result in more balanced and modest results, with 3% loss or gain per year.

As shown in Figure 3b, we estimate 32 models using KSY’s actual measures of sustainability. These include both binary and continuous forms of their variable, as well as different statistical model specifications. We have marked in red our replication of the estimate reported by KSY (their Table 6, Panel B, Model 2). Of the 32 estimates, two are strongly positive (5 & 8% per year). Both use the binary (top quintile) form of their orthogonalized measure of sustainability, and differ only in the sample used (the larger is estimated using the 2002-2009 time frame, and the smaller the 1991-2013 time frame). Of the remaining 30 estimates, 10 are positive, and 20 are negative (the strongest implies a 3% loss per year). To us, this analysis suggests that the inferential value of KSY’s estimate depends on users’ belief that material-sustainability is best measured by KSY’s process of orthogonalization and dichotomization.

-----------------------------------
Insert Figure 3 about here
-----------------------------------

The strength of KSY’s reported estimate might imply that some forms of our material-sustainability measures are more strongly associated with stock returns. To explore this possibility, we sort our models by the processing methods and functional form used in the measure of material-sustainability. This includes two forms of the orthogonalized measure (MatSust(orth, continuous) & MatSust(orth, topquint)), and two forms of the raw differenced scores (MatSust(raw, continuous) & MatSust(raw improved)). Within each group, we again order the data by marginal beta.

Figure 4a, shows that for MatSust(orth, continuous) and MatSust(orth, topquint) 60% and 40% of the models respectively result in a positive coefficient estimate. None of the positive estimates have a confidence interval not inclusive of zero. A few of the negative coefficients (5% for MatSust(orth continuous), and 4% for MatSust(orth topquint)) have confidence intervals that do not include zero. For data using the raw form of the measure, 54% of the models using MatSust(raw continuous) and 65% of
the models using the binary measure, MatSust(raw improved) result in a positive coefficient estimate. Not surprisingly, the sparseness of the binary form results in a larger dispersion of estimates.

Figure 4b shows that only the quintile form of KSY’s measure of material-sustainability results in a positive and “significant” estimate. Indeed, a model specification using the continuous form of their variable, but otherwise identical to the one they reported in their publication, results in a slightly negative coefficient estimate (marked in red as well). This suggests that an inference in a positive association between material-sustainability and stock return requires a prior belief that the binary form of their variable, and not its continuous form, best measures a firm’s material-sustainability.

To further explore our collection of estimates, we sorted our results by time periods. As shown in Figure 5a, estimates for the full sample have the lowest standard errors, and the 1991-2001 sample the highest. This is unsurprising since these are, respectively, the samples with the largest and smallest number of observations.

As shown in Figure 5b, the largest estimate for measures of Material-sustainability using KSY’s real data comes from the 2002-2009 sample (Model #9), and the second largest from the full time period (Model #25 marked in red; the other red mark refers to Model #32 with the continuous form). This pattern of results might suggest that the measure was more predictive in the middle ‘00’s, and we explore this conjecture later. Overall, however, Figure 5b again suggests that inference in a positive association between material-sustainability and stock return requires a prior belief that KSY’s binary variable best measures a firm’s material-sustainability.

To further explore whether temporal effects might influence our coefficient estimates, we separate our models by whether we assume that temporal patterns in stock returns are independent across sectors, or we allow each sector to have a different temporal pattern.
Figure 6a shows results for the variables we created separated by choice of fixed effects and estimated beta. Figure 6b shows the results for measures using KSY’s data, similarly sorted. To our eye, the two sets of estimates in Figure 6a appear similar.

Figure 6b shows that KSY’s reported estimate is contingent on the use of sector invariant time dummy. That is, had KSY assumed that temporal effects could vary by sector, and specified their effects accordingly, they would not have found a reported positive and significant association between material-sustainability and stock return.\footnote{Following KSY, we also conducted some additional analysis. First, we evaluated our model window using firm fixed effects: this reduces the number of models that can be estimated, but does not change the pattern of results reported above. Second we replicated the entire analysis using the Return on Sales (ROS) rather than stock return. Because these data are reported annually, we must move from a monthly analysis to a yearly analysis. Other than this change, we can estimate the same 800 models discussed above. In Appendix 1, we provide identical graphs for these models. We note that we are unable to replicate exactly KSY’s reported estimate, but we do calculate a coefficient that is similar in sign and significance (See Appendix 1, Figure A1). With this exception, it is our judgment that the pattern of results reported in Appendix 1 conforms to those discussed above, so we do not further describe them here.}

In summary, the preponderance of our models result in estimates that are much smaller than KSY’s and that are almost equally likely to be positive or negative. Separating the analysis by important empirical assumptions does not deliver a set of models with estimates that are reliably positive or negative. Some models do result in strong estimates, but these should be given additional weight in inference only if readers have strong priors that these model specifications best match the true data generation process.

For those readers who, like ourselves, are uncertain about which of our 800 models is best, we infer from the above evidence that it is rational to increase belief that the relationship between material-sustainability and stock return is small and of uncertain sign. We now turn to a more formal analysis of this inference.
Bayesian Analysis

Up to this point, we have looked at these data from multiple perspectives and tried to arise at a best explanation for the patterns we observe. This pragmatic process allows readers with strong priors to select particular models and update their thinking accordingly. For readers who have weak priors, we now provide a Bayesian analysis. As discussed in our methods section, if we assume that readers do not favor a particular set of assumptions, each model’s fit with the data can be used to assess its relative posterior probability. This allows us to select certain models for special attention and to calculate a probability-weighted synthesis of the entire set of estimates.

In Table 3, we show both “probability-weighted” aggregate estimates of coefficients and individual estimates from those models with the highest posterior probability. Since we know from our graphical analysis that the functional form of the predictor variable influences our estimates, we chose to form separate “Occam’s windows” for each form of the predictor variable. To allow inclusion of KSY’s real data in these windows, we are further forced to subdivide each window by samples using only the original six SASB sectors (where inclusion of KSY’s data is possible) and all 11 sectors (where it is not). The top part of the table reports information about those models selected as having the highest posterior probability, and the bottom part reports probability-weighted aggregate estimates.

Beginning with the top part of Table 3, we don’t see any evidence that a particular set of SASB-KLD links led to more predictive models, because the best models used varying links (Authors’ four times, KSY three, and TECH once). In contrast, all of the models used the connections between firm and industry created by SASB, and statistical specifications using both sector and year dummy variables. KSY’s actual measure and model stands out in its “Occam’s Window”. It is selected as fitting these data most closely, and it results in a very strong positive coefficient estimate. The best models within the other windows result in estimates with varying signs (four positive and 3 negative), but no coefficient

---

14 We take comfort in the selection of our SASB-KLD mapping because it suggests that our efforts at matching resulted in measures with better predictive power.
estimate approaches that obtained from KSY’s model\textsuperscript{15}. Also, in the three other windows where KSY’s real measure might be selected, it is never chosen as resulting in the most probable model. To us, this again suggests that just one form of their measure ($MSUS_{it}^{ksy, top}$) is particularly predictive.

Moving to the top part of Figure 3, we can consider the probability-weighted aggregate estimates. As is common in Bayesian model averaging, the probability estimates are sharply skewed, so a few models strongly influence these aggregate estimates. As a result, these aggregate estimates differ only moderately from those obtained by considering the best models alone. This is particularly true for the first window, where the aggregate estimate closely matches the one reported by KSY.

\begin{center}
Insert Table 3 about here
\end{center}

Thus, our Bayesian analysis suggests that inference depends on the window of analysis used. For one particular window, a reader with diffuse priors might infer that material-sustainability (as captured by KSY’s measure) is strongly and positively associated with stock return. For all other windows, however, such a reader might reasonably infer that the relationship between material-sustainability and stock return is small, absent, or even negative. Thus, the usefulness of KSY’s reported estimate, depends on belief that their particular method for calculating material-sustainability accurately captures the construct. We now turn to this question.

**Exploring KSY’s estimate**

As discussed above, we confirm that one form of KSY’s measure of Material-sustainability ($MSUS_{it}^{KSy, top}$) is strongly and positively associated with stock returns, but we also show that this result is not representative of broader patterns. Indeed, it is contingent on the use of 1) a dichotomized

\textsuperscript{15} In some cases, the coefficients for the selected models have T-statistics that exceed traditional cutoffs for significance, but they should not be interpreted as demonstrating “significance”, because such an interpretation would require pre- specification of a model and sampling plan. When models are selected ex-post for “fit”, coefficient estimates may be inflated and standard errors deflated.
variable based on the top quintile of orthogonalized scores, 2) a sample time period inclusive of 2002-2009, and 3) the assumption that industry effects do not differ across time. Might these contingencies suggest that KSY’s orthogonalization/dichotomization process resulted in a conflation between their measure and some temporal or industry effect?

From our experience constructing the measures, we know that most firms experience little change in their KLD’s scores from year to year. This means that KSY’s basic measure of material-sustainability is often zero. In fact, across the sample, KSY’s raw (non-orthogonalized) signal of sustainability is negative about 16% of the time, zero 69% of the time, and positive just 15% of the time. In some years, fully 90% of the firms have a signal with a zero value! As a result, their measure of a top quintile of firms with “high” material sustainability, must include both 1) firms with improved ESG scores, and 2) firms with unchanged ESG scores. Consider, for example, KSY’s measure of Material-sustainability (orthogonalized) in 2006. As shown in Figure 7a, more than 80% of the firms had unchanged scores that year. To form the top quintile of material-sustainability, KSY combined 5% of the sample that actually experienced an improved score (labeled MEASURED), with 15% that did not change, but had a slightly higher regression residual (labeled IMPUTED).

We note that in combining these two groups into one binary measure, KSY implicitly assumed that the two would have the same association with future stock returns. Since we have no strong prior that this is true, we decided to relax this assumption and estimate separate coefficients for the two parts. Model 1 in Table 4 reproduces KSY’s estimate, while Model 2 breaks their measure into its two component parts. Firms with improved material-sustainability scores (MEASURED) experienced slightly higher stock returns (0.076%/month or 1%/year, but IMPUTED members experienced much higher returns (0.99%/month or 12%/year)!

We know from our model uncertainty analysis that coefficient estimates for KSY’s measure vary across time-periods, so we next explored three different periods. As shown in Models 3, 4, & 5, the coefficient for KSY’s actual measure is positive and significant only in the 2002-2009 time frame. We also know from our model uncertainty analysis that KSY’s estimate is not robust to the inclusion of
sectorXyear. In Model 6, we confirm that relaxing the assumption of sector-independent time effects reduces the coefficient and significance for KSY’s measure. As shown in Model 7, the inclusion of such fixed effects reduce the coefficient estimate for the IMPUTED part, not for the MEASURED part. We will discuss model 8 later.

KSY also conduct, as part of their analysis, a Fama-French analysis in which they compared portfolio made of top and bottom firms. To confirm our interpretation of results in Table 4, we replicated this analysis and then again split the portfolios in to IMPUTED and MEASURED parts (see Appendix 2). We confirm that the IMPUTED part of their portfolio, and not the measured part, causes their reported result.

The importance of the IMPUTED part of KSY’s measure and its contingency on sector and time-period, suggested to us that it may be conflated with an industry or time event. To explore this hypothesis, we evaluated the sector affiliation of the firms categorized in KSY’s top quintile, and discovered that the IMPUTED part of their measure was not evenly distributed across sectors. From this, we inferred that KSY did not select their top quintile conditional on the sector (as we had assumed). Furthermore, in any given year, many of the IMPUTED firms came from a single industry, raising the possibility that KSY’s measure was conflated with an industry-year effect.

Insert Table 4 about here

From further analysis, we discovered that during the years 2002-2009, firms in the extractive-material sector formed a disproportionate part of KSY’s top quintile. Most of these firms experienced no change in performance, but were IMPUTED to be in the top quintile based on the residual of the orthogonalization process. In some years, 80% or more of the firms in the extractive materials sector were included in this manner, and these firms could comprise as much as 60% of KSY’s top quintile! Consider, for example, the graph of KSY’s Material-sustainability for the extractive materials sector in 2006 (Figure 7b). Even though only a small number of firms actually had improved scores (5), fully
80% of the sector (109 firms) were included in KSY’s top quintile, including ExxonMobil. These data were then associated with stock returns for April 2007 to March 2008. Over that period, prices for oil and some other resources climbed by more than 40%. Prices fell, of course, as the financial crisis took hold, but as luck would have it, KSY’s top quintile for the following year included only 33 firms from the extractive-materials sector. To confirm our hypothesis that KSY’s measure was conflated with a temporal effect in extractive-materials, we removed the sector from our sample, and reran KSY’s original model. As shown in Table 4, model 8, the estimated coefficient for KSY’s top quintile falls from 0.541%/month to .075%/month (s = .109), equivalent to a fall from more than 6% per year to less than 1%

We believe that our analysis strongly implies that KSY’s reported coefficient estimate is a statistical artifact. It is highly sensitive to changes in empirical assumptions, including the use of continuous forms of the same variable. It depends on the assumption that firms that experienced no improvement in material-sustainability will have the same regression coefficient as those that actually change. Finally, it lacks face validity as a measure of performance, because it requires that a large majority of the extractive-materials sector be judged to be materially sustainable just as the world became focused on the problems caused by excessive resource use – including the burning of fossil fuels and the resulting buildup of greenhouse gasses.

-----------------------------
Insert Figure 7 about here
-----------------------------

Can a predictive measure of material-sustainability be built using KLD data?

The totality of the evidence so far suggests to us that there is little reason to believe that existing SASB materiality scores, when combined with KLD data, provide a reliable predictor of stock returns. It does not imply that a predictive score of material-sustainability is infeasible. SASB’s materiality choices are based on the experience of particular experts, and these experts may have failed to identify which elements of sustainability are truly material. It seems reasonable to conjecture that further analysis by other experts might lead to a predictive sustainability scale. Given the tens of trillions of dollars now
invested in funds using corporate sustainability data, we would be remiss if we did not investigate the feasibility of building a predictive materiality-weighted scale.

To determine the feasibility of a predictive measure, we turned to machine-learning (ML) to speed the process of finding elements of KLD data that might materially influence returns. To conduct our ML analysis, we assume that human experts judge sustainability items to be material if they have observed them to be associated with stock movements. For example, if historical evidence suggests that, for firms in service industries, improved community relations are associated with subsequent higher returns, then we should conclude that community relations are material for this sector.

To search for historically predictive relationships, we conducted LASSO analysis of the association between every KLD measure and financial return. LASSO regressions select predictive variables and regularize them using a loss function. Consistent with Eccles and Serafeim’s (2013) contention that materiality is sector dependent, we searched separately in each sector for predictive correlates.

\[
\text{RET}_{i,y,m} = \beta \times \Delta ESG_{i,j,y-1} + \theta \{ \Delta [Firm Attributes_{i,y-1}] \} + D_{y,m} + e_{i,y,m} \quad \text{Eq. 5.1}
\]

where there are \( i \) firms, measured with \( j \) ESG attributes, in \( y \) years, and \( m \) months. For each ESG measure \( j \), we evaluate whether adding the variable exceeds the cost of inclusion. Once we have identified a possible set of predictive variables, we form them into scales. Where the estimated coefficient \( B \) is positive, increases in scores are added, where the coefficient estimate is negative, increases in scores are subtracted.

We first specified a loss function that maximized the predictive accuracy of the model, but found that this resulted in a sparse scale, because for several sectors, no KLD measures were selected as being predictive! To ensure a scale that included measures for all sectors, we also specified a loss function that ensured that we had at least 5 scale items per sector. We also discovered that many of the estimated coefficients did not match the anticipated direction. That is, ML predicted a KLD “strength” would precede a lower stock price or a “concern” a higher one. We chose not to constrain our scale to those
measures matching the assumed direction; had we done so, we would have had to remove more than half of the scale items.

In total, we created and tested four different scales (from two LASSO loss functions trained on two different time periods\textsuperscript{16}). We then tested these models on a holdout sample from 2014-2016. Figure 8 shows the coefficient estimates for material-sustainability from 144 models, trained on two sets of time data. Of all models tested, 19\% result in estimates of a positive association between material-sustainability and stock return in the subsequent 12 months, but none of these have a confidence interval that excludes zero. A predominance of negative coefficients is estimated, but 97\% are sufficiently uncertain that the confidence interval includes zero. We infer from this analysis that building a measure of material-sustainability, using KLD data, that predicts stock return will be difficult and may require going beyond mere observation of historical patterns of association.

\textsuperscript{16} Unsure whether more recent years’ would be more predictive, we trained the ML materiality identified using a longer sample (2003-2013) and a shorter sample (2010-2013).

Conclusion

In the research presented above, we continue the exploration of the link between corporate sustainability and stock return. We draw inspiration from Khan, Serafeim, and Yoon (2016), but assume that their estimates are contingent on the assumptions they employed in their analysis. Following previous research, we attempt to facilitate reliable inference by presenting their estimates in context with those made using other reasonable modeling assumptions. We define a space of 800 models, and calculate coefficient estimates for each one. For readers with strong beliefs about which models are most informative, we dissect and present our results contingent on various assumptions or beliefs. For readers who are less certain about the “true” model, we select models for particular attention based on their posterior probability, and we create probability-weighted aggregate estimates.
In our opinion, it is reasonable to infer from the totality of our evidence that materiality guidance from SASB does not allow the creation of a scale of KLD data that reliably predicts stock returns. We confirm the estimates reported by Khan, Serafeim, and Yoon (2016), but we show that they are neither representative of the totality of the estimates nor robust to changes in assumptions. Furthermore, we show that their measure confounds firm-level and sector-level effects. Specifically, we show that their portfolios of “high material sustainability” over-represent firms in the extractive materials sector during a historic resource boom, and that without these firms, KSY would estimate a small and insignificant effect.

Looking beyond SASB’s existing materiality measures, we use machine learning to simulate how experts might use past associations to create future materiality-weighted scales. We find that the created scales do not predict positive stock returns in a holdout sample. Thus, we extrapolate that scales created by selecting historically “material” measures from KLD ratings are unlikely to provide reliable predictors of stock return.

Our results suggest a solution to a puzzling conflict between KSY’s findings and the predictions of theories of weakly efficient markets (e.g., Keim and Stambaugh 1986). Those theories suggest that information can be used to predict superior return, but it must be closely held by a few investors (e.g., Basu 1983; Graham 1965; Fama 1970). Yet KSY’s results suggest that enormous returns could be had by scaling data that were readily available to investment managers. To put the enormity of KSY’s estimate in perspective, consider the effect of a historic sustainability event that occurred just before the start of KLD’s panel. On March 24, 1989, the Exxon Valdez struck a reef in Alaska and thereby caused one of the worst environmental accidents in history. Exxon’s stock price dropped 3.7% in the subsequent 12 months, but this is a smaller effect than KSY’s estimate for membership in the top quintile of their scaled measure, and they contend that this powerful association lasted for over 20 years! It seems improbable to us that such valuable information would have remained undiscovered for so long, and more probable that, as our analysis suggests, using SASB’s materiality guidance does not allow the creation of a scale of KLD’s measures that reliably predict stock returns.
Finally, we believe that our research demonstrates the importance of conducting model uncertainty analysis. Scholars have long argued that caution must be exercised in making inferences from a single empirical study, even one as well-designed as KSY, if model uncertainty has not been considered. As currently practiced, empirical analysis is often like a walk through a garden with forking paths. A few paths may lead to intriguing estimates, but these results cannot be understood without also observing estimates from equally reasonable paths. KSY’s assumptions and choices led to an estimate that seemed to suggest that portfolios made using materiality-weighted ESG measures would out-perform the market. Our model uncertainty analysis suggests that such a portfolio would provide no advantage.
References


Gelman, A., and E. Loken. 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.*


Figure 1: Final Space for Model Uncertainty Analysis

<table>
<thead>
<tr>
<th>KLD-SASB mapping</th>
<th>SASB to industry mapping</th>
<th>Measure Processing</th>
<th>Sample</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUTHmatch</td>
<td>AUTHOR</td>
<td>Orthogonized</td>
<td>Time Period</td>
<td>Fixed Effects</td>
</tr>
<tr>
<td>TECHmatch</td>
<td>SASB</td>
<td>• Top: quintile</td>
<td>• 1991-2013</td>
<td>• Year AND Sector</td>
</tr>
<tr>
<td>KSYmatch</td>
<td>KSY</td>
<td>• Continuous</td>
<td>• 1991-2001</td>
<td>• Year X Sector</td>
</tr>
<tr>
<td>(replicated*)</td>
<td>(actual**)</td>
<td></td>
<td>• 2002-2009</td>
<td></td>
</tr>
<tr>
<td>KSY (actual**)</td>
<td></td>
<td></td>
<td>• 2010-2013</td>
<td></td>
</tr>
</tbody>
</table>

*We attempt to replicate part of KSY’s mapping based on information in their publication.

**Because we have only KSY’s actual measure, we cannot observe the mappings used to create it. Thus, its individual elements cannot be combined with other mapping assumptions.

*Previous year’s annual returns, size, BTM, turnover, roe, analyst coverage, R&D, advertising intensity, SG&A, capital expenditure, leverage.
Estimates for the effect on stock return for all models and measures of material sustainability. Units are percent per month, so a value of one (1.0) represents a return of 0.01 per month for each of 12 months. Over the full collection of models, 54% of the coefficient estimates are positive and 46% negative; 80% of the estimates imply annualized returns less than +/-2%/year and 60% of the results imply annualized returns less than +/- 1% per year. 4.0% of the models result an estimate of a positive coefficient with a 95% confidence interval not inclusive of zero; 4.5% of models result an estimate of a negative coefficient with a 95% confidence interval not inclusive of zero.

A red line shows our replication of KSY’s estimate using their binary measure, MatSust(ortho, topquint), but our sample and other data.
Figure 3: Marginal Effects at Average by Source of SASB-ESG Matching

Figure 3a: All models using replicated data

<table>
<thead>
<tr>
<th>Source</th>
<th>Pos.</th>
<th>Neg.</th>
<th>CI n/i 0*</th>
</tr>
</thead>
<tbody>
<tr>
<td>KSYmatch</td>
<td>68%</td>
<td>32%</td>
<td>8%</td>
</tr>
<tr>
<td>AUTmatch</td>
<td>30%</td>
<td>60%</td>
<td>0%</td>
</tr>
<tr>
<td>TECHmatch</td>
<td>67%</td>
<td>3%</td>
<td>10%</td>
</tr>
</tbody>
</table>

CI n/i 0* confidence interval not inclusive of zero

Units are percent per month.

Figure 3b: All models using KSY’s actual materiality signal
Figure 4: Marginal Effects at Average by Form of Materiality Measure

Figure 4a: All models using replicated materiality signals

<table>
<thead>
<tr>
<th>Pos.</th>
<th>60%</th>
<th>40%</th>
<th>54%</th>
<th>65%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI n/i 0</td>
<td>0</td>
<td>0</td>
<td>4%</td>
<td>11%</td>
</tr>
<tr>
<td>Neg.</td>
<td>40%</td>
<td>60%</td>
<td>36%</td>
<td>35%</td>
</tr>
<tr>
<td>CI n/i 0</td>
<td>5%</td>
<td>4%</td>
<td>5%</td>
<td>5%</td>
</tr>
</tbody>
</table>

Figure 4b: All models using KSY’s actual materiality signal

Units are percent per month.
Figure 5: Marginal Effects at Average by Time Period

Figure 5a: All models using replicated materiality signals

Figure 5b: All models using KSY’s actual materiality signal

Units are percent per month.
Figure 6: Marginal Effects at Average by Sector and Time FE

Figure 6a: All models using replicated materiality signals

<table>
<thead>
<tr>
<th></th>
<th>KSYmatch</th>
<th>AUTmatch</th>
<th>TECHmatch</th>
<th>KSYmatch</th>
<th>AUTmatch</th>
<th>TECHmatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos</td>
<td>67%</td>
<td>69%</td>
<td>67%</td>
<td>66%</td>
<td>64%</td>
<td>64%</td>
</tr>
<tr>
<td>CI n/i</td>
<td>7%</td>
<td>9%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
</tr>
<tr>
<td>Neg</td>
<td>33%</td>
<td>31%</td>
<td>33%</td>
<td>36%</td>
<td>36%</td>
<td>36%</td>
</tr>
<tr>
<td>CI n/i</td>
<td>0%</td>
<td>3%</td>
<td>0%</td>
<td>11%</td>
<td>4%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Units are percent per month.

Figure 6b: All models using KSY’s actual materiality signal

Models

-2 -1 0 1 2

1 768

Models

-1 -.5 0 .5 1

1 32

Units are percent per month.
Figure 7: KSY’s Top Quintile Includes IMPUTED and MEASURED Parts

Figure 7a: All Firms in 2006

Figure 7b: KSY’s Top Quintile Oversamples on Extractive-Materials (2006)
Figure 8: Estimates for “Materiality” Scales Found Using Machine Learning

Units are percent per month.
Table 1 – Descriptive Statistics

<table>
<thead>
<tr>
<th>Variables</th>
<th>The 6 SASB sectors used by KSY</th>
<th>All the 11 SASB sectors sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
</tr>
<tr>
<td>Previous year’s annual returns</td>
<td>1.15</td>
<td>0.58</td>
</tr>
<tr>
<td>Size</td>
<td>21.19</td>
<td>1.60</td>
</tr>
<tr>
<td>BTM</td>
<td>317.48</td>
<td>1005.76</td>
</tr>
<tr>
<td>Turnover*</td>
<td>2.22</td>
<td>1.88</td>
</tr>
<tr>
<td>ROE</td>
<td>-0.01</td>
<td>1.21</td>
</tr>
<tr>
<td>Analyst Coverage</td>
<td>59.52</td>
<td>53.49</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>0.04</td>
<td>0.09</td>
</tr>
<tr>
<td>Advertising Intensity</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>SG&amp;A</td>
<td>0.24</td>
<td>0.22</td>
</tr>
<tr>
<td>Capital Expenditure</td>
<td>0.09</td>
<td>0.23</td>
</tr>
<tr>
<td>Leverage</td>
<td>0.04</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Number of observations: 11,610 (for 1991-2013 sample)

Observations with R&D and SG&A above 1 were excluded. R&D was imputed for those with missing observations (R&D=0).

*Monthly

Table 2 – Correlations among our Material Sustainability(orthogonalized) from different mappings

<table>
<thead>
<tr>
<th>SASB-KLD Mapping</th>
<th>SASB-SIC Mapping</th>
<th>Mean</th>
<th>SD</th>
<th>Q1</th>
<th>Q2</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Actual KSY</td>
<td>0.007</td>
<td>0.645</td>
<td>-0.076</td>
<td>0.083</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>KSY’s Replicated</td>
<td>-0.003</td>
<td>0.458</td>
<td>-0.019</td>
<td>0.026</td>
<td>0.5726</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>KSY’s Replicated</td>
<td>0.004</td>
<td>0.460</td>
<td>-0.023</td>
<td>0.043</td>
<td>0.5531</td>
<td>0.8595</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>AUTHOR’s SASB</td>
<td>0.001</td>
<td>0.717</td>
<td>-0.063</td>
<td>0.054</td>
<td>0.3565</td>
<td>0.4202</td>
<td>0.4008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>AUTHOR’s AUTHORS</td>
<td>0.001</td>
<td>0.662</td>
<td>-0.061</td>
<td>0.064</td>
<td>0.335</td>
<td>0.4024</td>
<td>0.4323</td>
<td>0.786</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>TECH’s SASB</td>
<td>0.001</td>
<td>0.629</td>
<td>-0.044</td>
<td>0.04</td>
<td>0.4674</td>
<td>0.4979</td>
<td>0.4706</td>
<td>0.4126</td>
<td>0.3529</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>TECH’s AUTHORS</td>
<td>0.006</td>
<td>0.619</td>
<td>-0.042</td>
<td>0.045</td>
<td>0.4333</td>
<td>0.4717</td>
<td>0.5122</td>
<td>0.3775</td>
<td>0.4116</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Obs.= 11,610

SASB and AUTHOR matching agree 85% of the time.
Table 3 – Bayesian Model Selection and Model Averaging

<table>
<thead>
<tr>
<th>IV process Sectors</th>
<th>Eight Occam’s windows within which models are compared</th>
<th>Specification with highest posterior probability within window</th>
<th>Coefficient estimate from highest probability model within window</th>
<th>Probability-weighted estimate of marginal effect at mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Orth. Top Quintile</td>
<td>Orth. Continuous</td>
<td>Raw Improved</td>
<td>Raw Continuous</td>
</tr>
<tr>
<td></td>
<td>Original 6</td>
<td>All</td>
<td>Original 6</td>
<td>All</td>
</tr>
<tr>
<td>KLD-SASB Ind Match</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N/A*</td>
<td>Authors</td>
<td>SASB</td>
<td>Authors</td>
<td>TECH</td>
</tr>
<tr>
<td>SASB</td>
<td>SASB</td>
<td>SASB</td>
<td>SASB</td>
<td>SASB</td>
</tr>
<tr>
<td>Fixed effect Year &amp; Sector</td>
<td>Year &amp; Year &amp; Sector</td>
<td>Year &amp; Year &amp; Sector</td>
<td>Year &amp; Year &amp; Sector</td>
<td>Year &amp; Year &amp; Sector</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Obs.</td>
<td>119,277</td>
<td>262,749</td>
<td>119,277</td>
<td>262,749</td>
</tr>
<tr>
<td>B_0</td>
<td>0.514</td>
<td>0.100</td>
<td>-0.095</td>
<td>0.050</td>
</tr>
<tr>
<td>s</td>
<td>0.095</td>
<td>0.061</td>
<td>0.043</td>
<td>0.029</td>
</tr>
<tr>
<td>Relative Post. Prob.</td>
<td>77%</td>
<td>43%</td>
<td>46%</td>
<td>26%</td>
</tr>
<tr>
<td>B_0</td>
<td>M_i</td>
<td>0.498</td>
<td>0.055</td>
<td>-0.048</td>
</tr>
<tr>
<td>s</td>
<td>0.094</td>
<td>0.061</td>
<td>0.048</td>
<td>0.037</td>
</tr>
</tbody>
</table>

Units are percent per month, all samples 1991-2013
* We cannot observe the firm-industry match KSY used in constructing their actual measure.
Table 4 – Exploring the Contingent Association Between KSY’s Top Quintile and Stock Return

<table>
<thead>
<tr>
<th>Predictors</th>
<th>(1) Returns</th>
<th>(2) Returns</th>
<th>(3) Returns</th>
<th>(4) Returns</th>
<th>(5) Returns</th>
<th>(6) Returns</th>
<th>(7) Returns</th>
<th>(8)* Returns</th>
</tr>
</thead>
<tbody>
<tr>
<td>KSY’s MatSust. (orth., top quintile)</td>
<td>0.514</td>
<td>0.205</td>
<td>0.687</td>
<td>-0.225</td>
<td>0.0410</td>
<td>0.0747</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0950)</td>
<td>(0.241)</td>
<td>(0.172)</td>
<td>(0.119)</td>
<td>(0.103)</td>
<td></td>
<td></td>
<td>(0.109)</td>
</tr>
<tr>
<td>IMPUTED</td>
<td>0.992</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0705</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.154)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.200)</td>
<td></td>
</tr>
<tr>
<td>MEASURED</td>
<td>0.0761</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0964</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.105)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.106)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>2.705</td>
<td>2.272</td>
<td>3.786</td>
<td>6.790</td>
<td>0.474</td>
<td>2.216</td>
<td>2.224</td>
<td>2.812</td>
</tr>
<tr>
<td></td>
<td>(0.877)</td>
<td>(0.879)</td>
<td>(2.621)</td>
<td>(1.242)</td>
<td>(1.262)</td>
<td>(0.948)</td>
<td>(0.920)</td>
<td>(0.997)</td>
</tr>
<tr>
<td>Controls</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Year FE</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Sector FE</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>SectorXYear FE</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Observations</td>
<td>119,277</td>
<td>119,277</td>
<td>13,543</td>
<td>64,761</td>
<td>40,973</td>
<td>119,277</td>
<td>119,277</td>
<td>100784</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.031</td>
<td>0.031</td>
<td>0.013</td>
<td>0.044</td>
<td>0.01</td>
<td>0.036</td>
<td>0.036</td>
<td>0.029</td>
</tr>
</tbody>
</table>

*Without the Extractive Materials sector

Standard errors in parentheses
Appendix 1 – Model Uncertainty Annual Analysis with ROS

Figure A1: Marginal Effects at Average – All Models

Figure A2a: Marginal Effects at Average by Source of SASB-ESG Matching

Figure A3a: Marginal Effects at Average by Source of SASB-ESG Matching

Figure A2b: All Replications using KSY’s Data

Figure A3b: All Replications using KSY’s Data
Figure 5: Marginal Effects at Average by Time Period
Figure 5a: All Replications

Figure 6: Marginal Effects at Average by Sector and Time FE
Figure 6a: All Replications

Figure 5b: All Estimates using KSY data

Figure 6b: All Estimates using KSY data
Appendix 2 – KSY’s Top Quintile and Stock Return – Equal Weighted Portfolio Analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Replication (table 4 Panel B)</th>
<th>Imputed Material Sustainability</th>
<th>Measured Material Sustainability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1a</td>
<td>Model 1b</td>
<td>Model 2a</td>
</tr>
<tr>
<td></td>
<td>Low Investment</td>
<td>High Investment</td>
<td>Low Investment</td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>t</td>
<td>Estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.184</td>
<td>(1.62)</td>
<td>0.187</td>
</tr>
<tr>
<td>Market</td>
<td>1.013</td>
<td>(36.75)</td>
<td>1.008</td>
</tr>
<tr>
<td>SMB</td>
<td>0.159</td>
<td>(4.40)</td>
<td>0.101</td>
</tr>
<tr>
<td>HML</td>
<td>0.330</td>
<td>(8.83)</td>
<td>0.0795</td>
</tr>
<tr>
<td>UMD</td>
<td>-0.126</td>
<td>(-5.45)</td>
<td>-0.117</td>
</tr>
<tr>
<td>LIQ</td>
<td>0.148</td>
<td>(4.91)</td>
<td>0.0885</td>
</tr>
<tr>
<td>N</td>
<td>261</td>
<td></td>
<td>228</td>
</tr>
</tbody>
</table>

$t$ statistics in parentheses

Annualized Alpha | 2.23% | 5.62% | 2.27% | 6.93% | 4.91% | 4.09%
Difference in Alphas | 3.39% | 4.66% | 0.81%

In Model 1a and 2a, we estimate, and confirm, KSY’s reported result using a Fama-French 5 factor analysis (see KSY Table 4, Panel B). In fact, the difference in the coefficients between the bottom and top quintile is actually slightly larger than the one they report (3.39% vs 2.69%). We then compare portfolios created from only the imputed part of the bottom and top quartiles; again, the top quintile has a higher return. However, when we compare portfolios formed from those firms in the top and bottom quintiles with real improvement in scores, the result changes. Now the return for the bottom part is actually larger. This again confirms that the imputed part of KSY’s measure seems to be the source of their result.