DO SUSTAINABLE COMPANIES HAVE BETTER FINANCIAL PERFORMANCE? REVISITING A SEMINAL STUDY

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

Abstract: Do high-sustainability companies have better financial performance? An extremely influential publication, “The Impact of Corporate Sustainability on Organizational Processes and Performance,” claims they do. Its 2014 appearance preceded a boom in sustainable investing, and scholars and practitioners have used it to explain these investments. Yet I report here that replication and extension of the original analysis fails to reveal evidence that corporate sustainability influences either stock-market or accounting returns. I then show that the original study also lacks evidence to support an inference that corporate sustainability leads to superior financial performance. Finally, I discuss the importance of my findings for the practical accretion of knowledge.

Monday, March 22, 2025

Corresponding author: Andrew A. King

Address Email: aaking@bu.edu

DO SUSTAINABLE COMPANIES HAVE BETTER FINANCIAL PERFORMANCE? REVISITING A SEMINAL STUDY

In this article, I replicate and extend an exceptionally influential research report: “The Impact of Corporate Sustainability on Organizational Processes and Performance” by Robert Eccles, Ioannis Ioannou, and George Serafeim (2014). Based on an analysis of an 18-year panel of data, the report concludes that “High Sustainability” companies have distinct organizational attributes and that, consequently, “High Sustainability companies significantly outperform their counterparts over the long-term, both in terms of stock market as well as accounting performance” (pg. 2835). In the decade after its publication, EIS has been cited extensively by both scholars[1] and policymakers (Crasi, 2015; Lee, 2020).

Eccles, Ioannou, and Serafeim (2014a) (hereafter: EIS) was not the first publication to investigate the connection between corporate sustainability[2] and corporate financial performance. Dozens of previous studies had identified a possible correlation, but few found evidence of a causal relationship (Berchicci & King, 2022). Indeed, three years before the publication of EIS, Marc Orlitzky noted that “narrative reviews of this literature” typically concluded that “the empirical evidence is too mixed to allow for any firm conclusions” (2011: 409).

EIS seemed to clarify the debate by reporting analyses that used a precisely defined construct, new data, and modern empirical methods. It defined its central construct, “corporate sustainability”, as integrating “social and environmental issues in … business models and daily operations … through the adoption of related corporate policies”. It measured this construct using data from a relatively new rating agency: Asset4 (now Refinitiv). Finally, it employed a sophisticated causal identification strategy: propensity score matching. Scholars remarked on both its methods and results (Aguilera et. al, 2021). In their review of the literature on “sustainable management”, Aragon-Correa, Marcus, Rivera, & Kenworthy (2017) concluded that EIS delivers “[O]ne of the strongest relationships between corporate environmental performance and corporate financial performance.” EIS quickly became a touchstone for the research community.

EIS also influenced business practice and policy. Its 2014 publication preceded a massive growth in “sustainable” investing, and EIS was used to promote these new investment strategies. Al Gore (former U.S. Vice President) and David Blood (co-founder of Generation Investment) used EIS to claim that “investors who identify companies that embed sustainability into their strategies can earn substantial returns.” Allison Herron Lee, former Commissioner of the US Securities and Exchange Commission, cited EIS to support her claim that corporate sustainability assessment has become “a core risk management strategy for portfolio construction” (Lee, 2020). More generally, EIS has been used widely in testimony on policymaking (Crasi, 2015; Blake, 2020).

Yet, despite the notable influence of EIS, no one has endeavored to replicate its analysis of the connection between corporate sustainability and financial performance. In this article, I attempt to fill this empirical gap. To do so, I must overcome a common problem in conducting replications – the lack of a fully specified empirical process in most research reports (Bloomfield, Rennekamp, and Steenhoven, 2018). To fill in the missing details, I contacted the original authors but received no response, so I proceeded cautiously using the information in the report itself while keeping a record of the assumptions and interpretations I was forced to make. I then specified models encompassing the combination of these assumptions and estimated coefficients for each. This kind of “model uncertainty analysis” is particularly well-suited to replication studies because it facilitates an understanding of the sensitivity of results to varying empirical assumptions (Simonsohn, Simmons, & Nelson, 2020; King, 2023).

In the remainder of this article, I first review the original study, paying particular attention to its empirical design and implementation. Given uncertainties about how to interpret the text, I widen the aperture of my investigation to include a multiverse of empirical models that might match the original analysis, and I calculate, report, and interpret estimates from all of these models. I then extend the original analysis by using alternative methods and measures.

In an attempt to harmonize my conclusions with those reached in the EIS analyses, I revisit the original report. I conclude that there are several reasons to be cautious in forming inferences from its evidence: employed measures lack precedent, an important method requires unusual conditions to operate, and the significance of one empirical test is misreported[3].

I conclude that neither my replication nor the original report provides sufficient evidence to support EIS's claim that "High Sustainability companies significantly outperform their counterparts over the long-term, both in terms of stock market return as well as accounting performance" (pg. 2835).

OVERVIEW OF THE ORIGINAL REPORT

Eccles, Ioannou, and Serafeim (2014) does not advance any formal hypotheses, but it does report an "overarching thesis" that organizations that voluntarily integrate environmental and social policies in their business model "represent an alternative and distinct way of competing for the modern corporation" (p. 2836). This claim was, and remains, a provocative one, because it suggests that "corporate sustainability" might be a set of reinforcing attributes and practices that provide sustained competitive advantage (Hart, 1995; Greening & Turban, 2000;

Casadesus-Masanell & Ricart, 2012). Hundreds of studies have tried to test this conjecture (Aguilera et al, 2021).

EIS adds to the previous literature by conducting a two-part analysis. It first supports its claim that sustainability is connected to other organizational attributes by showing that companies that adopt more "sustainability policies" (see EIS Appendix 1 for the descriptions) have different organizational practices. EIS then tests whether companies adopting more sustainability policies, "High Sustainability companies" in their terminology, have higher financial performance. It reports six tests of this conjecture, employing two measures of stock return and four measures of accounting return. EIS uses this evidence to justify its aforementioned claim that "High Sustainability companies significantly outperform their counterparts" (pg. 2835). It is this claim that has proven to be most influential[4], and it is the focus of my replication.

The EIS empirical strategy

EIS's strategy for causal identification is based on a simple but powerful idea: if a treatment can be randomly assigned, then the differences in post-treatment outcome measures will capture the effect of the treatment (Rubin, 1974). Unfortunately, for many research questions, including EIS's, random assignment is not possible. For these situations, scholars have developed methods where matching is used to mimic the effect of random assignment (Rosenbaum & Rubin, 1983). EIS's empirical strategy depends on such a matching technique: using observable accounting measures, it forms pairs of similar treated and control firms. It then assumes that having matched firms using observable attributes, the unobservable differences between these firms are also matched, and thus the treated firm of the pair was chosen "as if" by random.

Once matches have been made, EIS evaluates the performance effect of High Sustainability by comparing the means of the High and Low Sustainability groups relative to three different outcome variables (stock return, return on assets, and return on equity) using two weighting criteria (equal and market-value-weighted). Based on these six comparisons (three measures X two weightings), EIS concludes that: “during the 18-year period we studied, the High Sustainability firms outperformed the Low Sustainability ones in terms of both stock market and accounting measures” (p. 2853).

REPLICATING EIS’S ANALYSIS OF FINANCIAL PERFORMANCE

Replication Strategy

Replicating any empirical study can be challenging, and thus, few replications are conducted, and even fewer are published (Bloomfield et al., 2018). Fortunately, a few journa*ls* have taken steps to correct the situation by welcoming replication studies and issuing guidance on structuring replication reports (Bettis et al., 2016; Köhler et al., 2023). In my review, I follow recent recommendations for conducting a replication (Ethiraj, Gambardella & Helfat, 2016; King, 2023).

I first attempted to replicate the analysis as closely as possible but discovered several uncertainties and contacted the original authors for assistance. Receiving no reply despite multiple attempts, I proceeded with my replication while paying close attention to the uncertainties I uncovered about EIS’s measures and method (see Figure 1). When uncertainties arose, I created methods and models encompassing different ways these uncertainties could be resolved, and I added this multiverse of models to my analysis. I also identified how the original analysis could be enhanced and performed an empirical extension. Finally, I tried to reconcile the results from my replication, extension, and the original report.

Insert Figure 1

Replicating the Sample

To be in EIS's sample, firms had to be rated on sustainability by Refinitiv in 2003-05, not operate primarily in the "finance" sector, and report accounting data for each year from 1993-2010. EIS reports using a multistage process to winnow this eligible group to 90 High Sustainability and 267 Low Sustainability firms. Figure 2 shows a flow chart of EIS's process and my attempt to replicate it. Ghosting shows missing information about EIS's process.

I began my replication by accessing Refinitiv data and cleaning it of duplicate reports. Consistent with EIS, I identified 775 US firms with Refinitiv data in the window 2003-2005[5]. I then followed EIS in removing 100 banking, finance, and insurance companies, leaving me with 675[6].

EIS provides limited information about the next steps in the sample formation. It does not disclose what accounting data was used, how it was matched to the Refinitiv database, or how many firms were removed for failure to conform with the required sample frame (1993-2010). Out of caution, I used both Compustat and Worldscope business accounting data. Doing so allowed me to match 93% of the Refinitiv sample to accounting data – leaving me with 649 firms, but 12 of these firms had missing Refinitiv policy scores, and 7 had incomplete accounting data, leaving 630 firms.

Following EIS, I then removed firms not continuously operating between 1993 and 2010[7]. Enforcing this requirement caused me to lose 194 firms – 31%. Fearful that incomplete accounting data was causing me to eliminate some firms improperly, I conducted additional analyses using Amazon Mechanical Turk workers and identified a disqualifying event for 88%

of the removed firms (170).[8] I interpreted this to mean that these firms truly did not match the sample frame and should be removed.

I then followed EIS's use of Refinitiv data on sustainability policies in 2003-2005 to separate firms into four quartiles and labeled the 25% of firms with the most policies as "high sustainability" and the 50% with the fewest as "low sustainability." This left me with a sample of 327 firms (109 *HS* and 218 *LS*).

EIS reports using a sample of fewer HS firms (90) and more LS firms (267), but EIS notes that it winnowed its sample of *HS* firms by using 200 interviews to "ensure that firms in the *HS* cohort had "adopted a substantial number of these [sustainability] policies in the early to mid-90s" (p 2837). Based on these interviews, EIS disqualified 77 firms from inclusion in the *HS* group. EIS does not report how these interviews were conducted or processed, so I cannot replicate this step. I describe later how I use a Monte Carlo method to evaluate the potential effect of removing these unknown firms.

Insert Figure 2

Replicating the Matching Method

EIS reports using logistic regression to obtain propensity scores for matching each firm in its *HS* group to a firm in its *LS* group. The equation for this logistic regression appears to be:

$$Logit(P_i) = B * Ln(assets)_i + B * ROA_i + B * Turn_i + B * Leverage_i + B * MTB_i + \varepsilon_i \qquad \text{Eq. 1}$$

where P_{i} is the probability that the i^{th} firm is a member of the *HS* group and the predictor variables are "the natural logarithm of total assets (as a proxy for size), return on assets, asset turnover (measured as sales over total assets), market value of equity over book value of equity (MTB) as a proxy for growth opportunities, and leverage (measured as total liabilities over total assets)" (EIS: 2837). Using these scores, matches are made between each *High Sustainability* firm and a unique control firm operating in the “same industry classification benchmark subsector (or sector if a firm in the same subsector is not available)” (EIS: 2837).

For matching to allow inference about average treatment effects, paired firms should be very similar, and for this reason, it is standard practice to set a maximum distance (or caliper) allowed between the propensity scores of matched pairs. Calipers less than 0.2 or 0.25 have been shown to provide better estimates (Lunt, 2014), and many scholars follow Rosenbaum & Rubin (1985) in using a caliper of 25% of the standard deviation of the propensity score (for EIS, this would be a caliper of 0.06).

EIS provides information on the employed caliper in a footnote: “Using a caliper of 0.01 to ensure that none of the matched pairs is materially different reduces our sample by two pairs or four firms. All our results are unchanged if we use that sample of 176 firms” (EIS: 2837). Unfortunately, there are two ways to interpret these sentences: 1) the main analysis identified 88 pairs using a caliper of 0.01 and added two more found with a larger caliper, and 2) the main analysis formed 90 pairs using an unknown caliper and then conducted robustness tests using 88 pairs found with a caliper of 0.01. Given this uncertainty, I decided that my analyses should include matches made with various calipers (0.01, 0.1, 0.25, 0.5, and no limiting caliper at all).

Replicating Measures

EIS reports using stock returns and accounting ratios to create outcome measures of financial performance.

Stock Returns. EIS reports estimating portfolio stock returns using a three-factor Fama-French model augmented by the Carhart momentum factor.

$$R_t = \alpha + B * MKTRF_t + B * SMB_t + B\ HML_t + B * MTB_t + B * UMD\ _t + e_t \qquad \text{Eq. 2}$$

The outcome variable (R) is the portfolio stock return for low or high-sustainability firms minus the risk-free rate for that month. The first three predictor variables comprise the 3-factor

Fama-French model: MKTRF is the market return minus the risk-free rate. SMB is the return on a portfolio of small minus big firms. HML is the stock returns of low MTB minus high MTB firms. The fourth predictor variable, UMD, is the Carhart momentum factor. It captures the stock returns of firms with high prior returns minus firms with low prior returns. The value of α captures the variable of interest: the average month's abnormal stock return for the portfolio. The equation is estimated for a panel of 216 months (*t*) from 1993-2010.

EIS also reports analysis comparing returns from "value-weighted" portfolios. It does not disclose the details of the weighting process, but such weighting usually means that investments in the portfolio are rebalanced at the beginning of each period to be proportional to the market capitalization of the firms. I used this approach in my replication[9].

Accounting Returns. EIS does not use a standard accounting performance measure – such as annual Return on Assets or Return on Equity (ROA and ROE). It employs instead a measure of "cumulative" accounting performance from 1993 to 2010. Though it provides no formula for this calculation, its structure is implied by its discussion:

> *"Based on ROA, investing $1 in assets in the beginning of 1993 in a value-weighted (equal-weighted) portfolio of high sustainability companies would have grown to $7.1 ($3.5) by the end of 2010"* (EIS: 2851).

I interpret this as indicating that accounting returns were compounded in a manner similar to reinvested stock returns. Consistent with this conjecture, a longer draft of the EIS report (posted on SSRN and including the same results) has graphs showing evident compounding (Eccles, Ioannou, and Serafeim, 2014b). Thus, my best guess is that cumulative ROA and ROE were calculated using a formula like Equation 3.

$$Cumulative\ ROA_i = \prod_0^T (ROA_{it} + 1) \text{ and } Cumulative\ ROE_i = \prod_0^T (ROE_{it} + 1) \qquad \text{Eq. 3}$$

where there are *i* firms in a group being considered and the calculation is done for years *t* from 0 to T.

Equation 3 has an appealing parallel with cumulative stock return, but its use has no precedent in the scholarly literature, it is problematic to calculate, and it has no clear interpretation (see Supplement 2). Nevertheless, following guidance on constructive replication, I calculated cumulative ROA and ROE using Equation 3 (Köhler and Cortina, 2023).

Friendly reviewers of this manuscript suggested that EIS might have used a different calculation, based not on the compounding of accounting returns but on their aggregation.

$$Aggregate\ ROA_i = \sum_0^T ROA_{it} \text{ and } Aggregate\ ROE_i = \sum_0^T ROE_{it} \qquad \text{Eq. 4}$$

where there are *i* firms in a group being considered and the calculation is done for years *t* from 0 to T. I also created measures using Equation 4 and performed robustness analyses using these measures.

Design of the Model Uncertainty Analysis

There are three main sources of uncertainty in my replication of EIS: 1) uncertainty in identifying EIS's 90 top firms, 2) uncertainty in matching these 90 firms to controls, and 3) uncertainty in constructing EIS's accounting variables. To confine the uncertainty created by source #1, I used a Monte Carlo process to select many possible groups of 90 High Sustainability Firms. The probability of selecting the correct 90 (from 109) is very small in a single draw (< $3.3\text{x}10^{-8}$). Fortunately, the probability of getting 90% correct is much larger (0.017), and if I conduct 400 random draws, I have a 99.99% chance that one or more of my samples of 90 *HS* firms will be at least 90% correct, that is at least 81 of the 90 firms will match EIS's cohort of *HS* firms. I incorporated this approach into my uncertainty analysis.

To bind the second source of uncertainty, I matched firms in each of these 400 groups of HS firms to *LS* firms using different calipers: 0.01, 0.1, 0.25, 0.5, and 1 (i.e., no limiting caliper).

This meant that my uncertainty analysis now included 2000 portfolios of matched *HS* firms and *LS* firms.[10]

To limit the third source of uncertainty, I calculated estimates using both compounded and aggregated ROA. In an extension to the replication, I also analyzed more traditional measures of annual ROA and ROE.

Because I calculate estimates from many models, I can no longer interpret frequency statistics in the usual way[11]. However, I can follow the logic of frequency testing by comparing estimates from analyses using actual data with estimates obtained from data known to be random (Simonsohn, Simmons, & Nelson; 2020). I follow this approach by creating comparison portfolios where I randomly assign which firm of a matched pair is the *HS* one. I then compare estimates obtained from the actual data (where *B* is unknown) to those from the randomized data (where *B* is known to be zero).

RESULTS OF THE REPLICATION

As Köhler and Cortina (2023) discuss, replication often reveals hidden problems or uncertainties in the original empirical design, and that happened here. Specifically, I could not replicate EIS's success in matching high and low-sustainability firms. EIS reports that 88 of 90 *HS* firms were matched to *LS* counterparts in identical business sectors and differing in propensity score by less than 0.01, but using these criteria, I could match an average of only 9 *HS* firms. What could explain the discrepancy? By inspecting the distribution of firms across sectors and p-scores, I identified cases where matching would be infeasible (see Appendix A), such as treated firms outnumbering available controls in a given business sector or p-scores for treated firms not overlapping those for controls.

Because my sample may differ from EIS's, I conducted additional analyses using simulated data and idealized matching methods (see Appendix A). Estimates from these simulations were consistent with my experience attempting to match the actual sample data. They also suggested that matching 88 firms would be highly unusual - occurring fewer than once in 10^{53} samples.

Descriptive Statistics

Table 1 provides descriptive statistics for my multiple portfolios. Columns 1 and 2 provide data about the full unmatched samples of *HS* and *LS* firms. Columns 2-12 provide information on the matched cohorts (formed using differing *HS* firms and matching calipers) used in my analysis. To allow convenient comparison, columns 13 and 14 reproduce the statistics for the EIS sample.

For columns 1 and 2, there is only one sample, but for columns 3-12, there are 400 samples being evaluated. For all of the columns, I calculate the mean ($\overline{x}$)and standard deviation (s) for each variable for each sample *i*. For those cases where I have 400 samples, I report the median of $\overline{x}$ and s for all *i* samples from 1 to 400, and the 5%-95% interval for $\overline{x}_i$[12]. These are denoted respectively as: M:($\overline{x}_i$), M:(s_i), and I 90:($\overline{x}_i$).

Insert Table 1

Table 1 shows that matching reduces the imbalance in the *LS* and *HS* groups, even when a large caliper is used. The best correspondence between my samples and the EIS sample occurs when I use a caliper of 0.1, in which case all but one of the EIS variable means fit within the 90% interval for my 400-sample means. The exception is firm assets for LS firms. Also, regardless of the caliper used, I am never able to match all 90 HS firms. I discuss both issues further in Appendix A when exploring practical limits to matching.

Analysis of Return Differences Between *HS* and *LS* Portfolios

Table 2 provides information on return differences between portfolios of *HS* firms and matched *LS* firms – computed for stock return, compound ROA, and compound ROE. The first data row provides EIS's estimate of the difference in return (B_{EIS}). The next three data rows summarize the estimates from my 2000 replication portfolios – providing the 5%, 50% (median), and 95% levels of the 2000 estimates. Table 2 then provides information useful to interpreting these estimates: the count of times B is greater than zero, and the count when B is positive and has a 95% confidence interval that does not include zero. To allow direct comparison with the estimate reported by EIS (B_{EIS}), Table 2 also reports the count when B> B_{EIS} and the count when the 95% confidence interval does not include zero.

Rows of the table labeled #Br>B or %Br>B report how frequently (in counts and percentages) the estimate obtained from a portfolio with a randomized treatment was larger than the estimate from that portfolio using the actual treatment. In doing so, I provide a proxy for the traditional frequentist test — if estimates from the random data are larger for more than 5% of the portfolios, then one should not reject the null that the true B=0 across the 2000 portfolios (Simonsohn, Simmons, & Ne*ls*on; 2020).

Insert Table 2

Stock Returns

Table 2, columns 1 and 3, provide information on the replications of EIS's analysis of the difference in stock returns between the *HS* and *LS* portfolios. As shown in column 1, for equal-weighted portfolios, the median B is zero, and 90% of the estimates range between -0.002 and 0.005. The coefficient reported by EIS (B_{EIS} = 0.0017) fits comfortably in this interval. Table 2 also shows that 57% of my estimates of B are positive, indicating a larger return for the *HS* group than the *LS* one, and about a quarter are greater than the estimate reported by EIS. For value-weighted portfolios (column 3), about 80% of the returns are positive, and eight (0.4%) of these estimates suggest a greater performance gap than the one EIS reported.

Notably, the uncertainty of my estimates is such that they could not be used to reject the null hypothesis that the true B=0. For both equal- and value-weighted analyses, none of the estimates have a confidence interval exclusive of zero, and thus they would not be judged to pass a typical significance test. Thus, across 2000 portfolios, I can replicate the magnitude and sign but not the “significance” of the EIS results for equal- or value-weighted stock returns.

Columns 2 and 4 provide information on estimates obtained from portfolios where the *HS* and *LS* designation has been randomized, and thus, the null (B=0) is known to be true. As expected, about half of the estimates are positive. Interestingly, more estimates are “significant” when using the randomized data than the actual data. In other words, a researcher would be more likely to find support for EIS’s result using the randomized data than they would using the actual data.

Turning now to a frequentist analysis of all the estimates from the 2000 portfolios, the row labeled “#Br>B” reports the number of times the estimate from the randomized portfolio exceeds the estimate from the unrandomized equivalent. This provides an analog to a “p-value” by simulating how frequently random data would allow an estimate larger than the one obtained from the actual data. For unweighted returns, the randomized data results in a larger B for 44.45% of the portfolios. For weighted returns, the estimate from the randomized data is larger in 34.7% of the cases. Both counts fail to provide significant evidence that we can reject the null that the true B=0.

Insert Figure 3

Figures 3a-c show the estimates in graphical form – sorted from the largest positive difference to the largest negative difference. Grey spikes show each estimate's 95% confidence interval, with darker spikes indicating portfolios using larger calipers and thus including more firms. The dashed line shows the estimate reported in EIS.

The graphs show that the confidence intervals are large relative to the range of coefficient estimates, revealing the relatively low power of the empirical design. This is particularly true for portfolios comprised of fewer matched firms (see lighter spikes), where the confidence intervals are perceptibly larger. For the actual data, estimates from small portfolios are bunched at the positive extreme for the equal-weighted analysis and the negative extreme for the value-weighted analysis, and I speculate that a few frequently matched firms experienced more extreme stock performance over the 18 years of the study.

For equal-weighted returns, the EIS reported result fits well within the distribution of my estimates, but my calculation of its confidence interval suggests it includes zero – raising a question about its "significance" that I will return to later. For value-weighted results, EIS's reported estimate is larger than all but eight of my estimates, and all are derived from portfolios comprising fewer than 9 *HS* - *LS* pairs.

Cumulative ROA

Columns 5 and 7 in Table 2 show estimates of the difference in performance of the *HS* and *LS* firms for compound ROA. For equal-weighted ROA (column 5), positive coefficients outnumber negative ones, and 11 estimates would be deemed "significantly" different from zero in a classic frequentist test. Moreover, these 11 coefficients are also larger than the one reported by EIS. Thus, for equal-weighted ROA, I am able to replicate the magnitude, sign, and significance of the EIS report[13], but insignificant results predominate across my 2000 portfolios, raising questions about what inferences should be drawn.

For value-weighted ROA, I find a similar (though weaker) pattern (column 7). Positive estimates are more numerous than negative ones, and six of the estimates would be deemed "significantly" different from zero in a classic frequentist test, though none of these "significant" results is larger than the coefficient reported by EIS. In summary, I can replicate the sign and significance of the EIS report for value-weighted ROA[8], but the preponderance of insignificant results raises concerns about making inferences from these estimates.

To aid in forming inferences, it is helpful to compare results from actual data with those from randomized data. As shown in columns 6 and 8, randomized data provide fewer positive coefficient estimates but more that would be considered "significant" (37 for equal-weighted portfolios and 28 for value-weighted ones). Thus, a scholar would be more likely to find "significant" evidence conforming with EIS if they used randomized data. The row labeled "#Br>B" reports an analog to a "p-value" by counting the number of times the estimate from the randomized portfolio exceeded the estimate from the unrandomized equivalent. For equal-weighted portfolios, 25.00% of the estimates from the randomized data are larger, for value-weighted portfolios, 28.65% of the estimates are larger. Both exceed the usual 5% cutoff for significance, meaning I cannot reject the null hypothesis that the true B=0 across the 2000 portfolios.

Insert Figure 4

Figure 4 reveals the large uncertainty of the ROA estimates. This uncertainty is caused by the compounding of ROA, which expands variance and reduces the degrees of freedom by collapsing the 18-year panel into a cross-section. As was the case also with stock returns, the

EIS estimate fits well within the distribution of results for equal-weighted portfolios but appears at the extreme for value-weighted ones. Note that EIS's reported result is shown as a dot, without a confidence interval, because EIS does not report data that would allow the calculation of one. I will discuss this issue when I try to reconcile the findings of the EIS report with those from my replication.

Cumulative ROE

Columns 9 and 11 in Table 2 provide information about the performance difference of *HS* and *LS* firms for cumulative ROE. Across the 2000 portfolios, 45.00(42.35)% of the coefficient estimates for the equal(value) weighted analysis suggest *HS* firms have higher ROE, but these estimates are very uncertain: only 1(2) of these estimates would pass a traditional significance test. I can replicate the magnitude, sign, and significance of the EIS report[8] but show such estimates are rare because they are not robust to changes in the sampling or matching process.

For both equal and value-weighted ROE, the data with randomized treatments (columns 8 and 10) delivers more "significant" results than the actual data, and for both unweighted and weighted compound ROE, larger estimates are found more frequently in the randomized data than in the actual data (51.45% and 52.4% respectively). Thus, I cannot reject the null that B=0 across my 2000 portfolios.

Figures 5 a-d reveal one of the empirical difficulties caused by the use of cumulative ROE as an outcome variable: a very high variance in returns. As with cumulative ROA, this is caused by the amplification of differences through compounding and the simultaneous reduction of degrees of freedom by collapsing the panel into a cross-section. For ROE, it is further aggravated by the influence of leverage. Some firms regularly carry debt nearly equaling their assets, meaning their ROE can be 200% or more. When compounded over 18 years, this causes extremely large apparent cumulative returns.

Insert Figure 5

In summary, across 2000 equal or value-weighted portfolios, I find little support for EIS's conclusion that *HS* firms outperform their *LS* counterparts "both in terms of stock market return as well as accounting performance". For stock return, I can replicate the sign of EIS's estimate but not its "significance", and I find that portfolios with randomized treatments deliver more frequent "confirmatory" estimates. I also find that results from portfolios with randomized treatments outperform those with measured treatments 44.45% (equal-weighted) and 34.7% (value-weighted) of the time, far exceeding the usual cutoff for rejecting the null that B = 0.

For compound ROA and ROE, I can replicate the magnitude, sign, and significance of EIS's reported coefficients, but I show that these estimates are unusual and thus not robust to changes in sampling and matching. I show that randomized data deliver more such "confirmatory" estimates and that coefficients from portfolios with randomized treatment data result in larger coefficients between 25% (equal-weighted ROA) and 52.4% (value-weighted ROE) of the time. Thus, I do not find the evidence sufficient to reject the null hypothesis that the true B=0.

EXTENSION

Replicating previous research can reveal ways to extend that research (Bettis et al., 2016; Köhler et al., 2023). In the case of EIS, replication raised questions about the construction and interpretation of cumulative accounting measures, and it exposed uncertainties about the effectiveness of EIS's matching method. In this section, I extend the EIS analysis by substituting more standard accounting measures and by employing a different identification strategy. I also add an analysis where I extend the data panel from 2010-2021.

Analysis of Annual Accounting Measures

To select a more standard measure of ROA and ROE, I reviewed the finance and accounting literature and interviewed faculty at MIT, LBS, Harvard University, and the University of Pittsburgh. I discovered that annual measures of ROA and ROE are most commonly used.

Annual ROA is usually calculated as net income over assets or net income plus expenditures over assets (Singh et al, 2023), and I chose the former form. I measured Return on Equity as net income over shareholder equity.

$$ROA_{it} = \frac{Net\ Income_{it}}{Total\ Assets_{it}} \text{ and } ROE_{it} = \frac{Net\ Income_{it}}{Shareholder\ Equity_{it}} \qquad \text{Eq. 5}$$

The selection of annual ROA or ROE allows the use of time-varying panel data, and it also entails the selection of time-varying control variables. Previous scholars have used covariate controls and fixed effects to account for some sources of unobserved heterogeneity (c.f. Waddock & Graves, 1997). To remain consistent with the logic of EIS, I chose to use the covariates its authors used when matching *LS* and *HS* firms: Size (log assets), Turnover (revenues/total assets), Leverage (total liabilities/total assets), MTB (market value/ (total assets – total liabilities)), and fixed effects for the industry and year. I specify:

$$Y_{it} = B * High_i + B * Size_{it-1} + \\ B * Turn_{it-1} + B * Leverage_{it-1} + B * MTB_{it-1} + \delta_s + \mu_t + \varepsilon_{it} \qquad \text{Eq. 6}$$

where there are *i* firms in *t* years and δ_s and μ_t represent fixed effects for industry (two-digit sic) and year. The outcome variable Y_{it} is ROA or ROE and *High* indicates the firm is in the *High Sustainability* group. Since the sample only includes *HS* and *LS* firms, the coefficient for *High* captures the average performance difference between the groups. To replicate EIS's "value-weighted" analysis, I also specified a weighted-least-squares form of Equation 6, with the weighting calculated as the firm's market value at the end of the previous year divided by the market value of all the firms in the group.

Appendix B provides the results of my analysis in tabular and graphical form. The use of annual data reduces the variance in the estimates but does not change the empirical inference formed from the analysis of compound ROA and ROE. In all cases, the actual data perform little or no better than the randomized data. For ROA (both equal and value-weighted), the actual and randomized pairings result in a nearly equal number of positive and negative coefficient estimates, but the randomized data produces more positive estimates where the confidence interval does not include zero. For ROE, the actual data results in a smaller number of positive estimates and fewer “significant” estimates than are calculated using the randomized data.

Comparisons of coefficient estimates from the randomized and actual data show that the randomized data results in a larger estimate between 47% (for equal-weighted ROA) and 69% (for value-weighted ROE). These far exceed the usual cutoff for significance of 5%. Thus, I do not find evidence, using the totality of my accounting data, that justifies rejecting the null of no difference in accounting performance between *HS* and *LS* firms.

Analysis of Panel Data Without Pairing

Unlike earlier studies of corporate sustainability, EIS uses propensity score matching to reduce the threat of unobserved factors that may bias estimates of the connection between high sustainability and financial performance. Other authors have rightly praised this attempt (Aguilera et. al, 2021), but matching is not the only way to reduce such bias, and matching brings with it other problems. Its success depends on the similarity of the matched pairs, and its use can unintentionally increase bias (King & Nielsen, 2019).

For the EIS analysis, matching also causes a displacement in the timeline of the predictor and outcome variables. The EIS panel contains outcome information beginning in 1993 but does not include information on the predictor variable (corporate sustainability) until 2003 when Refinitiv was founded. To use data from 2003 to discern the sustainability of firms in 1993, one must make several assumptions: 1) the 2003 data accurately measure firm differences in 1993[14], 2) historical financial outcomes did not influence corporate sustainability in 2003, and 3) Refinitiv's assessment of sustainability was not influenced by past financial performance. The temporal displacement of the outcome and predictor variables also entails using a fixed measure of corporate sustainability, thereby precluding the use of firm fixed effects to control for unobserved firm differences.

Method

I extend the EIS analysis by restoring a panel structure to the data and by incorporating firm fixed effects. I also conduct a robustness analysis using a sample extended to 2021.

The form of my analytical model is:

$$Y_{it} = B * HS_i + B * LS_i + B * Survive_i + B * Rated_i + B * Size_{it-1} + B * Turn_{it-1} + B * Leverage_{it-1} + B * MTB_{it-1} + \mu_{i,s,t} + \varepsilon_{it} \qquad \text{Eq.7}$$

where there are *i* firms in *t* years, and $\mu_{i,s,t}$ is a vector of dummy variables capturing constant effects at the year, industry, or firm level. *Survive* is a dummy variable indicating that the firm was in business from 1993-2010, and *Rated* indicates Refinitiv rated the firm in 2003-2005. The other variables are the same as those specified in EIS (discussed earlier). When predicting stock return, I include a lagged measure of ROA.

Table 3 reports estimates from models predicting stock return, ROA, and ROE. The table is arranged horizontally, with the model number and the outcome variable indicated on the left and predictor coefficients and regression information presented on the right. The top three models use 11,850 firms that match EIS's sampling requirement of being a US firm and not primarily in finance, but unlike EIS, the sample does not require continued survival or rating by Refinitiv. Models 4-6 add EIS's restriction to the sample that firms must be rated by Refinitiv (at least one year from 2003 to 2005), but unlike EIS, the sample includes firms in the middle-performance category (i.e. between *HS* and *LS* firms). Models 7-9 restrict the sample to those years after the Refinitiv ratings began (2003). Models 10-12 expand the panel beyond 2010 to 2021.

Models 13-18 differ from the others in employing firm-fixed effects to control for unobserved constant firm attributes. Three of these models evaluate years where EIS could have performed a panel analysis (2003-2010), and three models extend the data panel to 2021. For all models, significance tests of coefficients (with respect to zero) are indicated with asterisks (*), and significant tests comparing the *HS* and *LS* coefficients are indicated by <u>underlining</u>.

Results

Turning first to the control variables *Survive* and *Rated*. The positive coefficient estimates for the variable *Survive* in Models 1-3 may suggest that low-performing firms are less likely to survive from 1993-2010. The coefficients for *Rated* may reveal information about Refinitiv's selection of firms in 2003 because it suggests Refinitiv's initial sample included firms with higher 10-year stock returns (though not higher ROA or ROE). Models 1-3 provide no evidence to support the idea that *HS* firms outperform. All coefficients for *HS* are negative, and for ROA the *HS* coefficient is significantly smaller than the *LS* one.

The pattern of results changes if the sample is limited to rated firms (Models 4-6). Now, I estimate a positive and significant coefficient for the connection between stock return and *HS* firms. This remains true if I limit the sample to those years after the start of Refinitiv's rankings (Models 7-9). I obtain a similar result if I extend the sample to include data up to the year 2021

(Models 10-12), though this time, I estimate that *LS* firms underperform (rather than *HS* firms overperforming). The coefficient for HS firms and ROA is negative in the extended sample, but I am reluctant to make too much of this single estimate.

Estimates from Models 4, 7, and 10 could be interpreted as suggesting that (among those firms that Refinitiv rated) *HS* firms had higher stock returns than *LS* firms, but caution is appropriate. All three models are cross-sectional and lack controls or corrections for unobserved firm differences that might jointly explain *HS* membership and stock return. In Models 13-18, I account for unobserved firm-level differences by including firm-fixed effects. As shown in Table 3, estimates from these models reveal no significant difference between the *HS* and *LS* stock returns or accounting performance.

In a final un-tabulated analysis, I conducted a robustness test by relaxing EIS's assumption about the binary form of the effect (the use of dummy variables indicating *HS* and *LS* firms). Dichotomizing a continuous variable can reduce its explanatory power and make it harder to estimate the true relationship. Consequently, I substituted a continuous measure of policy adoption and reran Models 16-18. I again failed to find evidence to support outperformance (relative to stock return, ROA, and ROE) by firms that had adopted a greater number of sustainability policies.

My extension does not provide evidence that *HS* firms have higher accounting performance. It does uncover evidence that supports the idea of a cross-sectional association between *High Sustainability* and stock return (among firms rated by Refinitiv), but evidence of this association disappears when analytical methods account for unobserved firm differences. These findings are consistent with previous evidence that an apparent correlation between sustainability and financial performance often disappears when methods correct for unobserved heterogeneity (Orlitzky, 2011; Berchicci and King, 2022).

Insert Table 3

RECONCILING THE REPLICATION AND ORIGINAL REPORT

In my replication and extension, I do not find evidence sufficient to justify EIS's claim that "High Sustainability companies significantly outperform their counterparts over the long-term, both in terms of stock market as well as accounting performance" (pg. 2836). Thus, the original study and my replication, though they employ the same data, seem to suggest conflicting conclusions.

How should scholars adjudicate the differences? Can the results be reconciled and synthesized in some way? Fortunately, I believe they can. As I show below, a detailed analysis of EIS reveals that it, too, does not report evidence sufficient to support the conclusion that High Sustainability firms have higher stock returns and accounting performance.

Missing or miscalculated Statistical Tests

EIS uses six comparisons to justify the inference that *HS* firms outperform their *LS* counterparts, but it reports significance tests for only two, and one of these tests appears to be based on a miscalculation.

EIS reports two tests of the difference in stock returns for *HS* and *LS* portfolios. It asserts that for value-weighted portfolios, the difference is "significant at less than 5% level," and for equal-weighted portfolios, it is "significant at less than 10% level" (p. 2849). However, a recalculation of the tests (see Supplement 5) suggests that the second statement is inconsistent with EIS's tabulated results. When I pointed this out to editors at *Management Science*, they contacted the authors and confirmed that the difference in the equal-weighted portfolios was insignificant. Subsequently, the authors published an Erratum Corrige acknowledging a "typo" and clarifying that "For the equal-weighted portfolio analysis, the difference in alphas between the low and high sustainability portfolios is *not* statistically significant "(Eccles, Ioannou, Serafeim, 2024). They did not opine about how this should change the interpretation of the report or modify the published claims

Additionally, the EIS report lacks critical information about other tests of relative performance. For accounting measures, EIS reports estimates of the difference in performance between *HS* and *LS* portfolios but provides no information on the variance of these estimates or the results of significance tests. For compound return on equity, it reports the number of years the *HS* portfolio outperformed, but it does not clarify if this comparison was made for equal or value-weighted returns, nor does it indicate how the analysis was conducted.

In total, for six financial performance tests, EIS reports conducting a significance test for only two, and only one of those tests meets the usual standard for significance. The joint probability of observing one significant result (or more) in six independent trials is 0.26 – five times the usual scientific standard of 0.05. Thus, the EIS report, like my analysis, does not support the conclusion that "High Sustainability companies significantly outperform their counterparts over the long-term, both in terms of stock market as well as accounting performance" (pg. 2836).

Uncertainties in EIS's Matching Method Make Its Estimates Hard to Interpret

EIS attempts to improve on previous research by better identifying the connection between corporate sustainability and outcome variables, but its report of the employed identification method is ambiguous, and despite many attempts, I could not replicate the matching success reported in EIS.

Because successful and repeatable matching is critical to interpreting EIS, I explored the issue further using computer simulation (Appendix A). To evaluate matching under ideal conditions, I assume that *HS* and *LS* groups are distributed similarly across business sectors and their propensity scores have the same distribution. I then draw 5000 distributions of *90 HS* and *269 LS* firms and identify matches. As shown in Appendix A Table 1, even under these idealized conditions, I can never obtain EIS's reported success of matching 88 firms with a propensity score difference of 0.01 or less. Indeed, I estimate such a match would occur fewer than once in 10^{53} samples. Even if the caliper is enlarged ten times to 0.1, I cannot match 88, and I calculate that such success would occur in fewer than one in 2.9 million samples.

My simulation adds to the uncertainty surrounding the matching method used in EIS. If the matching was done improperly, or if the matched pairs were insufficiently similar, the estimates in EIS may be biased. This adds to the difficulty of making inferences from EIS's reported estimates and further suggests that EIS should not be interpreted as showing that corporate sustainability leads to higher financial performance.

CONCLUSION

In this report, I replicate and extend a publication that has influenced scholarship and practice. It has been referenced on Wall Street and Capitol Hill and cited more times than any contemporary or subsequent article in *Management Science*. However, my replication reveals reasons to be cautious in forming inferences from its evidence or conclusions. I am unable to replicate or support its findings, and a close reading of the original report reveals that it, too, lacks evidence to justify the inference of a causal connection between corporate sustainability and financial performance.

My analysis reveals the multifaceted challenge of replicating archival research, where exact replication of a known empirical recipe may not be possible, so replication involves consideration of many alternative processes. This means that replication combines aspects of both review and robustness testing: determining possible empirical pathways requires careful review, and multiple interpretations of the text entail work that spans replication and robustness testing. In combination, such review, replication, and extension provide new perspectives (new camera angles, if you will) that clarify what we can and cannot infer from the data.

My replication reveals that “the market for ideas” does not always select clear and reliable results. As I conducted my work, I was often asked why a replication was needed: hadn’t such a well-known publication already been checked by readers, didn’t its popularity demonstrate its value? Perhaps, but I think this perspective understates the difficulty of reader review. Complex papers like EIS take time to understand in detail, and an individual reader may reasonably choose to delegate the adjudication responsibility to reviewers at the journal or the wisdom of the crowd. In the case of EIS, only by attempting to replicate the research did I come to better understand it. Deconstructing and reconstructing the method allowed me to see how it was made. For influential and sophisticated studies like EIS, replication provides critical guidance on how published evidence should be interpreted and thereby advances the accretion of knowledge.

I hope that the work I report here will encourage others to replicate influential research. I also hope it will buttress the work of scholars developing the institutions needed to ensure that management science remains trustworthy. The influence of studies like EIS demonstrates that our work on business management is important: it can influence investors, policymakers, and maybe even the health of the planet.

REFERENCES

Aguilera, R. V., Aragón-Correa, J. A., Marano, V., & Tashman, P. A. 2021. The corporate governance of environmental sustainability: A review and proposal for more integrated research. *Journal of Management*, 47: 1468-1497. DOI: 10.1177/0149206321991212

Aragon-Correa, J. A., Marcus, A. A., Rivera, J. E., & Kenworthy, A. L. 2017. Sustainability management teaching resources and the challenge of balancing planet, people, and profits. *Academy of Management Learning & Education*, 16: 469-483. DOI: 10.5465/amle.2017.0180

Berchicci, L., & King, A. A. 2022. Building knowledge by mapping model uncertainty in six studies of social and financial performance. *Strategic Management Journal*, 43: 1319-1346. DOI: 10.1002/smj.3374

Bettis, R. A., Helfat, C. E., & Shaver, J. M. 2016. The necessity, logic, and forms of replication. *Strategic Management Journal*, 37: 2193-2203. DOI: 10.1002/smj.2580

Blake, L. 2020. RE: Proposed rule on 'Fiduciary duties regarding proxy voting and shareholder rights' [RIN 1210-AB91]. Letter dated Oct 5, 2020.

Bloomfield, R., Rennekamp, K., & Steenhoven, B. 2018. No system is perfect: Understanding how registration-based editorial processes affect reproducibility and investment in research quality. *Journal of Accounting Research*, 56: 313–362. DOI: 10.1111/1475-679X.12208

Casadesus-Masanell, R., & Ricart, J. E. 2012. Competing through business models. Cheltenham, UK: Edward Elgar Publishing. DOI: 10.2139/ssrn.1115201

Crasi, T. 2015. Testimony on behalf of the National Association of Home Builders before the Senate Committee on Energy and Natural Resources. *Hearing on Energy Efficiency Legislation*, April 30.

Eccles, R. G., Ioannou, I., & Serafeim, G. 2014a. The impact of corporate sustainability on organizational processes and performance. *Management Science*, 60: 2835-2857. DOI: 10.1287/mnsc.2014.1984

Eccles, R. G., Ioannou, I., & Serafeim, G. 2014b. The impact of corporate sustainability on organizational processes and performance. *SSRN Working Paper*. DOI: 10.2139/ssrn.1964011

Eccles, R. G., Ioannou, I., & Serafeim, G. 2025. Erratum to "The Impact of Corporate Sustainability on Organizational Processes and Performance". *Management Science*. DOI:10.1287/mnsc.2025.00751

Ethiraj, S. K., Gambardella, A., & Helfat, C. E. 2016. Replication in strategic management. *Strategic Management Journal*, 37: 2191-2192. DOI: 10.1002/smj.2581

Gore, A., & Blood, D. 2011. A manifesto for sustainable capitalism: How business can embrace environmental, social and governance metrics. *Wall Street Journal*, December 14.

Greening, D. W., & Turban, D. B. 2000. Corporate social performance as a competitive advantage in attracting a quality workforce. *Business & Society*, 39: 254-280. DOI: 10.1177/000765030003900302

Hart, S. L. 1995. A natural-resource-based view of the firm. *Academy of Management Review*, 20: 986-1014. DOI: 10.5465/amr.1995.9512280033

King, A. A. 2023. Writing a useful empirical journal article. *Journal of Management Scientific Reports*, 1: 206-228. DOI: 10.1177/27550311231187068

King, G., & Nielsen, R. 2019. Why propensity scores should not be used for matching. *Political Analysis*, 27: 435-454. DOI: 10.1017/pan.2019.11

Köhler, T., & Cortina, J. M. 2023. Constructive replication, reproducibility, and generalizability: Getting theory testing for JOMSR right. *Journal of Management Scientific Reports*, 1: 75-93. DOI: 10.1177/275503112311760

Lee, A. H. 2020. Playing the long game: The intersection of climate change risk and financial regulation. *Keynote remarks at PLI’s 52nd Annual Institute on Securities Regulation*.

Lunt, M. 2014. Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *American Journal of Epidemiology*, 179: 226-235. DOI: 10.1093/aje/kwt212

Orlitzky, M. 2011. Institutional logics in the study of organizations: The social construction of the relationship between corporate social and financial performance. *Business Ethics Quarterly*, 21: 409-444. DOI: 10.5840/beq201121325

Rosenbaum, P. R., & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70: 41-55. DOI: 10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39: 33-38. DOI: 10.1080/00031305.1985.10479383

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66: 688. DOI: 10.1037/h0037350

Simonsohn, U., Simmons, J. P., & Nelson, L. D. 2020. Specification curve analysis. *Nature Human Behaviour*, 4: 1208-1214. DOI: 10.1038/s41562-020-0912-z

Singh, R. 2023. Defining return on assets (ROA) in empirical corporate finance research: A critical review. *Empirical Economic Letters*, 23(Special Issue 1). DOI: 10.5281/zenodo.10901886

Waddock, S. A., & Graves, S. B. 1997. The corporate social performance-financial performance link. *Strategic Management Journal*, 18: 303–319. DOI: 10.1002/(SICI)1097-0266(199704)18:4<303::AID-SMJ869>3.0.CO;2-G

.

Table 1: Descriptive statistics of pairs of firms matched at five caliper levels

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
		All pre-match		Caliper(0.01)		Caliper(0.1)		Caliper(0.25)		Caliper(0.5)		Caliper(1)		EIS Report	
		LS	*HS*	*LS*	*HS*	*LS*	*HS*	*LS*	*HS*	*LS*	*HS*	*LS*	*HS*	*LS*	*HS*
Assets	M:($\bar{x}_i$)	2.02	16.2	2.77	2.96	4.03	13.8	3.4	16.4	2.81	15.4	2.86	15.3	8.2	8.6
	I 90:($\bar{x}_i$)			1.54-3.35	1.66-3.61	3.56-4.44	6.59-15.2	3.1-3.66	10.9-18	2.57-3.01	11.7-16.6	2.58-3.08	12.1-16.5		
	M:(s_i)	(3.2)	(32.7)	(4.0)	(4.0)	(5.0)	(43.9)	(4.4)	(39.4)	(3.9)	(33.7)	(4.1)	(31.8)	(28.0)	(22.0)
Ln(Assets)	M:($\bar{x}_i$)	20.46	22.58	21.02	20.74	21.38	21.7	21.11	22.21	20.83	22.4	20.75	22.49	N/A	N/A
	I 90:($\bar{x}_i$)			20.7-21.3	20.4-21.2	21.2-21.5	21.5-21.9	21-21.2	22-22.4	20.7-21	22.3-22.5	20.6-20.9	22.4-22.6		
	M:(s_i)	(1.5)	(1.5)	(1.3)	(2.1)	(1.4)	(1.8)	(1.5)	(1.7)	(1.6)	(1.6)	(1.7)	(1.6)	N/A	N/A
ROA	M:($\bar{x}_i$)	6.10%	7.24%	9.34%	8.93%	7.64%	7.36%	6.35%	7.16%	6.57%	7.31%	6.46%	7.31%	7.54%	7.86%
	I 90:($\bar{x}_i$)			7.9-10.7	6.1-10.4	7.1-8.3	6.4-8.1	5.9-6.8	6.4-7.7	6.1-7.1	6.7-7.8	6.0-7.0	6.7-7.8		
	M:(s_i)	(9.28)	(6.30)	(5.94)	(9.32)	(6.51)	(6.37)	(7.44)	(6.59)	(8.36)	(6.99)	(8.46)	(6.53)	(8.02)	(7.54)
Leverage	M:($\bar{x}_i$)	0.53	0.61	0.69	0.51	0.58	0.58	0.56	0.6	0.55	0.6	0.55	0.61	0.57	0.56
	I 90:($\bar{x}_i$)			0.51-0.75	0.48-0.54	0.53-0.61	0.56-0.6	0.52-0.58	0.58-0.62	0.53-0.56	0.59-0.62	0.53-0.57	0.6-0.62		
	M:(s_i)	(0.25)	(0.16)	(0.53)	(0.11)	(0.33)	(0.15)	(0.30)	(0.15)	(0.29)	(0.17)	(0.29)	(0.17)	(0.19)	(0.18)
Turnover	M:($\bar{x}_i$)	1.02	1.03	1.35	1.52	1.06	1.08	1.15	1.14	1.18	1.07	1.12	1.04	1.05	1.02
	I 90:($\bar{x}_i$)			1.12-1.58	1.32-1.71	0.98-1.12	0.99-1.18	1.08-1.2	1.07-1.2	1.12-1.23	1.01-1.11	1.07-1.16	0.99-1.08		
	M:(s_i)	(0.70)	(0.64)	(0.91)	(0.75)	(0.75)	(0.71)	(0.77)	(0.78)	(0.76)	(0.69)	(0.74)	(0.68)	(0.62)	(0.57)
MTB	M:($\bar{x}_i$)	3.45	3.72	3.7	4.12	3.63	3.63	3.25	3.58	3.34	3.93	3.22	3.83	3.41	3.44
	I 90:($\bar{x}_i$)			2.96-4.3	3.21-4.73	3.37-3.84	3.24-3.98	3.08-3.38	3.26-3.8	3.16-3.5	3.68-4.18	3.03-3.37	3.6-4.04		
	M:(s_i)	(2.66)	(2.90)	(2.80)	(3.68)	(2.43)	(2.72)	(1.96)	(2.72)	(2.29)	(3.20)	(2.25)	(3.09)	(2.18)	(1.88)
	$\bar{N}_i$	215	109	9		34		49		68		78		90	
	Min:(N_i)			5		27		44		63		73			
	Max:(N_i)			11		38		55		76		83			

For single samples (i = 1), I report $\bar{x}$, s, and N. EIS do not report ln(Assets).

Table 2: Summary of results from actual and randomized data.

	Stock Return (HS-LS Portfolios)				**Compound ROA** (HS-LS Portfolios)				**Compound ROE** (HS-LS Portfolios)			
	Equal		Value		Equal		Value		Equal		Value	
B_{EIS}	0.0017		0.0037		0.2		2.7		6.5		6	
	1	2	3	4	5	6	7	8	9	10	11	12
	Actual	Random	Actual	Random	Actual	Random	Actual	Random	Actual	Random	Actual	Random
B(5%)	-0.002	-0.002	-0.001	-0.002	-0.163	-0.997	-0.199	-0.863	-77.448	-48.644	-54.401	-35.317
B(median)	0	0	0.001	0	0.439	0.013	0.328	0.005	-1.771	-0.332	-2.178	-0.271
B(95%)	0.005	0.002	0.002	0.003	1.672	1.108	1.308	0.887	11.067	51.095	8.256	40.306
#B >0	1141	1057	1606	1055	1763	1012	1682	1011	900	974	847	975
#B > 0*	0	25	0	5	11	37	6	28	1	25	2	12
#B < 0*	2	25	0	3	0	36	0	31	0	32	1	14
#B>B_{EIS}	497	211	8	7	1484	742	1	0	196	588	211	587
#B>B_{EIS}*	0	25	0	1	11	37	0	0	1	24	1	11
#Br>B	889		649		500		573		1029		1048	
%B >0	57.05%	52.85%	80.30%	52.75%	88.15%	50.60%	84.10%	50.55%	45.00%	48.70%	42.35%	48.75%
%B > 0*	0%	1.25%	0%	0.25%	0.55%	1.85%	0.30%	1.40%	0.05%	1.25%	0.10%	0.60%
%B < 0*	0.10%	1.25%	0%	0.15%	0%	1.80%	0%	1.55%	0%	1.60%	0.05%	0.70%
%B>B_{EIS}	24.90%	10.55%	0.40%	0.35%	74.20%	37.10%	0.05%	0%	9.80%	29.40%	10.55%	29.35%
%B>B_{EIS}*	0%	1.25%	0%	0.05%	0.55%	1.85%	0%	0%	0.05%	1.20%	0.05%	0.55%
%Br>B	44.45%		34.70%		25.00%		28.65%		51.45%		52.40%	

2000 models (400 portfolios X 5 matches) were analyzed for all cases. All Bs measure the difference in returns for the *HS* and *LS* portfolios. Cases with * indicate coefficients whose 95% confidence interval does not include zero.

Table 3: Estimates from Panel Regressions

M	Sample	DV	*HS*		*LS*		*Rated*		*Survive*		N	Firms	R^2	FE
			B	SE	B	SE	B	SE	B	SE				
1	All firms 1993-2010	Stock	-0.001	(0.018)	0.01	(0.015)	0.112**	(0.013)	0.053**	(0.005)	89,517	89517	0.156	Ind.
2		ROA	-0.026**	(0.007)	0.013*	(0.006)	-0.040**	(0.005)	0.053**	(0.002)	89,425	89425	0.238	Ind.
3		ROE	-0.026	(0.024)	0.016	(0.020)	-0.065**	(0.018)	0.107**	(0.007)	89,314	89314	0.07	Ind.
4	Rated firms 1993-2010	Stock	0.089**	(0.016)	-0.040**	(0.013)			-0.034**	(0.013)	9,480	600	0.239	Ind.
5		ROA	-0.006	(0.004)	-0.005	(0.003)			0.010**	(0.003)	9,478	600	0.198	Ind.
6		ROE	0.026	(0.017)	-0.006	(0.014)			0.029*	(0.013)	9,467	600	0.084	Ind.
7	Rated firms 2003-2010	Stock	0.066**	(0.021)	-0.018	(0.017)			-0.037*	(0.016)	4,508	600	0.374	Ind.
8		ROA	-0.003	(0.005)	-0.006	(0.004)			0.003	(0.004)	4,507	600	0.259	Ind.
9		ROE	0.058*	(0.026)	-0.003	(0.021)			0.029	(0.019)	4,502	600	0.118	Ind.
10	Rated firms 2003-2021	Stock	0.011	(0.010)	-0.041**	(0.008)					25,408	3,691	0.14	Ind.
11		ROA	-0.023**	(0.003)	-0.005	(0.003)					25,416	3,691	0.353	Ind.
12		ROE	-0.002	(0.014)	-0.011	(0.011)					25,391	3,690	0.078	Ind.
13	Rated firms 2003-2010	Stock	0.012	(0.021)	0.016	(0.017)					4,508	600	0.442	Firm
14		ROA	-0.004	(0.004)	-0.005	(0.003)					4,507	600	0.095	Firm
15		ROE	-0.058*	(0.027)	-0.043*	(0.021)					4,502	600	0.042	Firm
16	Rated firms 2003-2021	Stock	0.002	(0.012)	-0.012	(0.008)					25,408	3,691	0.219	Firm
17		ROA	0.003	(0.003)	0.003	(0.002)					25,416	3,691	0.056	Firm
18		ROE	0.001	(0.017)	0.004	(0.012)					25,391	3,690	0.009	Firm

All models use controls for lagged log assets, turnover, leverage, market-to-book, and fixed effects for the year. Models predicting stock return include lagged ROA. Standard errors in parentheses. For all models, significance tests for B=0 are indicated with asterisks (**p<0.01, * p<0.05), and significant differences between *HS* and *LS* coefficients are indicated by underlining. Fixed effects at SIC3 for models 1-12. Models 13-18 use firm fixed effects.

Figure 1: Flow Chart of Replication Process

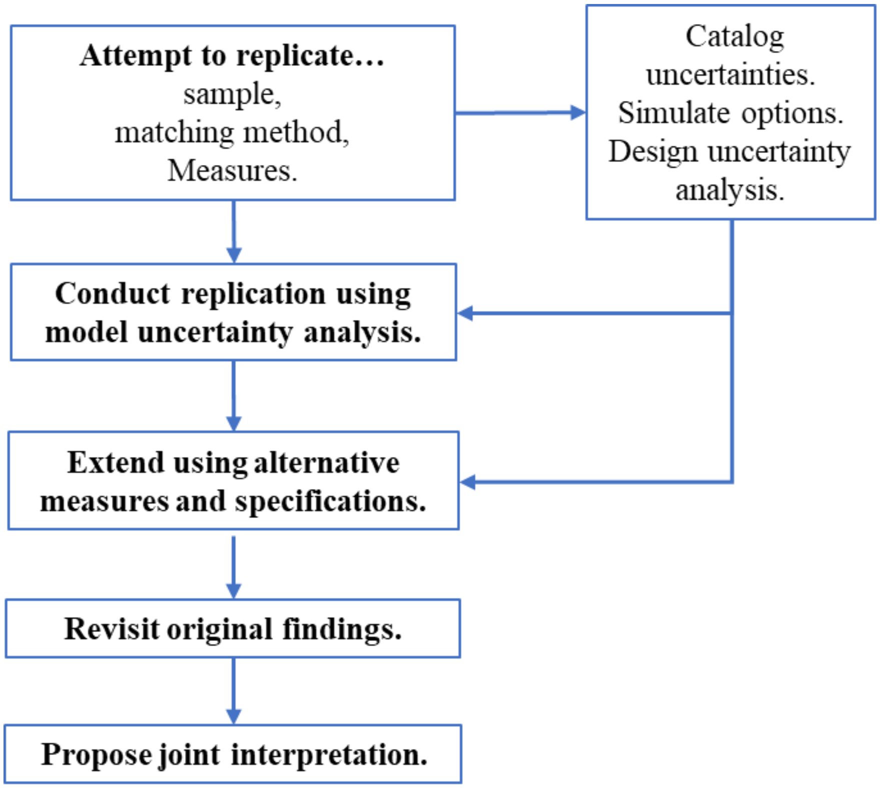
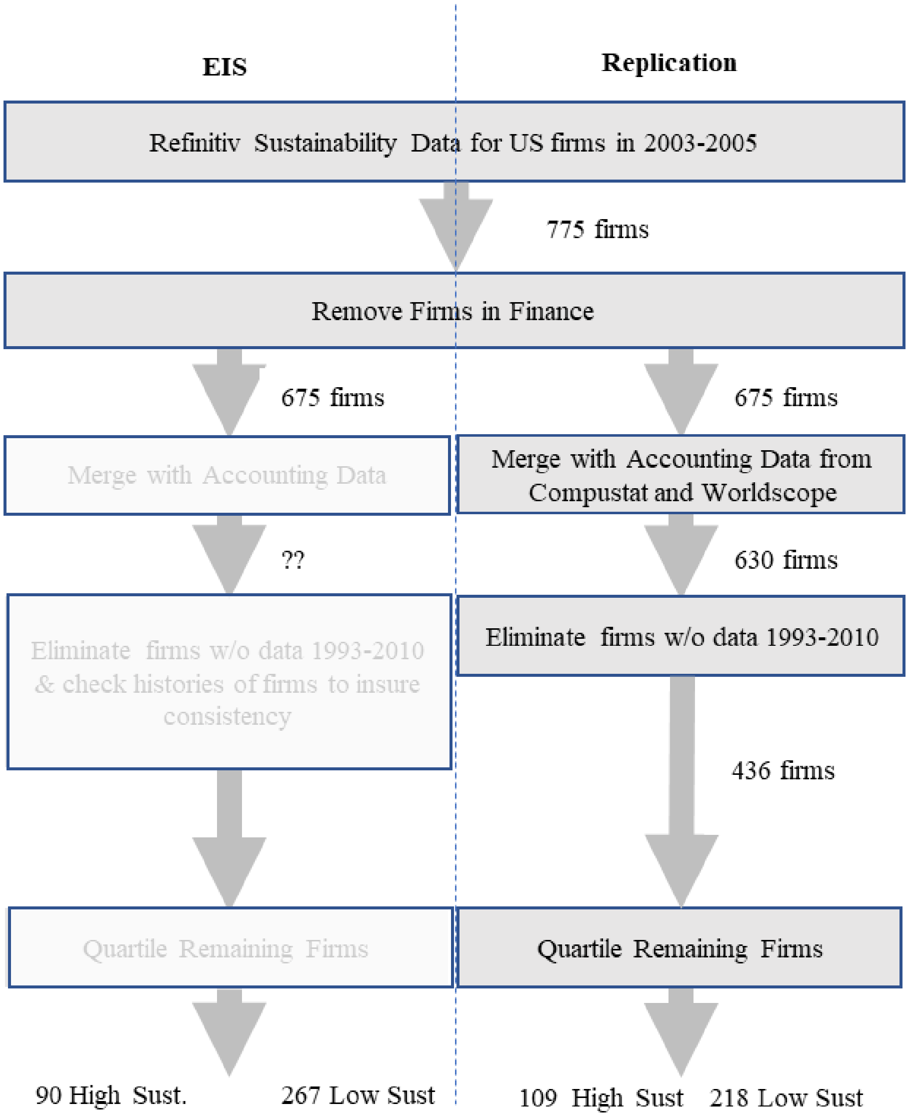

Attempt to replicate…
sample,
matching method,
Measures.
Catalog uncertainties.
Simulate options.
Design uncertainty analysis.
Conduct replication using model uncertainty analysis.
Extend using alternative measures and specifications.
Revisit original findings.
Propose joint interpretation.

Figure 2: Sample Formation Process.


EIS
Replication
Refinitiv Sustainability Data for US firms in 2003-2005
775 firms
Remove Firms in Finance
675 firms
675 firms
Merge with Accounting Data
Merge with Accounting Data from Compustat and Worldscope
??
630 firms
Eliminate firms w/o data 1993-2010 & check histories of firms to insure consistency
Eliminate firms w/o data 1993-2010
436 firms
Quartile Remaining Firms
Quartile Remaining Firms
90 High Sust.
267 Low Sust
109 High Sust
218 Low Sust

Note: steps unreported by EIS in ghosted letters.

Figure 3: Stock returns for high vs low sustainability portfolios

Figure 3a&b: Differences in returns from **equal-weighted** portfolios

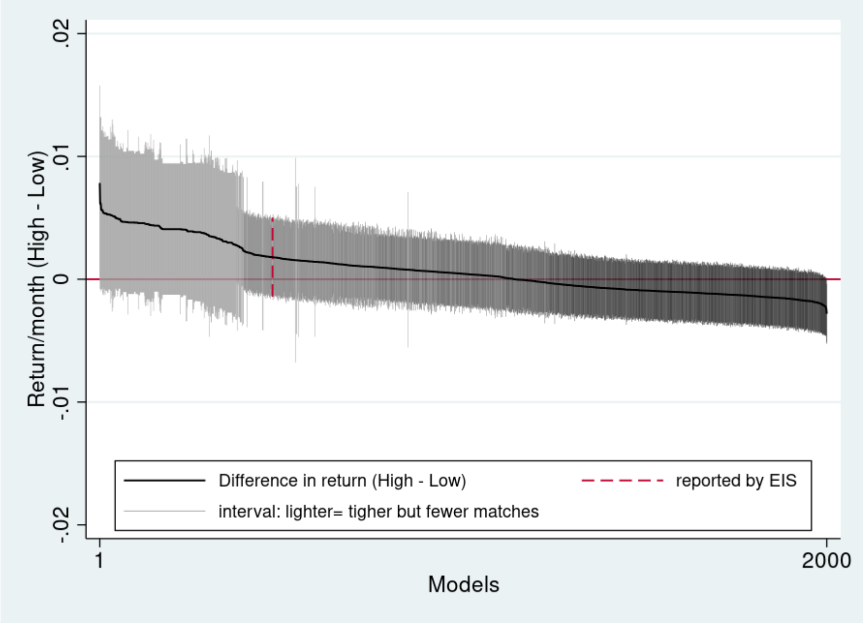
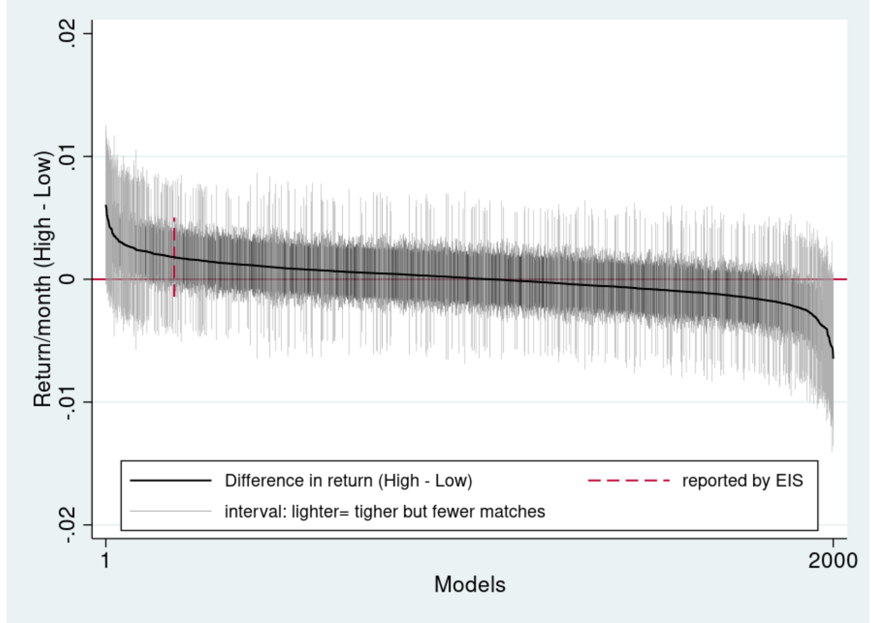
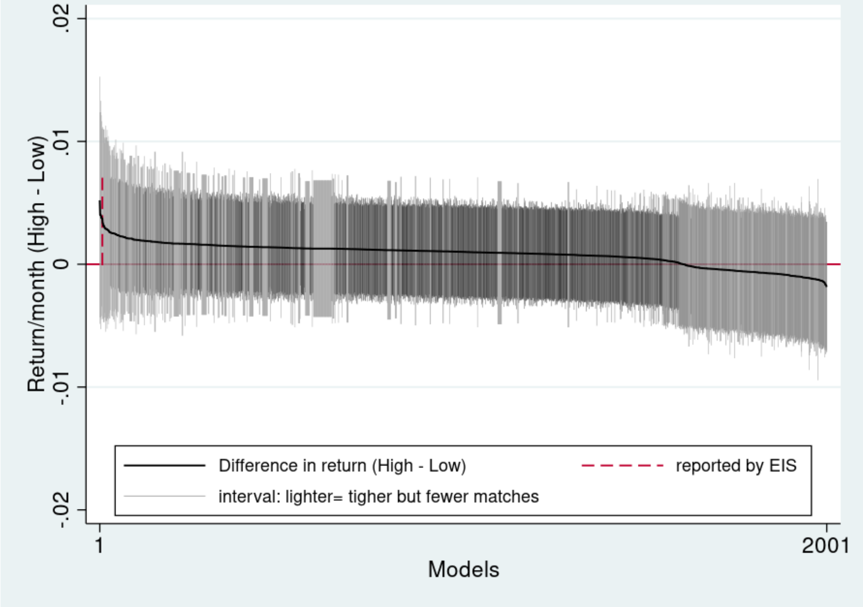
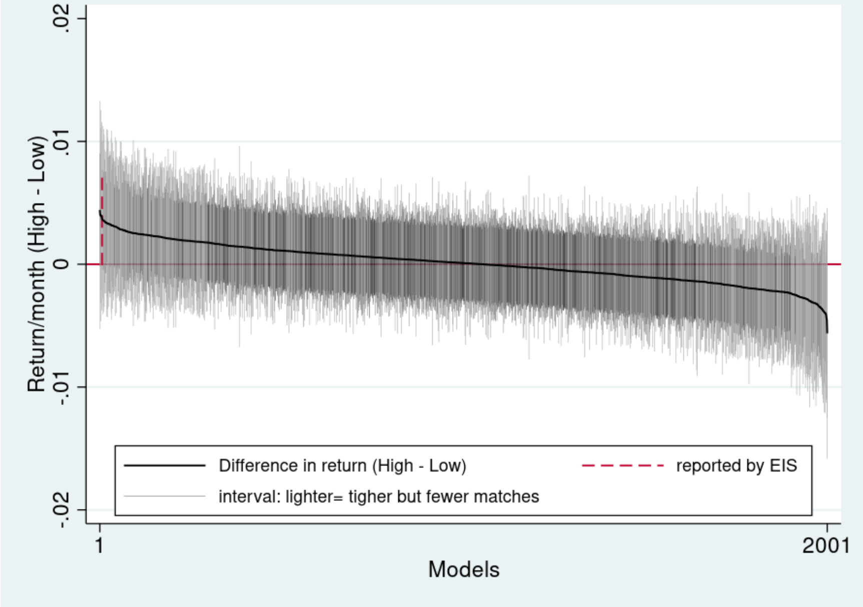
	3a: Actual Treatment	3b: Random (Shuffled) Treatment
		
Positive	57.05%	52.85%
Positive CI ni 0	0.0%	1.25%
Negative	42.95%	47.25%
Negative CI ni 0	0.1%	1.25%

Figure 3c&d: Differences in returns from **value-weighted** portfolios

3a: Actual Treatment



3b: Random (Shuffled) Treatment



Positive	80.3%	52.75%
Positive CI ni 0	0.0%	0.25%
Negative	19.7%	47.25%
Negative CI ni 0	0.0%	0.35%

Figure 4: Compound ROA for high vs low sustainability portfolios

Figure 4 a&b: Differences in **<u>equal</u>-weighted** compound ROA

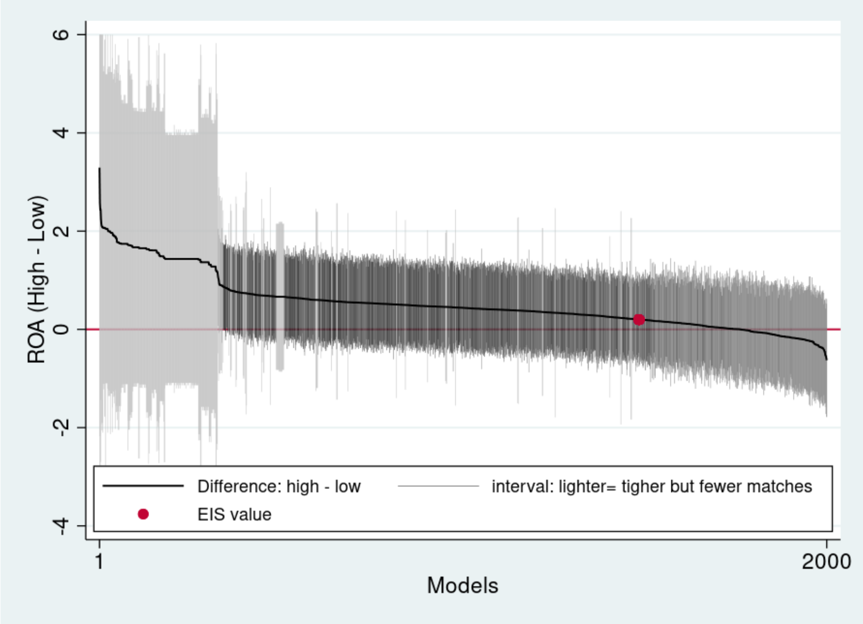
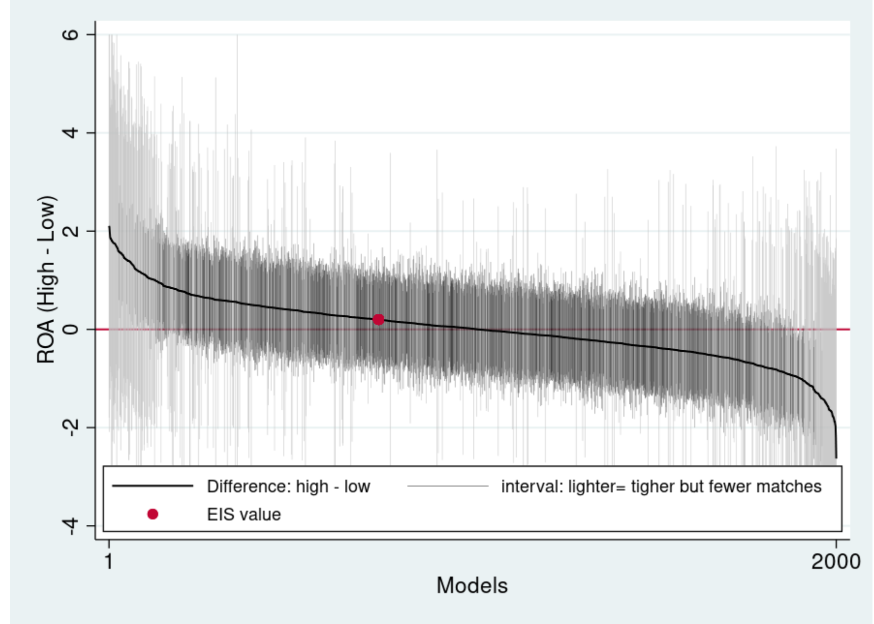
<table>
<tr><th colspan="2">3a: Actual Treatment</th><th>3b: Random (Shuffled) Treatment</th></tr>
<tr><td colspan="2">
</td><td>
</td></tr>
<tr><td>Positive</td><td>88.15%</td><td>50.6%</td></tr>
<tr><td>Positive CI ni 0</td><td>0.55%</td><td>1.85%</td></tr>
<tr><td>Negative</td><td>11.85%</td><td>49.4%</td></tr>
<tr><td>Negative CI ni 0</td><td>0.0%</td><td>1.80%</td></tr>
</table>

Figure 4 c&d: Differences in **value-weighted** compound ROA

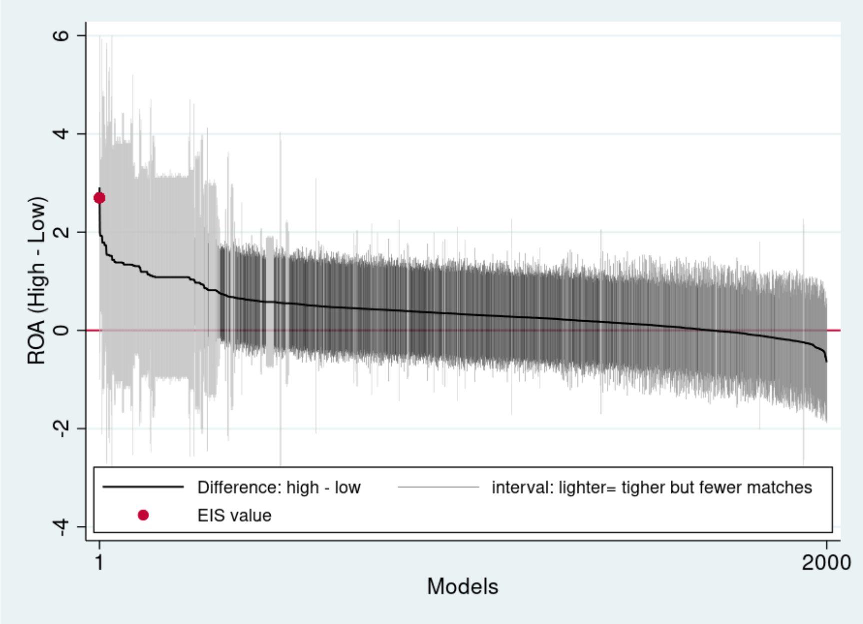
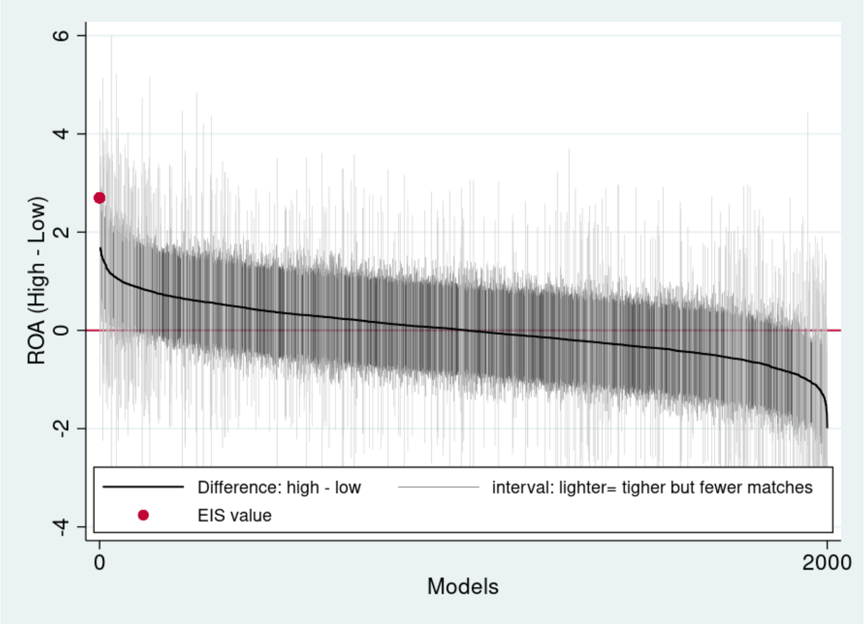
	3a: Actual Treatment	3b: Random (Shuffled) Treatment
		
Positive	84.1%	50.55%
Positive CI ni 0	0.30%	1.40%
Negative	15.9%	49.45%
Negative CI ni 0	0.0%	1.55%

Figure 5: Compound ROE for high vs low sustainability portfolios

Figure 5 a&b: Differences in **equal-weighted** compound ROE

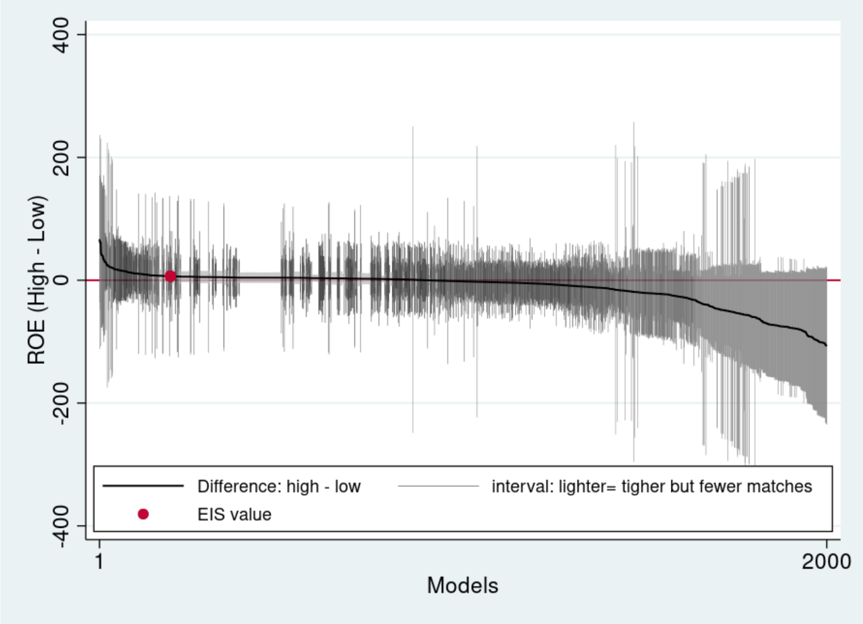
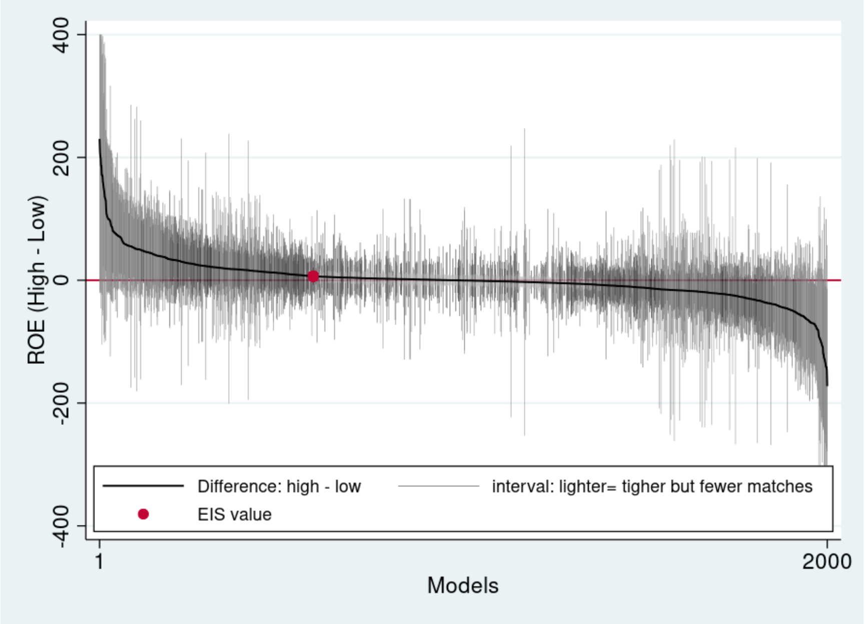
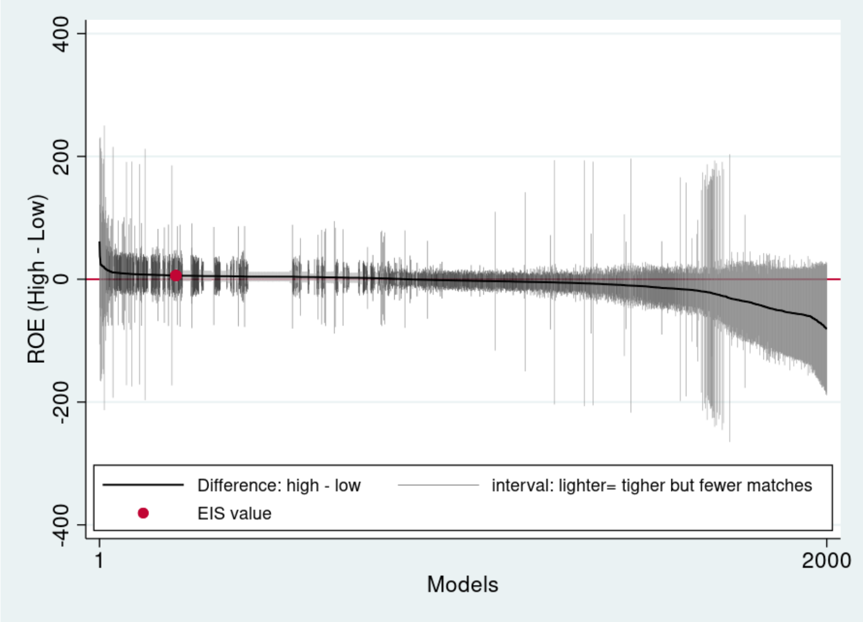
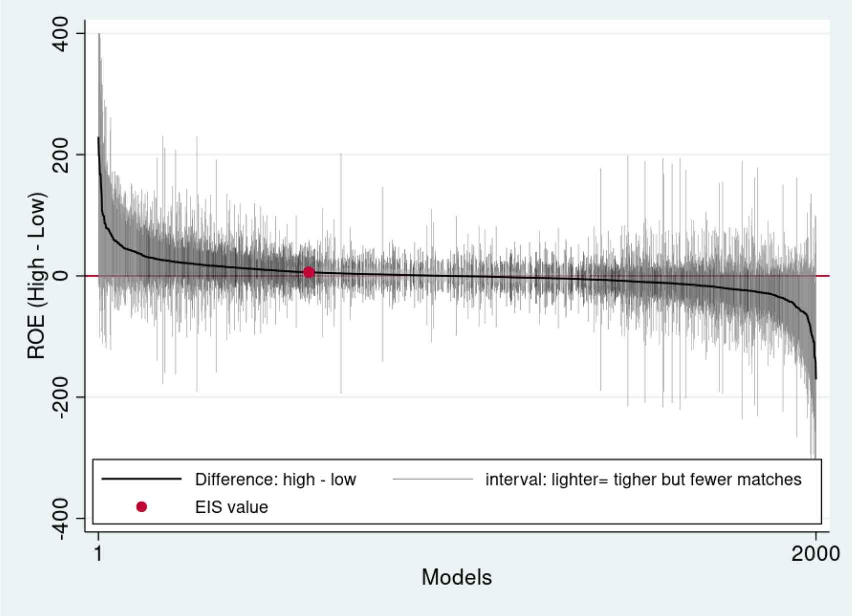
3a: Actual Treatment		3b: Random (Shuffled) Treatment
		
Positive	45.0%	48.7%
Positive CI ni 0	0.05%	1.25%
Negative	55.0%	51.3%
Negative CI ni 0	0.0%	1.60%

Figure 5 c&d: Differences in **value-weighted** compound ROE

3a: Actual Treatment		3b: Random (Shuffled) Treatment
		
Positive	42.35%	48.75%
Positive CI ni 0	0.10%	0.60%
Negative	57.65%	51.25%
Negative CI ni 0	0.05%	0.7%

APPENDIX A: EVALUATING THE FEASIBILITY OF EIS'S MATCHING PROCESS.

There are many apparent barriers to matching *HS* and *LS* firms. As shown in Appendix A Figure A1a, the *HS* and *LS* firms are not distributed equally across the 41 sectors, and there are some sectors where *HS* firms outnumber *LS* ones (Figure A1a) – thereby preventing matches for all *HS* firms. Furthermore, the distributions of propensity scores also differ for the *HS* and *LS* firms – as would be expected given the construction (A1b), and much of this difference is attributable to differences in assets between the two groups (A1c).

To evaluate the feasibility of EIS's reported matching success, I simulated matching under "best case" conditions. EIS report matching 88 *High Sustainability* (*HS*) firms (of 90) with one of 269 *Low Sustainability (LS)* firms where pairs must share the same sector and have propensity scores differing by less than 1% (i.e. caliper <= 0.01). In my simulation, I evaluated different distributions for firm allocation to sectors and for propensity scores (see Supplement 1 for the simulation program). To aid the chance of matches, I assumed both *HS* and *LS* firms had the same probability of being in any industry, so industries with larger numbers of *HS* firms also had larger numbers of *LS* firms. I assumed that the propensity scores for both *HS* and *LS* firms were drawn from the same distribution, though this is not usually the case for propensity scores. In a final simulation, I relaxed this assumption, allowing HS and LS firms to have p-scores that more closely correspond to those observed in the real data (see Figure A1b).

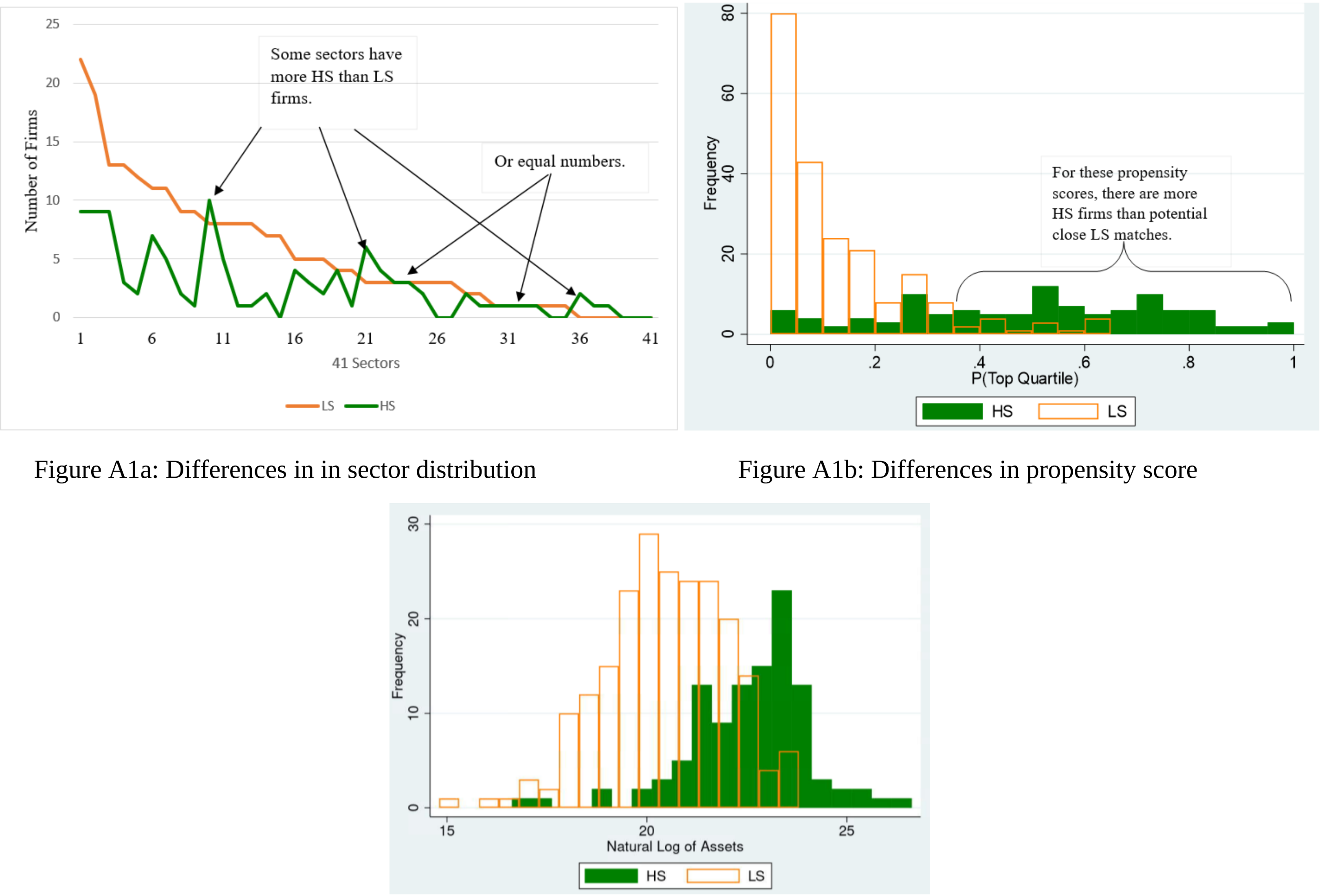
Table A1 shows that for EIS's reported caliper (0.01), I estimate that 88 *HS* & *LS* firms could be matched about once in 10^{53} samples or more. Even at 10 times the reported caliper (0.1), 88 matches would occur in fewer than one in a million samples. Only with a caliper at 0.5 does matching 88 become commonplace, but matching provides little benefit at this level.

My simulation conforms well with my own experience. Across my simulations, predictions for the number matched (caliper <0.01) range from 12.7 to 15.18. Using real data, I match fewer (9), but I have fewer *LS* firms available to match and some performance loss is expected when one moves away from ideal conditions.

Table A1: Expected matches for different conditions

Industry distribution	Propensity distribution	Caliper	m(matches) in 5000 trial samples	Prob of matching 88 in a given sample	N samples needed to expect 1 match of 88
Uniform	Beta(2,2)	0.01	12.7	4.89E-72	2.10E+71
		0.1	63.1	1.02E-11	9.70E+10
		0.25	81.6	7.93E-03	126.7
		0.5	87.8	0.621	1.6
Beta(2,2)	Beta(2,2)	0.01	14.8	5.46E-66	3.00E+65
		0.1	66.0	4.14E-10	2.50E+09
		0.25	82.2	1.35E-02	74.6
		0.5	87.5	0.545	1.8
Beta(1,3)	Beta(2,2)	0.01	20.5	8.96E-54	1.50E+53
		0.1	71.6	3.49E-07	2.90E+06
		0.25	83.7	4.38E-02	22.5
		0.5	87.5	0.545	1.8
Beta(1,3)	H:Beta(2,2) L: Beta(1,3)	0.01	15.2	4.40E-65	2.30E+64
		0.1	55.2	1.32E-16	7.50E+15
		0.25	76.0	4.20E-05	23813.6
		0.5	87.4	0.513	1.9

Figure A1: Practical barriers to matching High and Low Sustainability companies.



Some sectors have more HS than LS firms.

Or equal numbers.

For these propensity scores, there are more HS firms than potential close LS matches.

Figure A1a: Differences in in sector distribution

Figure A1b: Differences in propensity score

Figure A1c: *HS* firms tend to have more assets

APPENDIX B – ANALYSIS OF ANNUAL ACCOUNTING VARIABLES

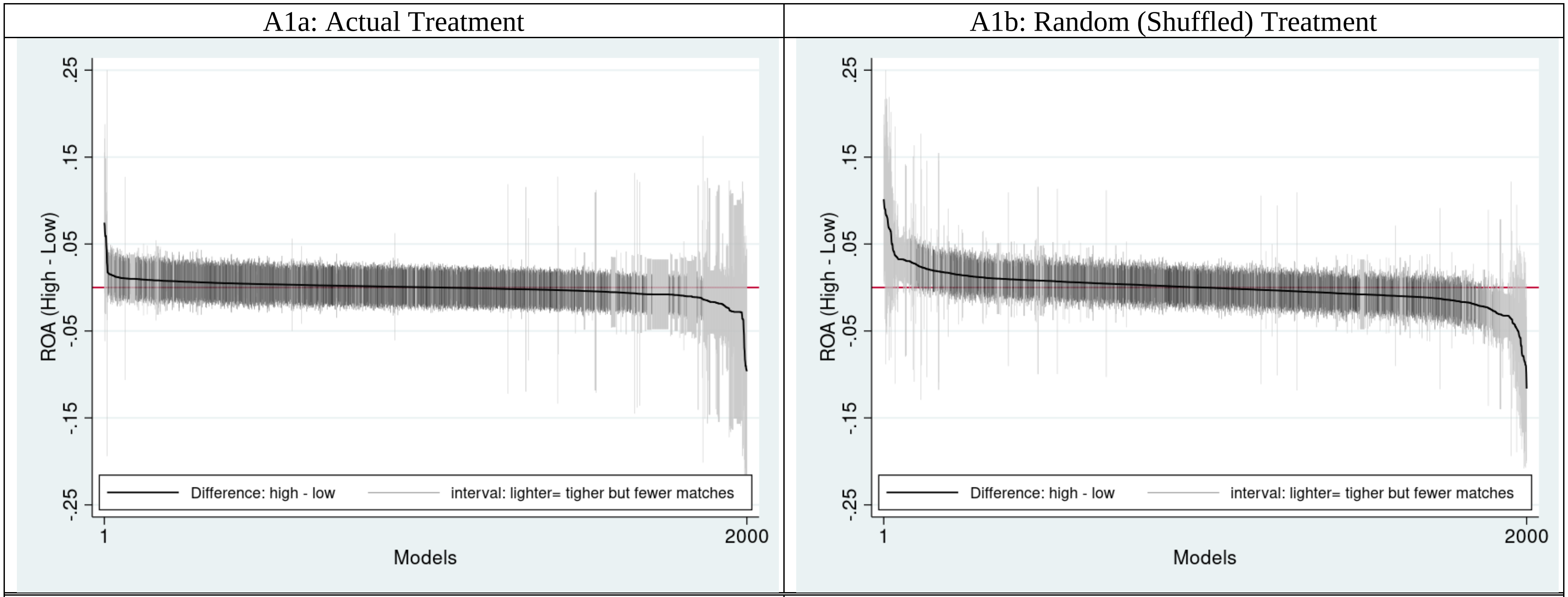
Appendix B – Table 1

	ROA				ROE			
	Equal		Value		Equal		Value	
	1	2	3	4	5	6	7	8
	Actual	Random	Actual	Random	Actual	Random	Actual	Random
B(5%)	-0.0169	-0.0274	-0.0088	-0.0159	-0.0559	-0.0691	-0.0469	-0.0565
B(median)	0.0003	-0.0001	0.0006	0.0000	-0.0152	0.0007	-0.0154	0.0010
B(95%)	0.0098	0.0271	0.0097	0.0164	0.0243	0.0647	0.0194	0.0569
#B >0	1037	990	1024	870	491	1013	405	884
#B > 0*	2	139	39	231	4	73	0	104
#B < 0*	0	133	1	242	39	79	12	97
#Br>B	917		897		1248		1199	
%B >0	51.85%	49.50%	54.47%	50.06%	24.55%	50.65%	21.54%	50.86%
%B > 0*	0.10%	6.95%	2.07%	13.29%	0.20%	3.65%	0.00%	5.98%
%B < 0*	0.00%	6.65%	0.05%	13.92%	1.95%	3.95%	0.64%	5.58%
%Br>B	46.85%		51.61%		62.40%		68.98%	
N	2000	2000	1880	1738	2000	2000	1880	1738

Note: EIS do not provide a coefficient estimate for this outcome variable so I cannot calculate #B>$B_{EIS.}$

Figure B1: ROA for high vs low sustainability portfolios

Figure B1 a & b: Differences in **equal-weighted** ROA



	A1a: Actual Treatment	A1b: Random (Shuffled) Treatment
Positive	51.85%	49.5%
Positive CI ni 0	0.1%	6.95%
Negative	48.15%	50.5%
Negative CI ni 0	0.0%	6.65%

Figure B1 c&d: Differences in **value-weighted** ROA

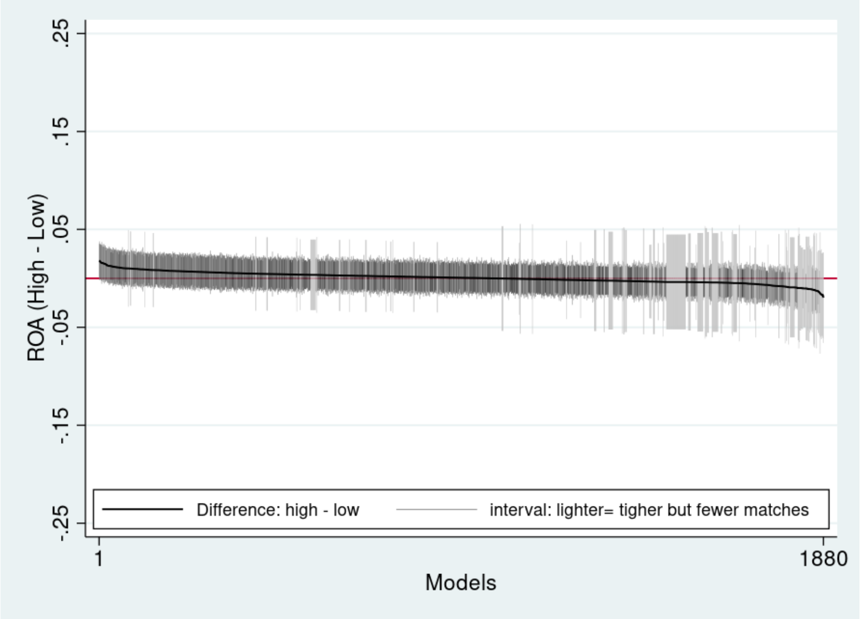
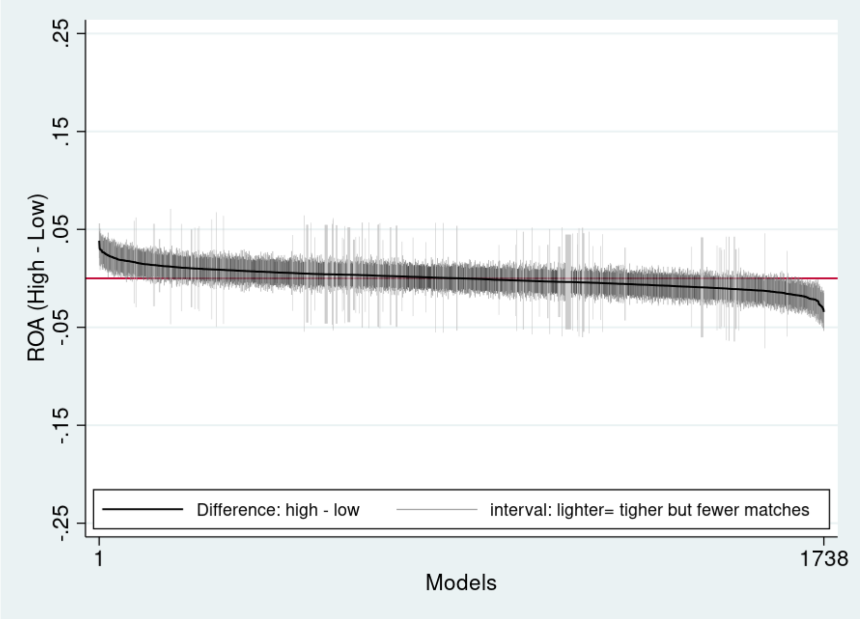
	A1c: Actual Treatment	A1d: Random (Shuffled) Treatment
		
Positive	54.47%	50.06%
Positive CI ni 0	2.07%	13.29%
Negative	45.53%	49.84%
Negative CI ni 0	0.05%	13.92%

Figure B1 e&f: Differences in **equal-weighted** ROE

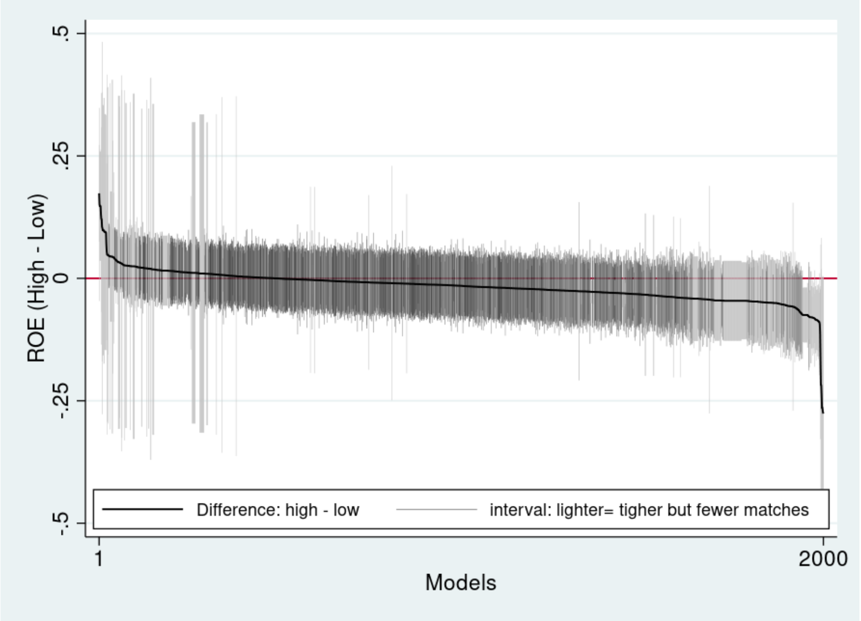
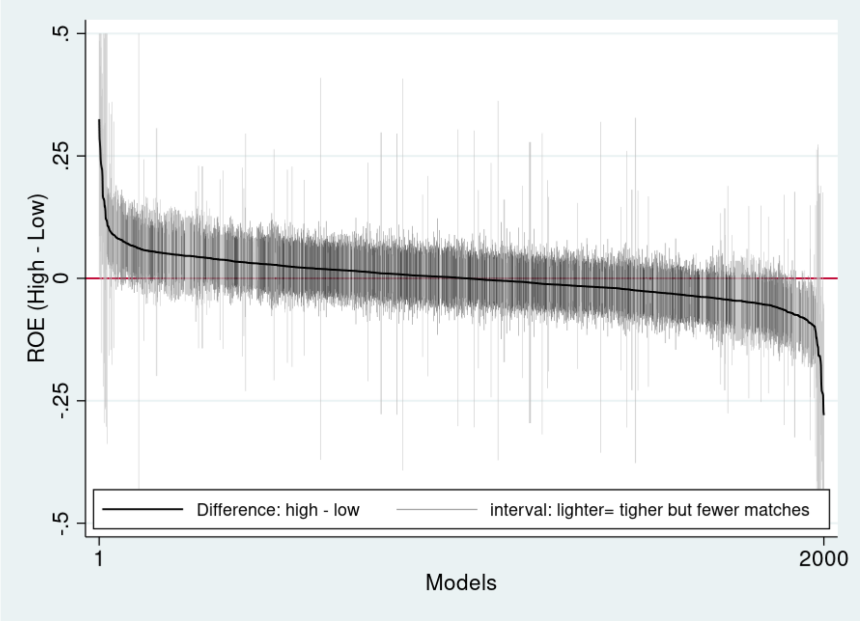
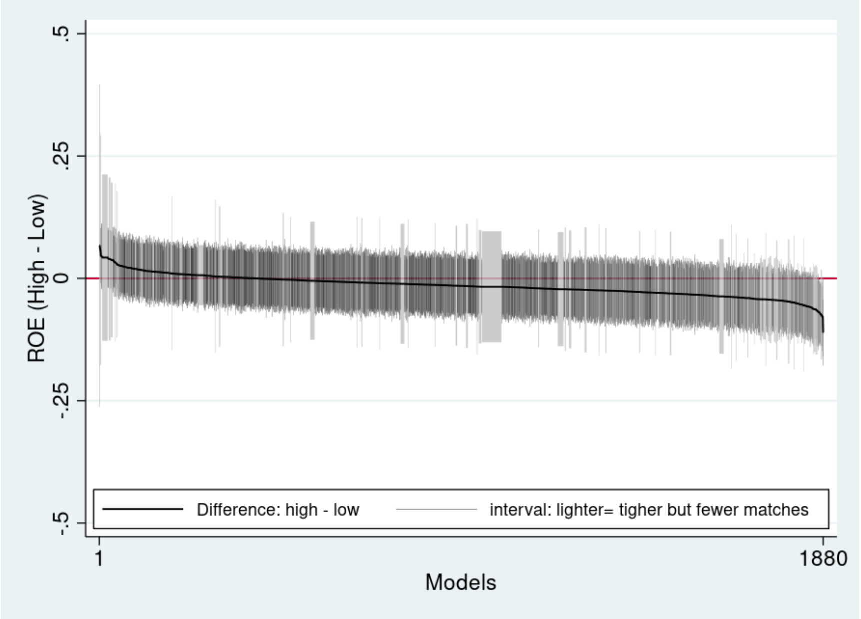
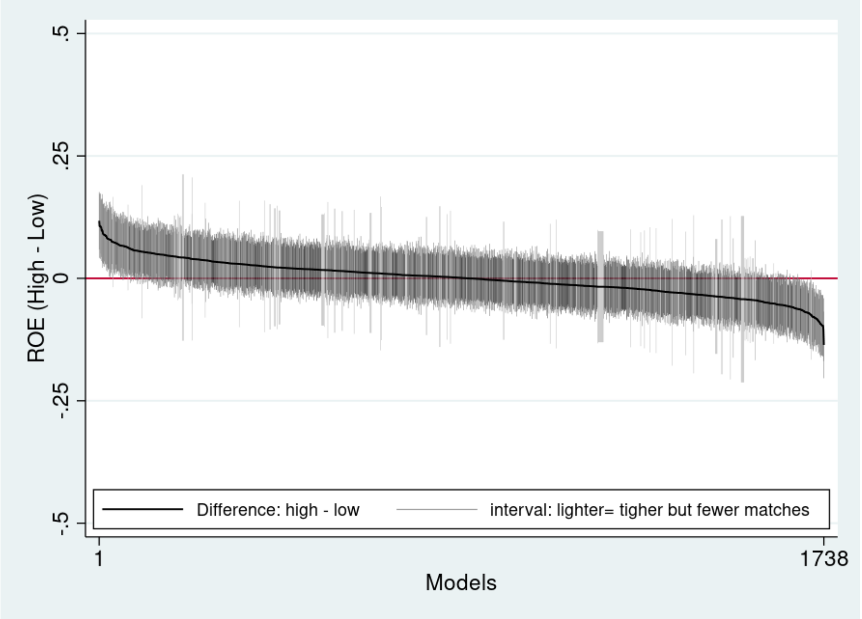
	A1e: Actual Treatment	A1f: Random (Shuffled) Treatment
		
Positive	24.55%	50.65%
Positive CI ni 0	0.2%	3.65%
Negative	75.4%	49.45%
Negative CI ni 0	1.95%	3.95%

Figure B1 g&h: Differences in **value-weighted** ROE

	A1g: Actual Treatment	A1h: Random (Shuffled) Treatment
		
Positive	21.54%	50.86%
Positive CI ni 0	0.0%	5.98%
Negative	78.46%	49.14%
Negative CI ni 0	0.64%	5.58%

SUPPLEMENT 1: PROGRAM FOR SIMULATING MATCHING

```
/************************************************************************
**
**  This stata program simulates a matching process where 90 treated firms
**  and 269 possible control firms are distributed across 41 sectors.
**  It varies the caliper distributions of the allocation
**  to sectors and the propensity scores.  It also varies the caliper
**  from 0.01 to 0.5
**
*************************************************************************/
program drop _all
program define m_sim
        global bfile = "match_sim2"
        postfile $bfile str20 sector str20 HS_dist str20 ls_dist calip p_nm p_sd str20 binom_88
using "match_sim_all.dta",replace
global sector =""
global HS_dist = ""
global ls_dist = ""
        forvalues j=1(1)4{

                        **set up conditions for the different simulations
                        if `j' == 1{
                                global sector = "runiform()"
                                global HS_dist = "rbeta(2,2)"
                                global ls_dist = "rbeta(2,2)"
                        }
                        if `j' == 2{
                                global sector = "rbeta(2,2)"
                                global HS_dist = "rbeta(2,2)"
                                global ls_dist = "rbeta(2,2)"
                        }
                        if `j' ==3 {
                                global sector = "rbeta(1,3)"
                                global HS_dist = "rbeta(2,2)"
                                global ls_dist = "rbeta(2,2)"
                        }

                        if `j' ==4 {
                                global sector = "rbeta(1,3)"
                                global HS_dist = "rbeta(2,2)"
                                global ls_dist = "rbeta(1,3)"
                        }
                forvalues k=1(1)4{
                        clear
                        ** set the 5000 simulations and give each an id
                        set obs 5000
                        gen run_id = _n
                        **add the 90 treated firms and give each an id
                        gen top = 1
                        expand 90
                        bys run_id: gen top_firm_id =_n
                        gen top_p_score = $HS_dist
                        hist top_p_score

                        **allocate them to 41 sectors, varying the distributions
                        gen temp = $sector
                        replace temp = temp*100
                        replace temp = temp*41/100
                        gen int sect = trunc(temp)+1
                        sum sect
                        drop temp
                        replace top_p_score = top_p_score + sect
                        save "top.dta",replace

                        **Repeat for 269 possible controls (Low Sustainability)
                        dis "before clear"
                        clear
                        set obs 5000
                        gen bot = 1
                        gen run_id = _n
                        expand 269
                        bys run_id: gen bot_firm_id =_n
                        gen bot_p_score = $ls_dist
                        hist bot_p_score
                        gen temp = $sector
                        replace temp = temp*100
                        replace temp = temp*41/100
                        gen int sect = trunc(temp)+1
                        sum sect
                        drop temp
                        replace bot_p_score = bot_p_score + sect
                        save "bot.dta",replace
                        **combine the two.
                        append using "top.dta"

                        **set the caliper for matching
                        if `k'==1 {
                                gen caliper = .01
                        }
                        else if `k'==2 {
                                gen caliper = .1
                        }
                        else if `k'==3 {
                                gen caliper = .25
                        }
                        else {
                                gen caliper = .5
                        }
                        gen matched=0
                **For each treated firm try to find a control with pscore within the caliper
                forvalue i = 1(1)90{
                        dis "run " `i'
                        *create p_score of treated firm to match
                        quietly{
                        gen temp = top_p_score if top_firm_id == `i' & top == 1
                        *spread the pscore around to that run_id
                        egen p_score_to_match = max(temp),by(run_id)
                        capture drop temp
                        *calculate the difference to all controls
                        gen dif_value_n = abs(p_score_to_match-bot_p_score)
                        *prevent any matches with other top firms
                        replace dif_value_n=. if top ==1
                        *sort so ordered by difference
                        gsort bot run_id  matched dif_value_n
                        quietly bys bot run_id: gen order = _n
                        **find the best match
                        replace matched =1 if order ==1 & dif_value_n<= caliper
                        **remove the matched control firm from the sample
                        replace bot_p_score=. if  matched == 1
                        drop p_score_to_match dif_value_n  order
                        }
                }
                **calculate how many not matched per simulation
                egen tot_match = sum(matched),by(run_id)
                        **report
                sum tot_match if top ==1
                local p_nm = 90 - r(mean)
                local p_sd = r(sd)
                local calip = caliper[1]
                local binom_88 =  (1/binomial(90,2,`p_nm'/90))
                dis `binom_88'
                post $bfile ("$sector") ("$HS_dist") ("$ls_dist") (`calip') (`p_nm') (`p_sd')
("`binom_88'")
                }
        }
postclose $bfile
end
```

SUPPLEMENT 2: THE MEANING AND FEASIBILITY OF CUMULATIVE ROA OR ROE.

EIS reports calculating “cumulative” ROA and ROE, and it appear this indicates the use of a compound formula to calculate cumulative ROA and ROE[15] similar to the one commonly used for stock returns.

$$Cumulative\ Stock\ Return = \prod_{0}^{T}(r_t + 1)$$

This measure works for stock return because it can be simplified mathematically to the value of an invested *V* at the end of the time *T* period over the value at the beginning.

$$r_t + 1 = \frac{V_t - V_{t-1}}{V_{t-1}} + 1 = \frac{V_t - V_{t-1} + V_{t-1}}{V_{t-1}} = \frac{V_t}{V_{t-1}} =$$

Substituting and expanding the product:

$$\prod\nolimits_{0}^{T}(r_t + 1) = \frac{V_T}{V_{T-1}} * \frac{V_{T-1}}{V_{T-2}} * \frac{V_{T-2}}{V_{T-3}} \ldots \frac{V_2}{V_1} * \frac{V_1}{V_0} = \frac{V_T}{V_0}$$

This simplification does not work for cumulative ROA or ROE. Assume:

$$Cumulative\ ROA\ = \prod_{0}^{T}(ROA_t + 1)$$

$$ROA_t + 1 = \frac{\pi_t}{A_{t-1}} + 1 = \frac{\pi_t + A_{t-1}}{A_{t-1}}$$

$$= \frac{\pi_T + A_{T-1}}{A_{T-1}} * \frac{\pi_{T-1} + A_{T-2}}{A_{T-2}} \ldots * \frac{\pi_1 + A_0}{A_0}$$

Net income (π) from one year do not directly influence assets in subsequent years and assets change for other reasons, so unless net income is always and completely reinvested in assets (i.e. $A_t = \pi_t + A_{t-1}\ \forall\ t$) the product continues to be a complex combination of attributes. A simple empirical analysis of the relationship between net income and the change in assets reveals that it explains only about 2% of the variance of assets. So, in practice $A_t \neq \pi_t + A_{t-1}$.

The calculation has other difficulties, because ROA (and ROE) cannot be calculated for about 1% of the observations – for example, where shareholder equity is negative, or ROA (or ROE) are < -1. This may not seem like a large number, but in an 18-year panel about 10% of the firms are affected.

For the above reasons, compounded ROA or ROE is not a common measure in accounting or finance. A review of the literature failed to uncover any precedent among FT50 journals, and interviews with leading accounting and finance faculty confirmed that the measure lacks any known construct validity.

SUPPLEMENT 3: ALTERNATIVE INTERPRETATION OF EIS SAMPLING FRAME

EIS specifies the beginning of its required data sampling frame (1993) but not its end. By construction, all firms must be alive and rated by Refinitiv in 2003-5, and the sample ends in 2010. In my main analysis, I assumed that EIS required all firms to have stock and accounting information from 1993-2010, but it is possible that EIS's sampling process only required such data from 1993-2003. I rejected this sampling frame for my main analysis because it conflicts with the EIS empirical design and seems to contradict the text. The EIS analysis is based on a system of matched pairs; if one of these pairs is censored during the analysis, the portfolios lose balance, ruining this matching process. Moreover, if right-censoring is allowed, the EIS report should discuss how it was handled. At a minimum, this would include the handling of non-censored pairs and the method for calculating terminal values. EIS also includes a discussion of a robustness test using pairwise comparisons without mentioning any censoring. Thus, I think it is most likely that EIS required existence, as I did, across the 1993-2010 sampling frame.

Nevertheless, to ensure the robustness of my inferences, I reconducted my main replication analysis using a sample frame that required operation only from 1993-2003. Doing so made more firms available for matching, but it did not change the inferences reported in the main analysis. It did, as expected, add to the noise in estimates of accounting coefficients because some firms in a balanced pair are lost to attrition, thereby unbalancing the comparison.

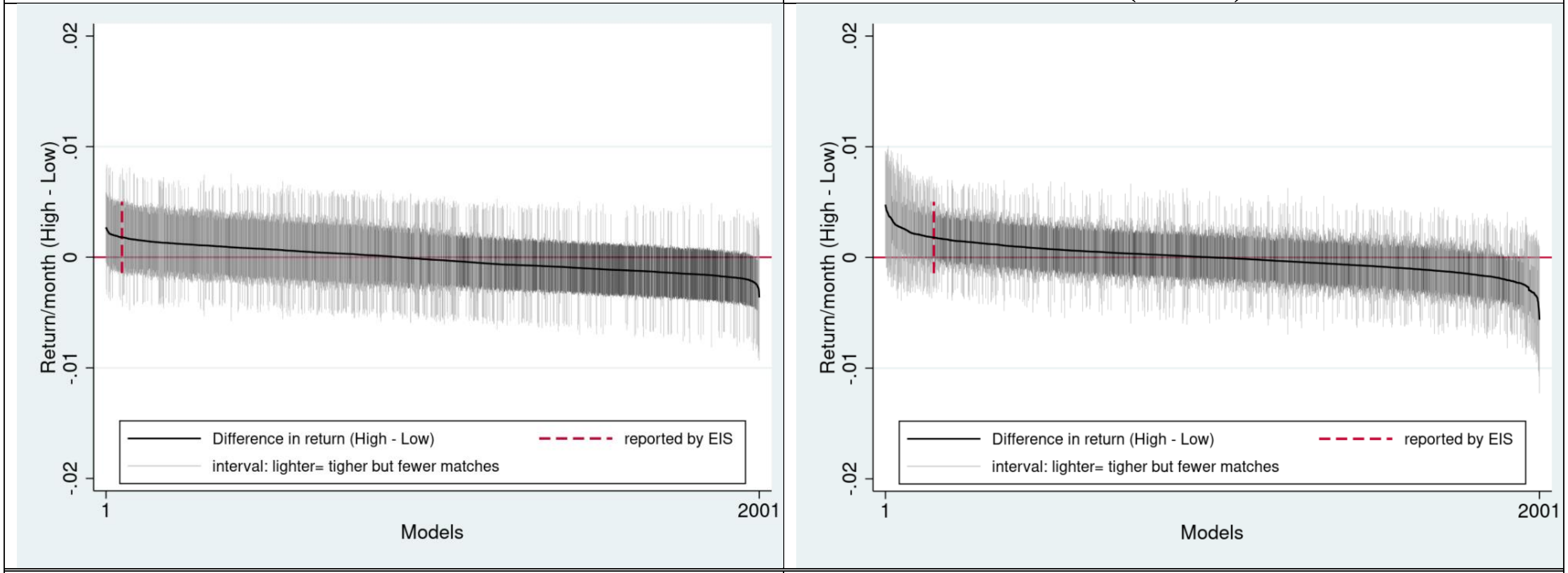
Supplement 3 Table 1: Descriptive statistics of matched pairs for alternative required sample frame (1993-2003)

Caliper	0.01		0.1		0.25		0.5		1	
Cohort	*LS*	*HS*	*LS*	*HS*	*LS*	*HS*	*LS*	*HS*	*LS*	*HS*
Assets	3.19	3.87	5.02	8.32	4.35	16.70	3.42	15.60	3.37	16.20
	(3.70)	(3.51)	(9.86)	(16.50)	(8.69)	(39.10)	(7.51)	(33.40)	(7.26)	(34.20)
Ln(Assets)	21.28	21.63	21.35	21.76	21.16	22.24	20.81	22.45	20.74	22.52
	(1.21)	(1.11)	(1.45)	(1.67)	(1.55)	(1.73)	(1.69)	(1.59)	(1.73)	(1.56)
ROA	8.06%	6.23%	7.42%	6.83%	7.14%	6.60%	7.56%	6.98%	7.48%	7.22%
	(0.05)	(0.04)	(0.08)	(0.07)	(0.08)	(0.06)	(0.09)	(0.07)	(0.09)	(0.06)
Leverage	0.52	0.58	0.59	0.59	0.58	0.61	0.56	0.61	0.56	0.61
	(0.19)	(0.09)	(0.32)	(0.14)	(0.30)	(0.15)	(0.29)	(0.16)	(0.29)	(0.16)
Turnover	1.35	1.04	1.23	1.09	1.23	1.09	1.24	1.05	1.15	1.04
	(0.86)	(0.49)	(0.95)	(0.66)	(0.90)	(0.67)	(0.86)	(0.61)	(0.78)	(0.66)
MTB	3.55	2.64	3.95	3.38	3.71	3.37	3.70	3.77	3.47	3.66
	(2.48)	(1.46)	(3.71)	(2.66)	(3.34)	(2.68)	(3.25)	(3.08)	(3.17)	(2.96)
N	For each caliper, I identify 400 pairs of 90 *HS* & *LS* firms and calculate means & std errors for each variable. I report the median value of these 400 estimates.									
Median Matches	13		37		51		71		80	
Min Matches	8		31		43		65		75	
Max Matches	17		43		58		78		85	

Supplement 3 Table 2: Summary of results from actual and randomized data from alternative required sample frame (1993-2003)

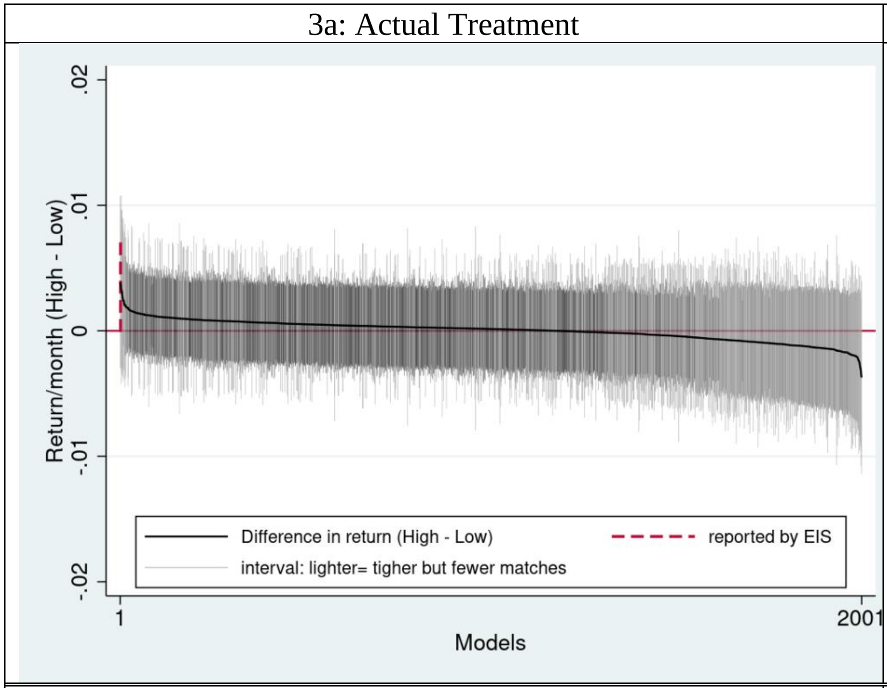
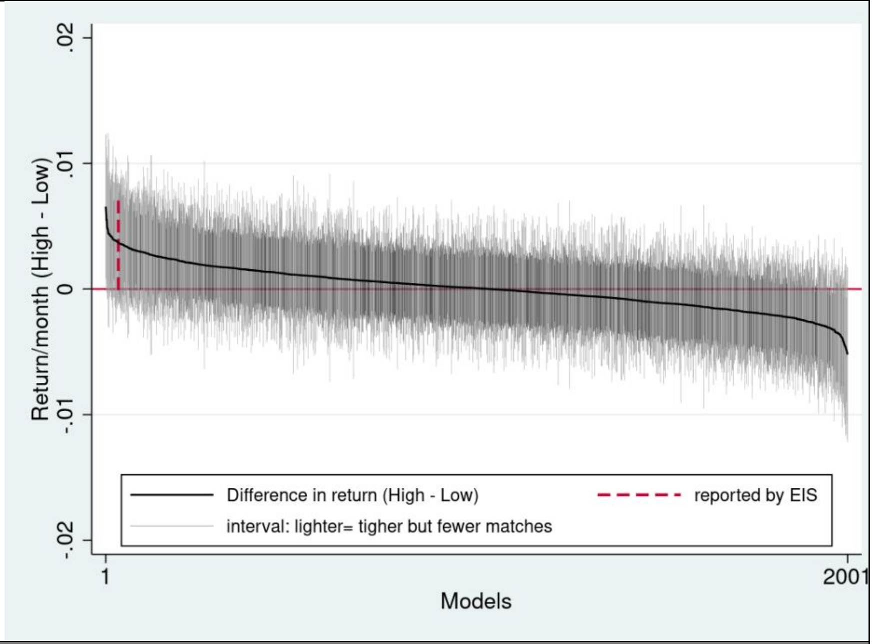
	Stock Return				**Compound ROA**				**Compound ROE**			
	Equal		Value		Equal		Value		Equal		Value	
B_{EIS}	0.0017		0.0037		0.2		2.7		6.5		6	
	1	2	3	4	5	6	7	8	9	10	11	12
	Actual	Random	Actual	Random	Actual	Random	Actual	Random	Actual	Random	Actual	Random
B(5%)	-0.0018	-0.0020	-0.0014	-0.0026	-0.1219	-0.9864	-0.1335	-0.9907	-47.0685	-173.286	-27.601	-62.3726
B(median)	-0.0002	0.0000	0.0001	0.0001	0.5274	0.0241	0.4632	0.0182	7.3518	1.2575	2.9614	0.6033
B(95%)	0.0015	0.0021	0.0011	0.0029	1.0310	0.9966	0.9577	0.9877	175.7927	156.8728	53.6715	62.3068
#B >0	901	1003	1179	1039	1818	1029	1778	1024	1440	1065	1222	1041
#B > 0*	0	19	0	13	77	46	24	41	8	21	2	18
#B < 0*	7	30	0	12	0	40	0	40	0	17	2	10
#B>B_{EIS}	0	169	0	34	1624	775	0	0	1436	1054	1029	851
#B>B_{EIS}*	0	19	0	6	77	46	0	0	8	21	2	18
#Br>B	1018		968		460		525		750		880	
%B >0	45.05%	50.15%	58.95%	51.95%	90.90%	51.45%	88.90%	51.20%	72.00%	53.25%	61.10%	52.05%
%B > 0*	0%	0.95%	0%	0.65%	3.85%	2.30%	1.20%	2.05%	0.40%	1.05%	0.10%	0.90%
%B < 0*	0.35%	1.50%	0%	0.60%	0%	2.00%	0%	2.00%	0%	0.85%	0.10%	0.50%
%B>B_{EIS}	0.00%	8.45%	0.00%	1.70%	81.20%	38.75%	0.00%	0%	71.80%	52.70%	51.45%	42.55%
%B>B_{EIS}*	0%	0.95%	0%	0.30%	3.85%	2.30%	0%	0%	0.40%	1.05%	0.10%	0.90%
%Br>B	50.90%		48.40%		23.00%		26.50%		37.50%		44.00%	

Supplement 3 Figure 1: Stock returns for *HS* vs *LS* portfolios (using data from alternative required sample frame)
Figure 1a&b: Differences in returns from **equal** weighted portfolios



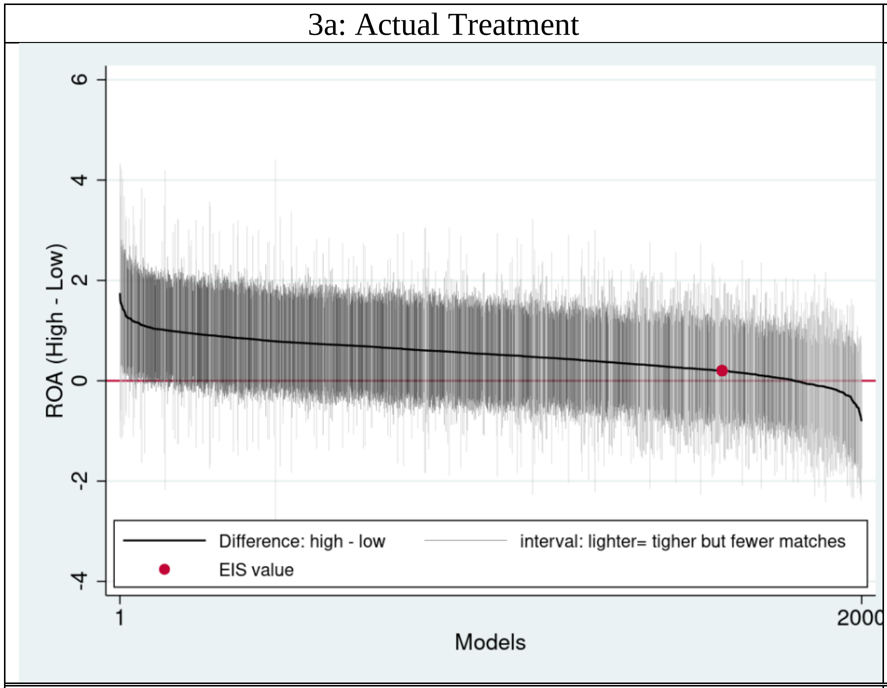
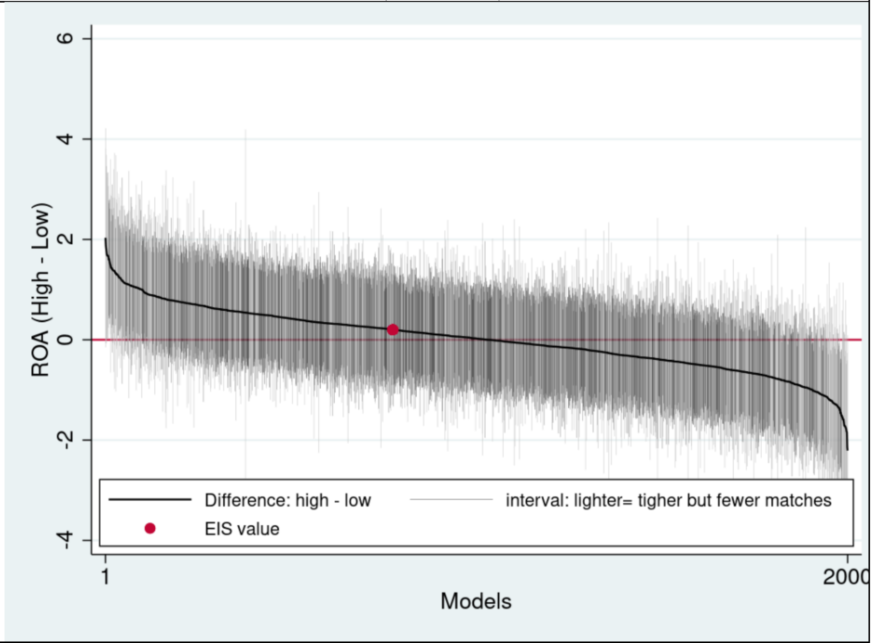
	3a: Actual Treatment	3b: Random (Shuffled) Treatment
Positive	45.05%	50.15%
Positive CI ni 0	0%	0.95%
Negative	54.95%	49.85%
Negative CI ni 0	0.35%	1.50%

Supplement 3 Figure 1c&d: Differences in returns from **value** weighted portfolios

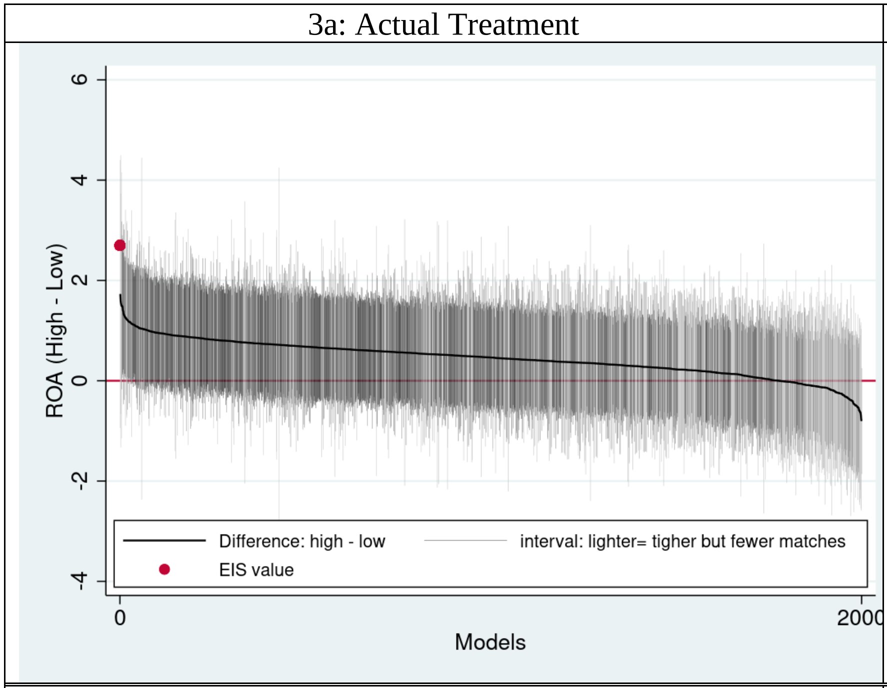
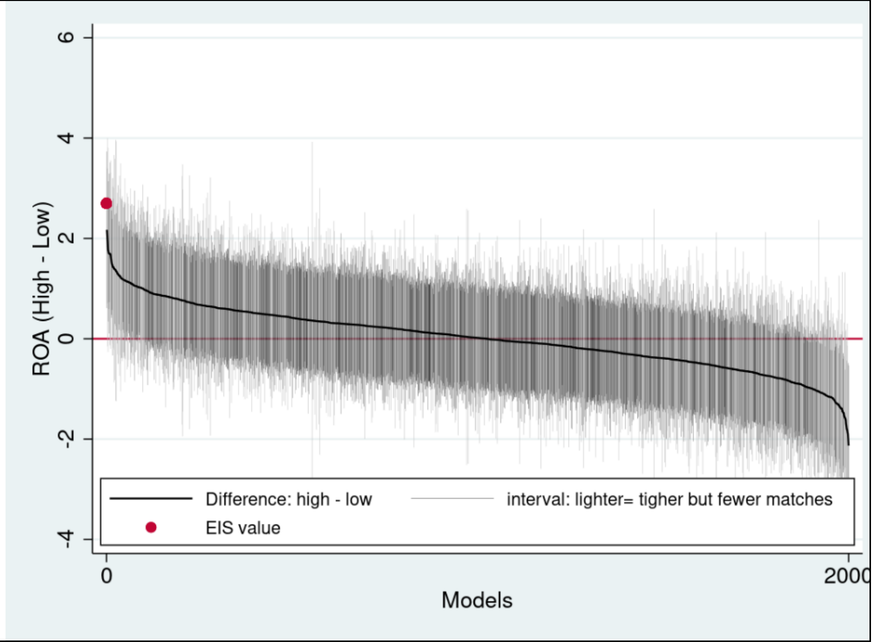
	3a: Actual Treatment	3b: Random (Shuffled) Treatment
		
Positive	58.95%	51.95%
Positive CI ni 0	0%	0.65%
Negative	41.05%	48.05%
Negative CI ni 0	0%	0.60%

Supplement 3 Figure 2: Compound ROA for high vs low sustainability portfolios (using data from alternative required sample frame)

Figure 2 a&b: Differences in **equal** weighted compound ROA

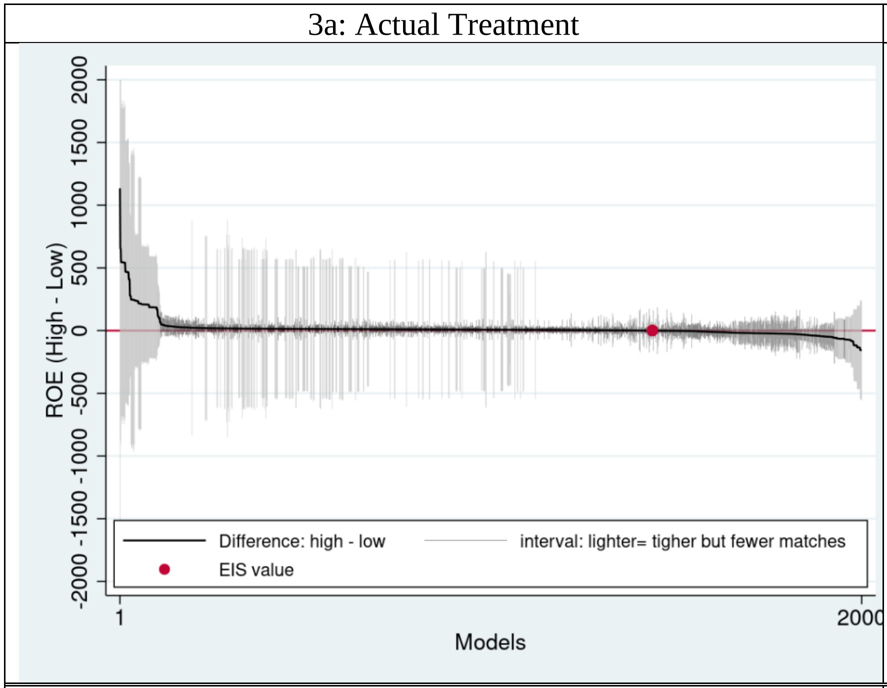
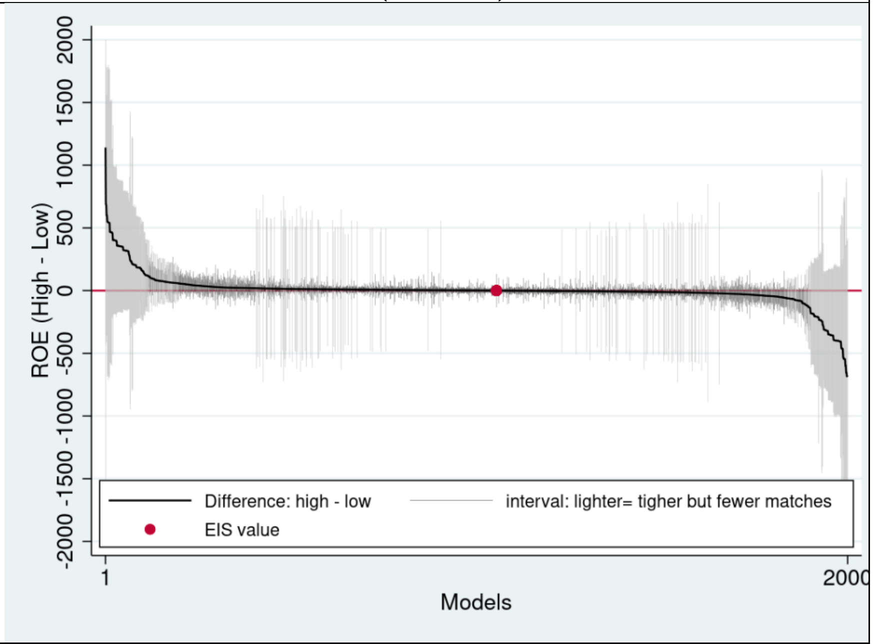
	3a: Actual Treatment	3b: Random (Shuffled) Treatment
		
Positive	90.90%	51.45%
Positive CI ni 0	3.85%	2.30%
Negative	9.10%	48.55%
Negative CI ni 0	0%	2.00%

Supplement 3 Figure 2 c&d: Differences in **value** weighted compound ROA

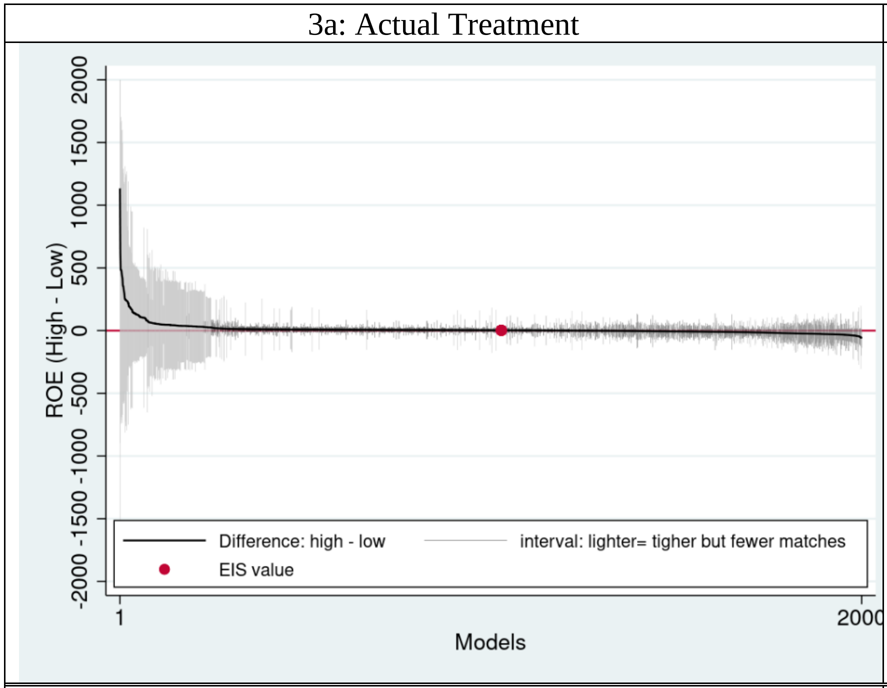
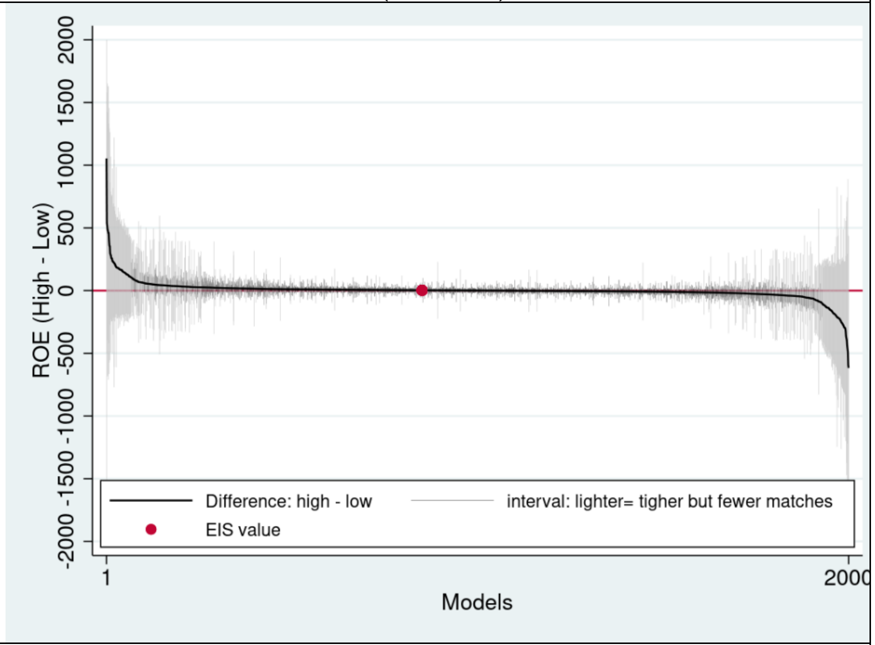
	3a: Actual Treatment	3b: Random (Shuffled) Treatment
		
Positive	88.90%	51.20%
Positive CI ni 0	1.20%	2.05%
Negative	11.10%	48.80%
Negative CI ni 0	0%	2.00%

Supplement 3 Figure 3: Compound ROE for high vs low sustainability portfolios (using data from alternative required sample frame)

Figure 3 a&b: Differences in **equal** weighted compound ROE

<table>
<tr><th colspan="2">3a: Actual Treatment</th><th>3b: Random (Shuffled) Treatment</th></tr>
<tr><td colspan="2">
</td><td>
</td></tr>
<tr><td>Positive</td><td>72.00%</td><td>53.25%</td></tr>
<tr><td>Positive CI ni 0</td><td>0.40%</td><td>1.05%</td></tr>
<tr><td>Negative</td><td>28.00%</td><td>46.75%</td></tr>
<tr><td>Negative CI ni 0</td><td>0%</td><td>0.85%</td></tr>
</table>

Supplement 3 Figure 3 c&d: Differences in **value** weighted compound ROE

3a: Actual Treatment		3b: Random (Shuffled) Treatment
		
Positive	61.10%	52.05%
Positive CI ni 0	0.10%	0.90%
Negative	38.90%	47.95%
Negative CI ni 0	0.10%	0.50%

ENDNOTES

[1] As of this printing, the published paper and its penultimate NBER working paper have been cited over 5000 times.
[2] Terminology for "corporate sustainability" varies with time, discipline, and author preference. Scholars refer to closely related ideas using a variety of terms, including "corporate social responsibility/performance" or "environmental-social-governance (ESG) performance".
[3] As a result of this replication, the authors published an Erratum Corrige correcting a statistical test of unweighted portfolio returns and the calculation method and reported estimates for weighted portfolio returns.
[4] In an analysis of the top 50 articles citing EIS, 41 (82%) cite its finding of a financial connection.
[5] Because Refinitiv updates and corrects past records, my Refinitiv data may vary slightly from that used by EIS.
[6] Refinitiv data for the period include 118 companies in finance and insurance (ICB supersector: 3010, 3020, and 3030), suggesting that EIS included 18 of these in their sample, but which 18? I guessed that EIS retained "financial data providers," "mortgage REITs," and reinsurance companies).
[7] In robustness tests, I consider alternative sampling criteria requiring firms to exist from 1993-2003 (rather than 2010). See Supplement 3.
[8] Disqualifying events included being founded after 1993, acquired, merged, failed, or taken private before the end of 2010. Such events were not identified for 24 firms, but these firms lacked accounting data in Compustat and Worldscope data for the full frame.
[9] In a recent Erratum Corrige, the authors confirmed this interpretation.
[10] Reviewers suggested another interpretation of the report: EIS matched with replacement. If I follow this protocol (and no caliper requirement), I can indeed match 90 HS firms, but because some LS firms are used more than once, not all HS firms have a unique pairing, and this seems to contradict the EIS report. Nevertheless, I also performed my full analysis for these matches and again confirmed the inferences reported here. Results were submitted for review and are available on request.
[11] Significance calculations are predicated on the perfect repeatability of sampling, measurement, and analysis.
[12] The 5-95% interval cannot be calculated for the single samples reported in columns 1 and 2.
[13] Note that none of these estimates are based on 88 or 90 pairs as reported by EIS.
[14] As noted earlier, EIS reports using 200 interviews to check conformance with this assumption.
[15] The unpublished penultimate draft of the paper includes graphs showing an evident compounding effect in the measure of cumulative ROA and ROE

ACKNOWLEDGEMENTS: I would like to thank the anonymous reviewers at JOMSR for their guidance. I would also like to thank Kenneth French and Myles Shaver for their assistance. This publication would not have been possible without the patience and insight of Xavier Martin.