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 than their low-sustainability counterparts? An extremely influential publication, "The Impact of Corporate Sustainability on Organizational Processes and Performance"**,** claims that they do. Its 2014 publication preceded a boom in sustainable investing, and both scholars and practitioners have used it to explain these investments. Yet I report here that I cannot replicate the original study's methods or results, and I show that a close reading of the original report reveals its evidence is too weak to justify its claims concerning financial performance. In concluding remarks, I discuss the importance of my findings for both sustainable business and the practical accretion of knowledge.

August 4, 2024
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 review and replicate an exceptionally influential publication: "The Impact of Corporate Sustainability on Organizational Processes and Performance" by Robert Eccles, Ioannis Ioannou, and George Serafeim (hereafter EIS). Its most influential empirical claim is that "high sustainability companies" have higher stock returns and accounting performance than their "low sustainability" counterparts. Published in 2014, EIS appeared to provide definitive evidence of a link between sustainability, organization form, and financial performance.

EIS was not the first (or last) to investigate the connection between corporate social performance (variously labeled as Corporate Social Responsibility, Environmental-Social-Governance, or Corporate Sustainability) and corporate financial performance. Dozens of previous studies had evaluated the connection and reported a diverse range of results – positive, negative, curvilinear, moderated, mediated, and NULL effects (Barnett & Salomon, 2012; Hillman & Keim, 2001; Hull & Rothenberg, 2008; McWilliams & Siegel, 2000; Waddock & Graves, 1997). Reviews of the literature and meta-analysis also resulted in uncertain conclusions (Aguinis & Glavas, 2012; Henisz, Dorobantu, & Nartey, 2014; Horváthová, 2010; Margolis & Walsh, 2003; Orlitzky, Schmidt, & Rynes, 2003.). Summarizing the state of research in 2009, Marc Orlitzky concluded "the empirical evidence is too mixed to allow for any firm conclusions."

EIS used alternative methods and data to clarify the debate. It employed ESG measures from a relatively new rating agency (Refinitiv), a modern identification strategy (propensity score matching), and methods of portfolio analysis from the finance literature. Scholars remarked on both its results and methods. In their review of the literature, Aragon-Correa,

Marcus, Rivera, & Kenworthy (2017) concluded that EIS delivers "[O]ne of the strongest relationships between corporate environmental performance and corporate financial performance." Other scholars praised its use of matched pairs and its longer data panel (Aguilera et. al, 2021). EIS quickly became a touchstone for the research community, and as of today it has been cited more times than any contemporary or subsequent article in *Management Science*

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 the assessment of corporate sustainability has become "a core risk management strategy for portfolio construction" (Lee, 2020). EIS also has been used in government lobbying efforts and testimony before the U.S. Senate (Crasi, 2015; Blake, 2020).

Despite this outsized importance, no one has replicated EIS's analysis of financial performance, and yet replication is a critical step in the validation and refinement of management theory (Kohler and Cortina, 2023). In this article, I fill this empirical gap by replicating a key part of EIS – the investigation of financial performance. In attempting to do so I must overcome several difficulties – a key analytical procedure cannot be replicated, and the construction of four outcome variables is not described. To bracket the range of possible replications, I construct hundreds of portfolios of high and low-sustainability firms and compare their performance using different methods. I am unable to confirm the estimates in EIS and instead find no reliable

evidence for a connection between sustainability and financial performance. I then revisit EIS and show that its reported evidence also provides an insufficient basis to conclude that high-sustainability firms have better financial performance.

I set the stage for my replication by reviewing the original study and discussing its importance. I then focus attention on the design and application of the original analysis. Given the many uncertainties uncovered, I widen the aperture of my analysis to include a multiverse of possible models and comparisons. I then reconcile my replication and the original study by reconsidering the original analysis. Finally, I briefly discuss the implication of my research for theories of sustainable business and the accumulation of knowledge on business management.

OVERVIEW OF THE ORIGINAL REPORT

Eccles, Ioannou, and Serafeim (2014) does not report 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 might provide sustained competitive advantage (Hart, 1995; Greening & Turban, 2000; Casadesus-Masanell & Ricart, 2012). Hundreds of studies have tried to test this conjecture, with mixed results.

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 adopted more of Refinitiv's "sustainability policies" (see Appendix A for description) measure and reward executives differently, engage stakeholders more systematically, and disclose more non-financial information. It then tests whether these

sustainability policies are linked to higher financial performance. EIS reports using six outcome measures in these tests - two measures of stock return and four measures of accounting return. Based on these tests, EIS advances the conclusion 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). It is this claim that has proven most influential, and it is the focus of my replication.

The EIS empirical strategy

EIS's analytical strategy is based on a simple but powerful idea. If a treatment (e.g. adoption of sustainability policies) can be randomly assigned, then the difference in outcomes of the two groups (treated and control) will measure the effect of the treatment (Rubin, 1974). Unfortunately, for many research questions, random assignment is not possible, and for these situations, scholars have developed methods designed to match treated subjects to control subjects that are so similar that it is "as if" the treatment had been randomly assigned (Rosenbaum & Rubin, 1983). The analysis in EIS is built around the idea that matching will allow "as if" randomization: all of their statistical tests compare the means for the treated and control groups, and if these groups differed on other dimensions, these other factors could explain the observed differences.

To evaluate whether sustainability is connected to financial return, EIS compares the means of three different outcome variables (stock return, return on assets, and return on equity) using two weighting criteria when forming estimates (equal and market-value-weighted). Based on these six comparisons (3 measures X 2 weightings), EIS concludes that: "high sustainability companies significantly outperform their counterparts over the long term, both in terms of stock market and accounting performance" (p. 2835).

REPLICATING EIS'S ANALYSIS OF FINANCIAL PERFORMANCE

Replication Strategy

Replicating an empirical study can be challenging, because few reports provide the detail needed to conduct every step of the analysis. Indeed, "a lack of detail" is commonly cited as the main barrier to conducting replication studies (Bloomfield, Rennekamp, and Steenhoven, 2018). To clarify the details of EIS's method, I sent multiple emails to its authors, but they did not respond[1]. Thus, I was forced to use uncertainty analysis to calculate estimates from a multiverse of ways EIS might have conducted the analysis.

Insert Figure 1

Figure 1 shows the process I used in my replication. I first tried to replicate the sample as closely as possible, but found I could not identify the exact groups used in the EIS analysis. Thus, I had to design a method to allow testing of a range of possible groups. Then, I tried to replicate the matching method but found that the report could be interpreted in different ways. Consequently, I again created a method covering a range of approaches that might have been implied by the text. I then tried to replicate the measures used in EIS but discovered that some could not be calculated. Based on a mathematical analysis and literature search, I concluded that the reported measures were uninterpretable and unprecedented. I thus substituted more

[1] I first sent the original authors an email complimenting them on the influence of their publication, explaining my efforts to understand their manuscript, describing my attempts to replicate it, and asking for assistance. I later contacted them after completing a more polished draft, asked for comments, and expressed an openness to cooperating in a joint statement. I sent a third message asking for specific information about their calculation of cumulative accounting measures. None of the authors responded to any of my messages.

common methods. Ultimately, having created different methods for mapping the uncertainty about the treated group and the matching method, I analyzed a multiverse of possible models.

My first discovery was that my matching method did not deliver the performance described by EIS, and hence I had to investigate the source of the difficulty. I simulated matching under a variety of idealized conditions and calculated that the matching success reported in EIS is extremely improbable (see Appendix C). Thus, I continued with my analysis of the financial performance of 1600 alternatively matched groups of treated and control firms.

I analyzed my estimates both graphically and by comparison with estimates from samples where the NULL hypothesis was known to be true (Simonsohn, Simmons & Nelson, 2020). Finally, as an extension and robustness test, I conduct an analysis using a method that does not require matching.

In total, my results differ substantially from those reported by EIS. In an attempt to harmonize the two analyses, I revisited the evidence presented in the EIS report. I found that a critical significance test appears to have been miscalculated. I also found that no test was performed for four of the six tests on the connection between sustainability and financial performance. Thus, I conclude that both my evidence and that of the original report do not support the empirical claim advanced by the report, 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).

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. I began my replication by accessing Refinitiv data and cleaning it of duplicate reports.

Consistent with EIS, I identified 775 US firms with data in the window 2003-2005 (see Figure 2). EIS's second step is the removal of 100 banking, finance, and insurance companies. Here, a lack of reporting detail made it difficult to match EIS's method, but after some consideration, I too was able to remove 100 firms in finance, leaving me with 675[2].

Insert Figure 2

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, or how many firms were removed for failure to match the required sample frame (1993-2010). Out of caution, I used both Compustat and Worldscope business accounting data. Using both, I was able to match more than 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.

The next sample selection criterion is that firms must have continued existence and report accounting information for the entire 1993-2010 sample frame. Failure to meet this requirement caused me to lose 194 firms – a loss of 31%. Fearful that I was losing firms in error, I conducted additional analysis. Using CRSP data and Amazon Turks workers, I was able to find evidence that 170 of these 194 lost firms (88%) had an evident disqualifying event, such as being founded after 1993, or being acquired, merged, failed, or taken private before the end of 2010.

EIS reports losing 28% of its *Low Sustainability* group because of failure to match the sample frame, but it does not report losing any of its *HS* group for this reason. Instead, it reports

[2] 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).

that 200 interviews were conducted to gather histories of the *High Sustainability* firms and that based on these interviews, 46% (78 firms) of the 168 firms were disqualified. EIS does not report how these interviews were conducted or processed, so it is possible that some firms were removed for failure to match the 1993-2010 frame.[3]

Once my sample was complete, I placed the remaining firms into quartiles based on the same Refinitiv score used by EIS. This left me with 109 firms in the *High Sustainability* group (top quartile), and 218 in the *Low Sustainability* group (bottom two quartiles). EIS reports identifying a group of 90 *HS* firms, but as discussed, EIS reports some firms were removed following interviews with executives. If so, this could explain my larger group and suggest that EIS's 90 is a subset of my 109. Without access to the interview data, I cannot be sure which 18 firms to remove. The probability of selecting the correct 90 (from 109) is very small in a single draw ($< 3.3x10^{-8}$), but the probability of getting 90% correct is much larger (0.017). If I conduct 400 random draws from my candidate group of 108 *HS* firms, I have a 99.99% chance that one or more of my portfolios will be at least 90% correct (i.e., 81 of the 90 firms will match EIS's sample). Thus, I decided to conduct my analysis using 400 random draws of 90 firms from my 109 candidates.

Replicating the Matching Method

EIS reports using logistic regression to obtain propensity scores for firms in the sample to be in the high sustainability quartile. The equation for this logistic regression appears to be:

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

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), ROA, asset turnover

[3] A large difference in attrition between the HS and LS groups would raise empirical concerns, so I think EIS's historical analysis removed a similar percentage of firms for failure to match the sampling frame.

(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)" (pg. 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)" (pg. 2837).

For matching to allow inference about average treatment effects, matched firms should be very similar. 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's study, that would mean a caliper of 0.06.

EIS provides information about 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" (pg. 2837). Unfortunately, the correct interpretation of this note is unclear. It could mean that the main analysis used 90 matched pairs (88 with a caliper of 0.01 and 2 with a larger caliper) and then additional robustness analysis was done using just the 88 tightly matched pairs, OR it could mean no caliper was used in the main analysis and then robustness analysis was conducted using 88 matches made using a caliper of 0.01. The former interpretation corresponds to best practice, but the latter is possible. Given the uncertainty about the caliper, I chose to analyze matches made using four different calipers -- 0.01, 0.1, 0.25, and 0.5. This meant that my uncertainty analysis will need to consider 400 draws of 90 top firms, matched according to four different caliper criteria.

Replicating Measures

EIS reports the use of two methods for evaluating financial performance concerning stock returns and accounting measures.

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

$$PR_t = \alpha + B * MKTRF_t + B * SMB_t + B\ HML_t + B * MTB_t + B * UMD_t + e_t \quad \text{Eq. 2}$$

The outcome variable (*PR*) is the portfolio stock return for low or high sustainability 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 for that month. 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 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 abnormal stock return for the portfolio for the average month. The equation is estimated for a panel of 216 months (*t*) for the period 1993-2010.

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

Accounting Returns. EIS reports the use of an unusual measure of accounting returns in its analysis. Rather than compare accounting performance in each year, EIS compares the "cumulative" accounting performance for the entire 1993-2010 frame. For example, EIS states "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" (pg. 2851). EIS provides no formula for this calculation, and the authors did not

respond to requests for details. To try to gain additional information on the process, I reviewed drafts and presentations of their research. The penultimate draft of EIS (with results identical to those eventually published) includes graphs revealing that accounting returns were compounded.[4] Thus, my best guess is that cumulative ROA and ROE were calculated using a formula like Equation 3.

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

where there are *i* firms in a group being considered and the calculation is done for years *t* from 0 to T. Although this formula has an appealing parallel with cumulative stock return, it cannot be calculated for all the observations in the sample (see Appendix 1) and has no clear interpretation, since there is no way to administer the hypothetical (investment of annual returns into equity or assets). A search of the literature reveals no examples of precedence for its use in top journals, and interviews with finance faculty at MIT, LBS, Harvard University, and the University of Pittsburgh failed to uncover a precedent for the measure or justification for its use.

Given the uncertainty surrounding cumulative accounting return, I chose to substitute measures and methods more consistent with the literature. With rare exceptions, accounting returns are measured and analyzed annually (Singh et al, 2023). Most commonly ROA is calculated as net income over assets or net income plus expenditures over assets. I chose the latter form. I calculate Return on Equity as net income over shareholder equity.

When predicting ROA or ROE, scholars often use covariate controls and fixed effects to try to account for some sources of unobserved heterogeneity (c.f. Waddock & Graves, 1997). Elsewhere in the analysis, EIS reports the use of several such controls: Size (log assets), Turnover (revenues/total assets), Leverage (total liabilities/total assets), MTB (market value/

[4] For example: "Investing $1 in book value of equity in the beginning of 1993 in a value-weighted (equal-weighted) portfolio of High Sustainability firms would have grown to $31.7 ($15.8) by the end of 2010."

(total assets – total liabilities), and fixed effects for the industry and year. I use these measures and fixed effects as control variables in my analysis of accounting returns. I specify:

$$Y_{it} = B * High_{it} + B * Size_{it-1} + B * Turn_{it-1} + B * Leverage_{it-1} + B * MTB_{it-1} + \delta_s + \mu_t + \varepsilon_{it} \quad \text{Eq. 4}$$

where there are *i* firms in *t* years and δ_s and μ_t represent fixed effect 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 this variable 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 4, weighted by the market value of the firm at the end of the previous year.

Model Uncertainty Analysis

The two most important sources of uncertainty in my replication relate to 1) ambiguity about EIS's 90 top firms and 2) doubt about the matching criteria used. To contain the problems created by uncertainty about the top 90 firms, I use 400 random draws of 90 from the eligible sample of 108. To bound the problems created by uncertainty about matching, I perform the matches using 4 different matching calipers[5]. This means that the aperture of my analysis is 1600 model estimates (400 treated cohorts X 4 matchings at different caliper settings).

RESULTS OF THE REPLICATION

As discussed by Kohler and Cortina (2023) replication often reveals hidden problems or uncertainties in the original empirical design, and that happened here. Specifically, I was

[5] I also conducted robustness tests where I allowed matches at the supersector level, and in an extreme robustness test, I created matches using a different approach (Coarsened Exact Matching) and employing different sector classifications (Refinitiv's Business Industry Codes). The results of these tests support the inferences reported here.

unable to replicate EIS's success in matching high and low-sustainability firms, and this led me to simulate the matching process under a variety of idealized conditions.

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. Using these criteria, I can match an average of only 9 *HS* firms across my 400 *HS* draws. What could explain the discrepancy? By inspecting the distribution of firms across sectors and p-scores, I identified cases where matching would be difficult or even impossible (see Appendix C, Figure C1). In some sectors there were more *HS* firms than *LS* ones, making matching all of them impossible. I also noted the expected distributional differences in p-scores for *HS* and *LS* firms. This problem of shared "support" is common in p-score matching.

The above analysis is contingent on my sample and could have been less binding for EIS. Could a more favorable sample have allowed 88 matches within a caliper of 0.01? To find out, I conducted simulations of matching using a variety of distributions of firms across sectors and propensity scores. To increase the probability of matching, I drew both target and control groups from the same distribution. As shown in Appendix C, I can find matches for only about 12 to 20 matches firms using EIS's criteria (same sector and propensity score difference < 0.01), and this corresponds more closely with my matching experience using actual data. The simulation also suggests that the EIS report of the matching processes is missing a step or condition. At a caliper of 0.01, I never succeeded in matching 88 pairs across 500 simulations, and I estimate that I would need to run $1 \text{ X } 10^{64}$ simulations to do so.

The apparent uncertainty of EIS's matching process reinforces the importance of evaluating portfolios using multiple caliper settings.

Descriptive Statistics

Table 1 provides descriptive statistics for my multiple portfolios. To allow comparison with EIS's single sample, I show the median value of the means and standard deviations for the descriptive variables (assets, ROA, leverage, turnover, and market to book) for the 400 runs at the extreme matching calipers (0.01 and 0.5). These values can be interpreted as providing information on the "average" portfolio given these calipers.

Table 1 reveals the expected tradeoff between the similarity of the matched pairs and the number of pairs identified. A tighter caliper results in fewer matches, yet more similar *HS* and *LS* pairs. Larger calipers allow more matches, but less similar *LS* and *HS* pairs.

The descriptive statistics for my samples conform well with the statistics EIS reports. The largest apparent difference between the two (other than the sample size) occurs for the variable turnover when comparing the *LS* sample from EIS and the *LS* sample formed using the 0.01 caliper. If a means test is done between this hypothetical average sample and the EIS sample, this difference is significant, but this is not the case for any other comparison.

Insert Table 1

Analysis of Stock Returns

Using the same Fama-French model employed by EIS, I compared the difference in stock returns for the *HS* and *LS* portfolios for each of 1600 paired portfolios (400 draws of the *HS* group X 4 matches with different caliper settings). Table 2 shows results from a typical run and Figure 3 provides graphical information capturing the results from all the runs.

Table 2 provides the median values for the coefficients and standard errors for Fama-French analysis of the *LS* and *HS* portfolios (using the extreme caliper settings: 0.01 and 0.5). It gives a sense of what estimates from a "typical" run look like. In this implementation of Fama-French

analysis, the intercept is of the most interest because it provides a measure of the abnormal performance of the portfolio. Positive and larger intercepts imply higher-than-expected performance, and the difference between the *HS* and *LS* groups provides an estimate of the relative performance of *HS* companies over the matched *LS* counterparts.

For all of the "typical" runs, the intercept coefficient is positive and statistically significant – matching EIS's report. This is likely caused by the selection process for forming both the *LS* and *HS* groups. Both had to exist from 1993-2010 and be chosen for rating by Refinitiv – thereby adding a survivor bias and a selection process distinguishing them from the average firm. In addition, the removal of financial firms may also have dampened the negative effects of the 2008 financial crisis.

For equal-weighted portfolios, the "typical" run using the 0.01 caliper results in estimates of a larger difference in performance than that reported by EIS (0.0041 vs 0.0018) but the high uncertainty in these estimates means that the difference is not statistically significant. For the looser caliper setting, the "typical" run estimates a higher return for the *LS* group compared to the *HS* one, but the difference is again not statistically significant. For value-weighted portfolios, estimates of the intercept are larger for the *HS* group for both tight and loose calipers, but the differences are again not statistically significant.

Insert Table 2

Figure 3 provides graphs of the difference between the *HS* and *LS* pairs for all the models run. Estimates are sorted from the largest positive difference (those most supporting the hypothesis of superior performance by *HS* firms) to the largest negative difference. Grey spikes show a 95% confidence interval, with darker spikes indicating portfolios using less stringent

matching criteria and thus including more firms (thus they also have smaller confidence intervals). The bold dashed line shows the estimate reported in EIS[6].

For the analysis of unweighted portfolios (Figure 2a), 71.5% of the estimates of return are higher for the *HS* group than for the *LS* group. Some of the differences are larger (particularly for the smaller but tighter portfolios) than those reported by EIS, but all are so uncertain that the confidence intervals include zero. Thus, from the perspective of classic frequentist statistical inference, had a scholar selected any one of my replication portfolios to analyze, he/she would not have been justified to reject the NULL hypothesis that *HS and LS* firms have equivalent stock returns. Note that my calculation for the confidence interval for EIS's estimate overlaps zero. Later I will explore this in more detail.

Figure 2b, shows the results for value-weighted portfolios. The majority (75%) result in positive coefficient estimates, but none of these estimates can be confidently differentiated from zero. Moreover, portfolios with a smaller number of matched firms (lighter interval lines) dominate the extremes of the estimates. This is unsurprising given the higher variance caused by smaller portfolio size. EIS's reported estimate is near the extreme. Only 6 of my portfolios result in return differences larger than the one reported by EIS, and these portfolios have only 7 or 8 pairs. As a result, the estimates are uncertain, and the confidence intervals include zero.

In summary, across 1600 cases, using both unweighted and weighted portfolios, I am unable to find a single case where I can reject the NULL hypothesis that *HS and LS* firms have equivalent stock returns.

Insert Figure 3

---

[6] Because EIS does not report standard errors, calculation of the confidence interval requires conversion from p-value ranges. I later discuss the details of this calculation.

Analysis of Accounting Performance

Figures 3a-d show my analysis of accounting performance. Graphed is the coefficient for "High" from equation 4. This coefficient provides an estimate of the relative performance of the *HS* group while controlling for other observable corporate attributes. As before, the black line is the estimated difference in returns, and the grey spikes show the 95% confidence interval, with darker spikes indicating portfolios using less stringent matching criteria and thus including more firms. EIS provides no estimate of annual accounting performance (only cumulative), so I cannot include EIS's estimates in the graphs.

Figure 4a and 4b show results for unweighted ROA and ROE. Most estimates are negative (lower performance for HS firms), but very few of the models provide estimates where the confidence interval does not include zero. Not surprisingly, small portfolios provide the most extreme estimates and the largest confidence intervals. A few positive estimates have confidence intervals that do not include zero (ROA:3, ROE:7).

Analysis of weighted portfolios reveals a similar pattern. Negative estimates outnumber positive ones, and a few would pass traditional significance tests. Note that the number of runs graphed has fallen from 1600 to 1480. This is because a weighted least squares analysis could not estimate coefficients for some of the smallest portfolio groups. All positive estimates have confidence intervals that include zero,

In summary, across 1600 unweighted portfolios and 1480 weighted ones, I find a total of ten portfolios where a scholar could reject the NULL hypothesis that *HS* and *LS* firms do not have superior accounting returns. For more portfolios (109), a scholar could reject the NULL hypothesis that *HS* and *LS* firms do not have <u>inferior</u> accounting returns. In the next section, I will compare these results to estimates from randomly assigned portfolios.

Insert Figure 4

Comparison with Estimates where NULL is known to be true.

In the usual frequentist analysis, single coefficients are tested by comparison with a hypothetical where the NULL is true. Can something similar be done for multiple estimates? Simonsohn, Simmons, & Nelson (2020) recommend comparing a multiverse of analyses to an identical multiverse where the NULL hypothesis is known to be true. This is done by randomly shuffling the treatment variable, reconducting the analysis, and then comparing the shuffled results (where B is known to be 0) to the unshuffled results (where B is unknown). I implemented this process by using a random number generator to determine the sustainability identifier (*HS* & *LS*) for each pair of firms.

Table 3 shows a summary of the results, and Appendix C shows the results in graphical form. The shuffled data deliver estimates similar to those from the replication data. The mean and centile of estimates from shuffled treatments are similar and none of the differences are statistically significant. For stock returns, the shuffled data results in fewer positive estimates, but neither original nor shuffled data deliver any estimates that would be deemed "significant" using the standard frequentist rule of $p< 0.05$.

For the accounting data, the shuffled data results in fewer negative estimates than for the original data. In all cases, the shuffled data delivers the same or more estimates that are both positive and "statistically significant". Thus, it appears that the original data are less supportive than the shuffled (no effect) data of the inference that sustainability is positively connected to accounting return.

A hint of a signal that *HS* firms outperform *LS* ones comes from the Fama-French analysis, where the unshuffled data deliver more positive (but not significant) results than the

shuffled data. This possible signal is not strong enough to reject the NULL, but it suggests an avenue for further inquiry.

Insert Table 3

Extension: Analysis Without Pairing

The matching method that EIS employs has both advantages and disadvantages (Guo, Fraser, & Chen, 2020; King & Nielson, 2019). Most scholars believe that matching should be just one of several methods of estimation. Indeed, recent work suggests that if matches are loose or limit the sample, propensity score matching increases bias (King & Nielson, 2019). Thus, I extend EIS's analysis by dropping the matching requirement and substituting instead a series of panel regressions. In doing so, I also reduce concerns about inconsistencies between my matching process and EIS's because now no matching is needed.

Because I have panel data at the annual level, the unique observation is for an *I* firm in *t* year.

Eq.2: $$Y_{it} = B * HS_{it} + B * LS_{it} + B * Survive_{it} + B * Rated_{it} + B * Size_{it-1} + B * Turn_{it-1} + B * Leverage_{it-1} + B * MTB_{it-1} + \mu_t + \delta_s + \varepsilon_{it}$$

Because I do not limit the sample by survival or matching, *HS* denotes all firms that are in Refinitiv's top quartile, and *LS* denotes all firms in Refinitiv's bottom two quartiles. *Survive* is a dummy indicating that the firm was in business from 1993-2010, and *Rated* indicates the firm was rated by Refinitiv in 2003-2005. The other variables are the same as those in Equation 1. When predicting stock return, I include a lagged measure of ROA.

Table 4 reports estimates from models predicting stock return, ROA, and ROE. Across all three, there appears to be no evidence that *HS* companies outperform *LS* ones. For model 2, the results contradict EIS's hypothesis, because *LS* firms appear to significantly outperform *HS* ones. In total, I believe that the results of Table 4 can best be interpreted as further reinforcing a lack of

robust evidence for the claim that *High Sustainability* firms outperform their counterparts concerning both stock returns and accounting performance.

Estimates shown in Table 4 also provide insight into why both LS and HS groups outperform. For all relevant models, the coefficient for *Survive* is strongly positive, suggesting that survival selects for stronger firms. When predicting stock return, *Rated* is strongly positive – perhaps suggesting that Refinitiv selected higher-performing stocks to rate first.

Insert Table 4

RECONCILING RESULTS FROM REPLICATION AND EIS REPORT

In my replication, I find no reliable evidence to reject the NULL hypothesis of no connection between high sustainability and financial performance. In contrast, EIS claims "High Sustainability companies significantly outperform their counterparts over the long-term, both in terms of stock market as well as accounting performance" (pg. 2836). Can the evidence and conclusions of the two reports be reconciled? I think they can. A close reading of EIS reveals that its reported evidence is insufficient to reject the NULL in five of its six tests of financial performance.

I believe the EIS report contains an error in the calculation (or an error in reporting) of the significance of the difference between the stock returns for the unweighted *HS* and *LS* portfolios.

EIS reports the intercept from the Fama-French analysis is larger for the *HS* portfolios whether value-weighted or equal-weighted (0.0096 to 0.0059 and 0.0057 to 0.0039). It asserts that the former difference is "significant at less than 5% level" and the latter is "significant at less than 10% level" (p. 2849). My analysis reveals the first statement is consistent with the reported evidence, but the second is not.

EIS does not report the standard errors for the coefficient estimates, but it is still possible to conduct a significance test of the intercepts from the Fama-French analysis (assuming no covariance between the samples). Recalculating the significance of the reported coefficient differences requires several steps: 1) calculating the range of potential coefficient differences, 2) converting p-value ranges into standard error ranges, 3) calculating the range of standard errors for the coefficient difference, and 4) recalculating the feasible range of significance.

For the value-weighted portfolios, I calculate a p-value range of about 0.0005 to 0.21, and EIS's report of $p<0.05$ is included in this region. For the equal-weighted portfolios, I calculate a feasible range from 0.17 and 0.39. This region does NOT contain the original claim of 0.1 or less, meaning that an error has occurred. When I pointed out this issue to the editors at *Management Science*, they contacted the authors and confirmed there was an error: "the sentence commenting Table 7 (p.2849 of the original manuscript) suffers from a typo and should read 'NOT statistically significant at less than 10% level'" (Editor, 2024). A typo had cause "not" to be left out.

Insert Table 5

For accounting performance, EIS does not report the results of any significance tests. As discussed earlier, it provides values of "cumulative" ROA and ROE, but it does not explain the calculation of this measure or provide any information on its certainty. It states that "The portfolio of High Sustainability firms outperforms the portfolio of control firms in 14 out of 18 years", but it does not disclose whether this concerns ROA or ROE, nor does it explain if this outperformance is cumulative or annual.

In summary, for four measures of accounting performance, EIS does not report any tests of statistical significance. Of six tests of financial performance, EIS provides a correctly

calculated significance test for just one. The joint probability of observing one significant result (or more) in six trials is 0.26 – five times the usual scientific standard of 0.05. Thus, the evidence provided in the EIS report (like that from my replication) does not refute the NULL hypothesis that *HS* and *LS* firms have the same financial performance.

CONCLUSION

Eccles, Ioannou, and Serafeim (2014) is an extremely influential publication – highly cited and employed as evidence from Wall Street to Capitol Hill. It has been cited more times than any contemporary or subsequent article in *Management Science.* Yet, my replication reveals difficulties with its methods and a lack of support for its findings.

For scholars of sustainable business, my analysis reconnects EIS to the preponderance of evidence on the connection between sustainable business and financial performance. A close reading of EIS itself reveals that it matches Mark Orlitzky's comment on the literature as a whole: its empirical evidence concerning financial performance does not allow strong conclusions. My replication reiterates this result and reveals that multiple models of the connection reveal a weak or noisy association. This is consistent with more recent evidence questioning the accuracy and value of ESG measures – such as those from Refinitiv used in this analysis.

For management scholars more generally, my research reveals the importance of replication. We must be able to learn from each other, and yet we cannot trust peer review to fully adjudicate the trustworthiness of what we read. EIS is a complicated publication, and its results seem convincing. Some of its problems are evident, but others are hidden. More replications (and more journals like JOMSR) are needed to safeguard the reliability of what we read.

My study also reveals that "the market for ideas" does not provide a credible signal of trustworthiness. We may hope that the wisdom of the crowd will divert attention from unreliable findings, but that has not happened with EIS. Its citation rate has grown almost every year, and it continues to be cited at a very high rate. Indeed, its very fame may insulate it from analysis, because readers may reasonably assume that among the thousands of people citing the article, at least one carefully checked its results.

My replication also reveals the value of a multiverse analysis. No single sample or empirical model can provide the basis for claims of knowledge or truth, because other samples and empirical methods may provide supporting or conflicting evidence. These may come from replications and extensions published in different reports, or they may come in the form of a multiverse of estimates within a single report. Interpreting any estimate is limited by both sample and model uncertainty. It provides just a pinhole view of the full picture. We must guard against excessive confidence based on a single study or finding.

Management scholars sometimes joke that their work has little impact, but this is surely not the case for research on sustainable business. Trillions of dollars in investments can be influenced, so too important policies, and maybe even the health of the planet. We have a responsibility to provide guidance justified by scientific evidence. That means we must do more original research and we must also employ replication studies to guide the use of the published record.

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(6), 1468-1497.

Aguinis H, Glavas A. 2012. What We Know and Don't Know About Corporate Social Responsibility: A Review and Research Agenda. *Journal of Management* 38(4): 932–968.

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 and Education*, 16(3): 469-483

Barnett ML, Salomon RM. 2012. Does it pay to be really good? addressing the shape of the relationship between social and financial performance. *Strategic Management Journal* 33(11): 1304–1320.

Blake, Lynn (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(2):313–362.

Casadesus-Masanell, R., & Ricart, J. E. (2012). *22 Competing through business models1* (Vol. 460). Cheltenham, UK: Edward Elgar Publishing.

Crasi, Tony (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. (2014). The impact of corporate sustainability on organizational processes and performance. *Management Science*, 60(11), 2835-2857.

Editor (2024). Report to [author of this manuscript] from a department editor at *Management Science*.

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

Guo, S., Fraser, M., & Chen, Q. (2020). Propensity score analysis: recent debate and discussion. *Journal of the Society for Social Work and Research*, *11*(3), 463-482.

Greening, D. W., & Turban, D. B. (2000). Corporate social performance as a competitive advantage in attracting a quality workforce. *Business & Society*, *39*(3), 254-280.

Hart, S. L. (1995). A natural-resource-based view of the firm. *Academy of Management Review*, *20*(4), 986-1014.

Henisz WJ, Dorobantu S, Nartey LJ. 2014. Spinning gold: The financial returns to stakeholder engagement. *Strategic Management Journal* 35(12): 1727–1748.

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

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

Hull CE, Rothenberg S. 2008. Firm performance: The interactions of corporate social performance with innovation and industry differentiation. *Strategic Management Journal*.

King, A. A. (2023). Writing a useful empirical journal article. *Journal of Management Scientific Reports*, *1*(3-4), 206-228.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political analysis*, *27*(4), 435-454.

Köhler, T., & Cortina, J. M. (2023). Constructive replication, reproducibility, and generalizability: Getting theory testing for JOMSR right. *Journal of Management Scientific Reports*, *1*(2), 75-93.

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*(2), 226-235.

Margolis JD, Walsh JP. 2003. Misery Loves Companies: Rethinking Social Initiatives by Business. *Administrative Science Quarterly*. Cornell University 48(2).

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

Orlitzky, M (2009) Corporate Social Performance and Financial Performance: A Research Synthesis, in *The Oxford Handbook of Corporate Social Responsibility* (Crane, Andrew, and others (eds), Oxford Academic, Oxford, UK.

Qin, S. (2011). Comparing the matching properties of Coarsened Exact Matching, Propensity Score Matching, and Genetic Matching in a Nationwide Data and a Simulation Experiment. *University of Georgia. ATHENS, GEORGIA*.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55

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*(1), 33-38.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208-1214

Singh, R. (2023). Defining Return on Assets (ROA) in empirical corporate finance research: a critical review. *Empirical Economic Letters* 23 (Special Issue 1): (January 2024)

Waddock SA, Graves SB. 1997. The corporate social performance-financial performance link. *Strategic Management Journal* 18(4): 303–319.

.

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

	Caliper < 1% difference in p-score				Caliper < 50% difference in p-score				EIS Report			
	LS		HS		LS		HS		LS		HS	
	Mean*	SD	Mean	SD	Mean	SD	Mean	SD	Mean	Std. dev.	Mean	Std. dev.
Assets	2.8E+09	4.0E+09	3.0E+09	4.0E+09	2.8E+09	3.9E+09	1.5E+10	3.4E+10	8.2E+09	2.8E+10	8.6E+09	2.2E+10
Ln(Assets)	21.02	1.26	20.74	2.07	20.83	1.64	22.40	1.63	22.83		22.87	
ROA	9.34%	5.94%	8.93%	9.32%	6.57%	8.37%	7.31%	6.99%	7.54%	8.02%	7.86%	7.54%
Leverage	0.69	0.53	0.51	0.11	0.55	0.29	0.60	0.17	0.57	0.19	0.56	0.18
Turnover	1.35	0.91	1.52	0.75	1.18	0.76	1.07	0.69	1.05	0.62	1.02	0.57
MTB	3.70	2.80	4.13	3.68	3.34	2.29	3.93	3.20	3.41	2.18	3.44	1.88
#Matches	9.09	9.09			68.5	1.81			90		90	

* all means and std deviations for replication data are medians from the means and std deviations for the 400 portfolios

Table 2: Stock market performance – Fama-French analysis

Table 2a: Equal-Weighted Portfolios

	Replication				EIS	
	Cailper 0.01		Cailper 0.5			
	Low	*High*	*Low*	*High*	Low	High
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Intercept	0.0090	0.0131	0.0094	0.0086	0.0039	0.0057
	(0.0026)	(0.0024)	(0.0013)	(0.0010)		
MKTRF	0.9077	1.0505	0.9594	0.9723	0.9977	0.9557
	(0.0608)	(0.0566)	(0.0313)	(0.0224)		
SMB	0.3062	0.0961	0.2661	-0.0261	0.1598	0.0366
	(0.0781)	(0.0727)	(0.0403)	(0.0288)		
HML	0.3806	0.0554	0.2552	0.1521	0.4053	0.2204
	(0.0796)	(0.0741)	(0.0410)	(0.0293)		
UMD	-0.1678	-0.1296	-0.1411	-0.1271	-0.1436	-0.1239
	(0.0491)	(0.0457)	(0.0253)	(0.0181)		
N	216	216	216	216	216	216
R-squared	62.3%	68.5%	86.4%	92.1%	88.9%	91.0%

Table 2b: Value-Weighted Portfolios

	Replication				EIS	
	Cailper 0.01		Cailper 0.5			
	Low	*High*	*Low*	*High*	Low	High
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Intercept	0.0063	0.0076	0.0061	0.0070	0.0059	0.0096
	(0.0023)	(0.0019)	(0.0017)	(0.0010)		
MKTRF	0.8194	0.7976	0.8650	0.8393	0.9839	0.9360
	(0.0540)	(0.0450)	(0.0399)	(0.0226)		
SMB	0.1871	0.0022	0.0687	-0.2390	-0.2076	-0.1776
	(0.0694)	(0.0578)	(0.0512)	(0.0290)		
HML	0.4904	0.3674	0.2716	0.1200	0.1982	-0.2727
	(0.0707)	(0.0589)	(0.0522)	(0.0296)		
UMD	-0.1322	-0.0440	-0.1063	-0.0514	-0.0156	-0.0266
	(0.0436)	(0.0363)	(0.0322)	(0.0182)		
N	216	216	216	216	216	216
R-squared	61.7%	64.9%	74.5%	88.8%	85.6%	86.6%

Table 3: Summary of replication results and shuffled (NULL) results.

	Replication Data						Shuffled Data					
	Stock Return*		ROA		ROE		Stock Return		ROA		ROE	
	Equal	Weight	Equal	Weight	Equal	Weight	Equal	Weight	Equal	Weight	Equal	Weight
B	0.0013	0.0008	-0.0032	-0.0023	-0.0254	-0.0263	0.0000	-0.0001	-0.0005	0.0001	-0.0011	-0.0004
B>0	1144	1204	589	552	246	179	772	774	805	671	799	654
B>0 & sig	0	0	3	0	7	0	0	3	65	84	86	82
B<0& sig	0	0	0	0	69	40	0	4	58	77	86	71
N	1600	1600	1600	1480	1600	1480	1600	1600	1600	1330	1600	1330
%>0	71.50%	75.25%	36.81%	37.30%	15.38%	12.09%	48.22%	48.34%	50.31%	50.45%	49.94%	49.17%
%>0 & sig	0.00%	0.00%	0.19%	0.00%	0.44%	0.00%	0.00%	0.19%	4.06%	6.32%	5.38%	6.17%
%<0& sig	0.00%	0.00%	0.00%	0.00%	4.31%	2.70%	0.00%	0.25%	3.63%	5.79%	5.38%	5.34%

*the difference in the performance of the high-low sustainability portfolios. When using weighted analysis, some results are lost due to a lack of sufficient degrees of freedom.

Table 4: Panel regression of financial performance

	(1) Stock Return	(2) ROA	(3) ROE
High Sustainability	0.0111	-0.0176**	-0.0203
	(0.0176)	(0.00694)	(0.0223)
Low sustainability	0.0104	0.0131**	0.00896
	(0.0145)	(0.00570)	(0.0183)
Survive	0.0519***	0.0525***	0.107***
	(0.00515)	(0.00202)	(0.00650)
Rated	0.105***	-0.0419***	-0.0639***
	(0.0116)	(0.00458)	(0.0147)
Lag ROA	0.0508***		
	(0.00984)		
Lag Assets	-0.0240***	0.0461***	0.0727***
	(0.00146)	(0.000538)	(0.00173)
Lag Turnover	0.0195***	0.0642***	0.0984***
	(0.00271)	(0.00103)	(0.00331)
Lag Leverage	0.0311***	-0.122***	-0.0687***
	(0.00684)	(0.00255)	(0.00820)
Lag MTB	-4.34e-05***	-4.42e-06***	1.63e-05***
	(1.30e-06)	(5.12e-07)	(1.65e-06)
Observations	90,168	90,076	89,966
R-squared	0.158	0.237	0.074

All models use fixed effects for industry (SIC3) and year. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1 two-tailed. For model 2, the difference between HS and LS is significant at p< 0.05.

Table 5: Reanalyzing statistical significance of differences reported in EIS.

	Weighted			Unweighted		
	Low	**High**	**Difference**	**Low**	**High**	**Difference**
Reported estimate	**0.00590**	**0.00960**		**0.00390**	**0.00570**	
Lower bound	0.00585	0.00955	0.00360	0.00385	0.00565	0.00170
Upper bound	0.00595	0.00965	0.00380	0.00395	0.00575	0.00190
Reported p-value	**<.0001**	**<.0001**	**<0.05**	**0.00400**	**<.0001**	**<0.1**
Smaller bound*	1.00000E-20	1.00000E-20		0.00300	1.00000E-20	
Upper bound	0.00010	0.00010		0.00400	0.00010	
	Recalculating Original Estimates					
Recalculated T-stat						
Larger bound	10.40308	10.40308		3.00260	10.40308	
Smaller bound	3.96626	3.96626		2.91015	3.96626	
Recalculated Std Error			Pooled			Pooled
Lower bound	0.00056	0.00092	0.00108	0.00128	0.00054	0.00139
Upper bound	0.00150	0.00243	0.00286	0.00136	0.00145	0.00199
Recalculated p-value						
Smaller bound			0.00046			0.17316
Upper bound			0.20856			0.39248

*Theoretically, when p is provided as < 0.0001, it could approach an arbitrarily small number. Practically, there is no p-value < 1.0 E^{-20} over 1600 replications.

.

Figure 1: Flow Chart of Replication Analysis

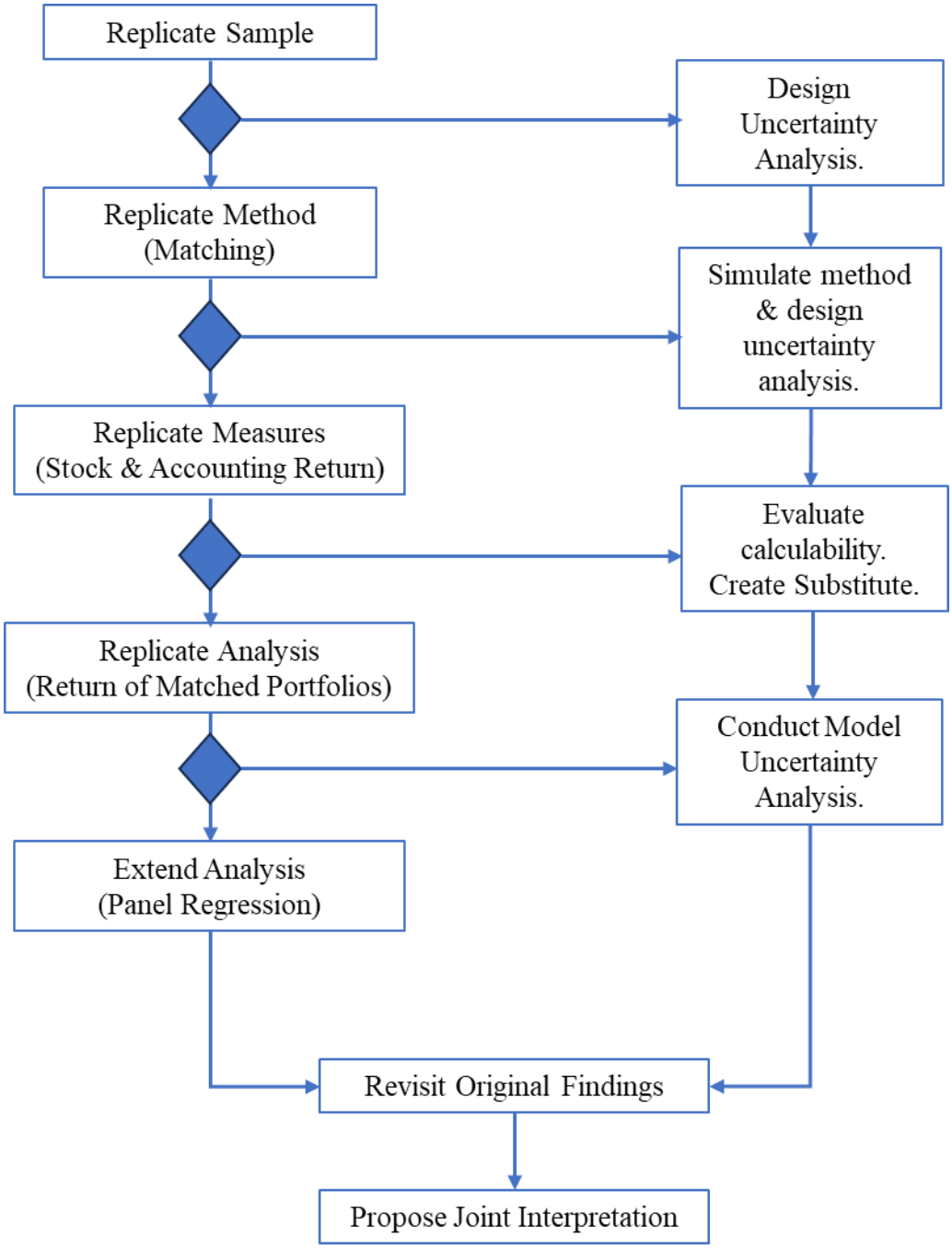
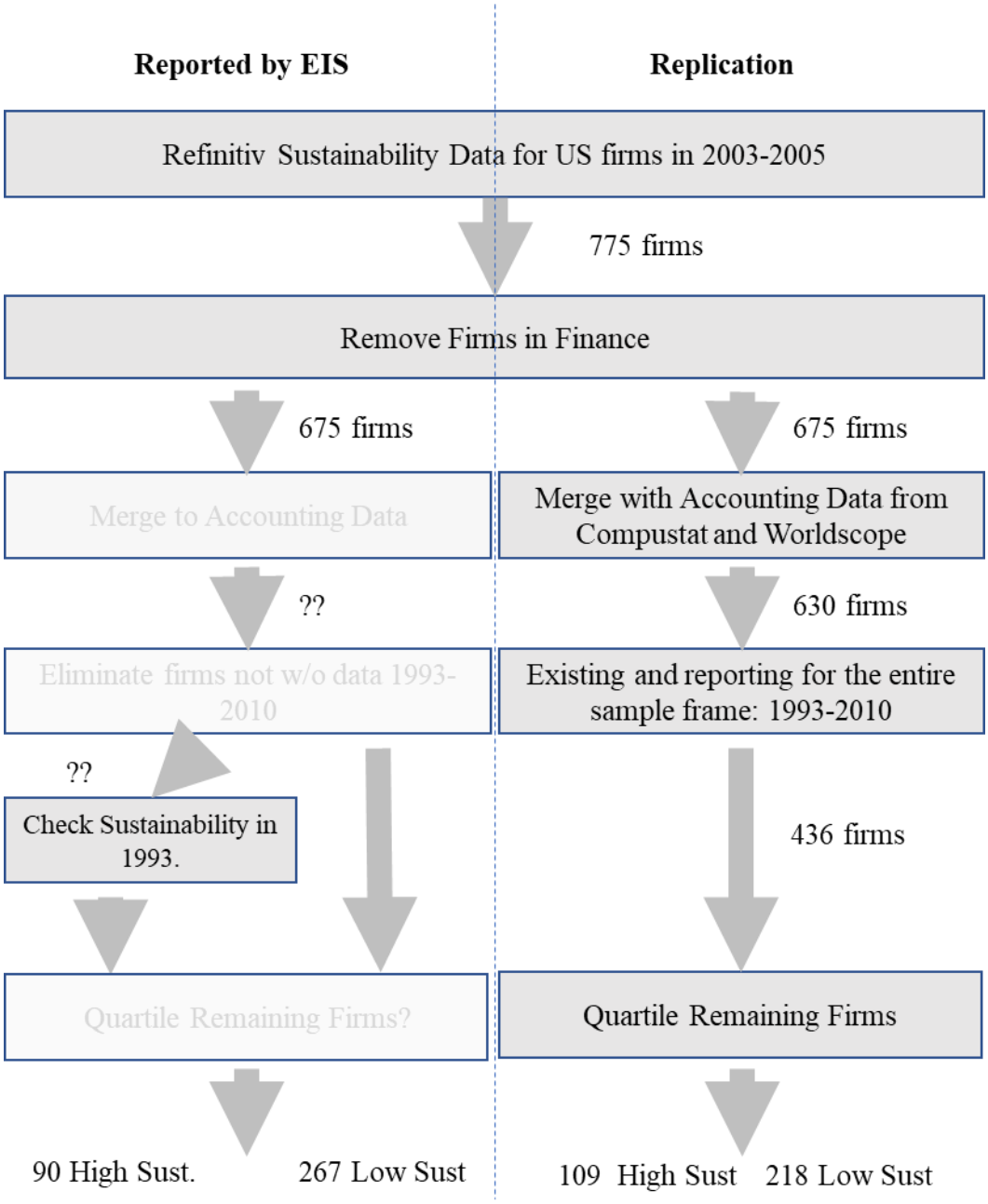


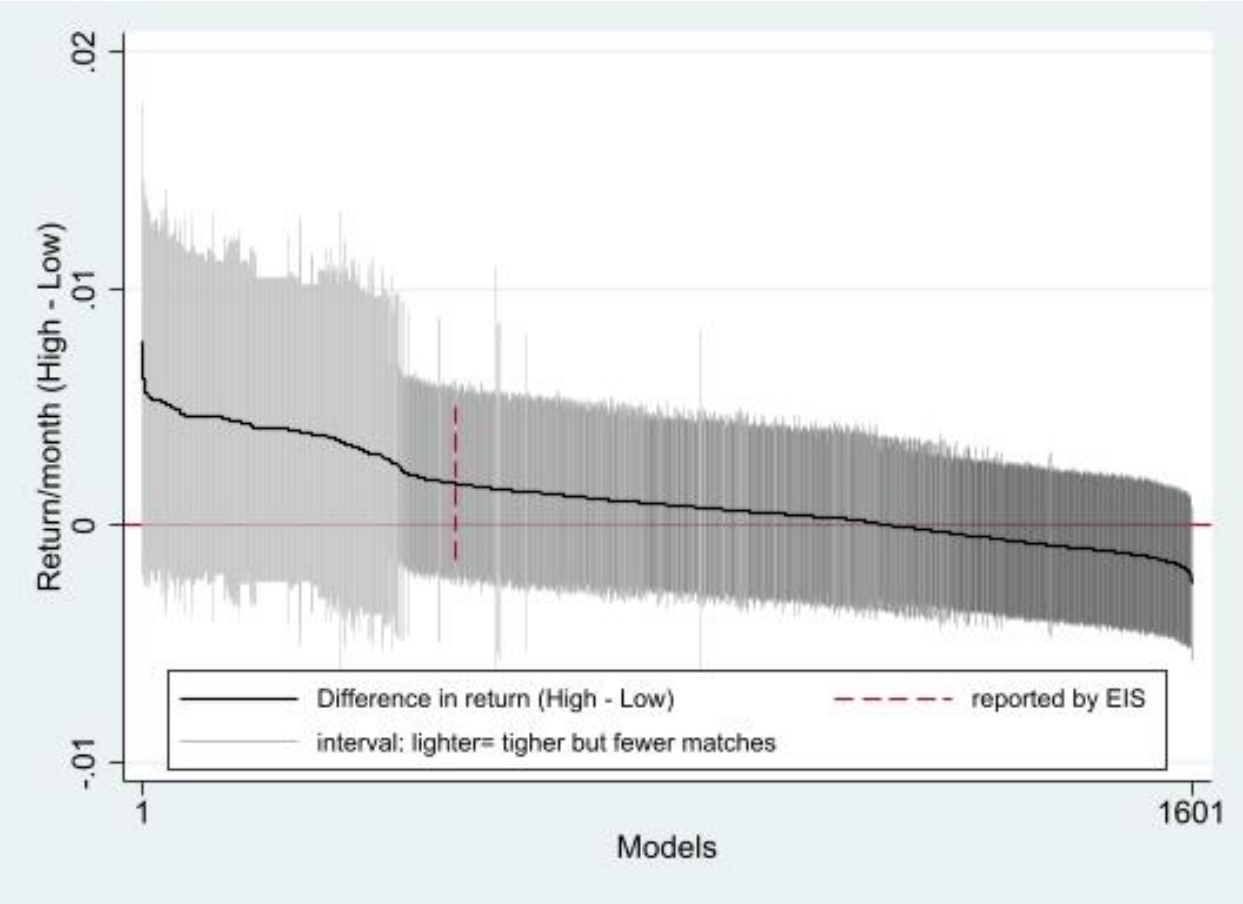
Figure 2: Sample Formation Process.



Note: steps unreported by EIS in ghosted letters.

Figure 3: Stock returns for high vs low sustainability portfolios

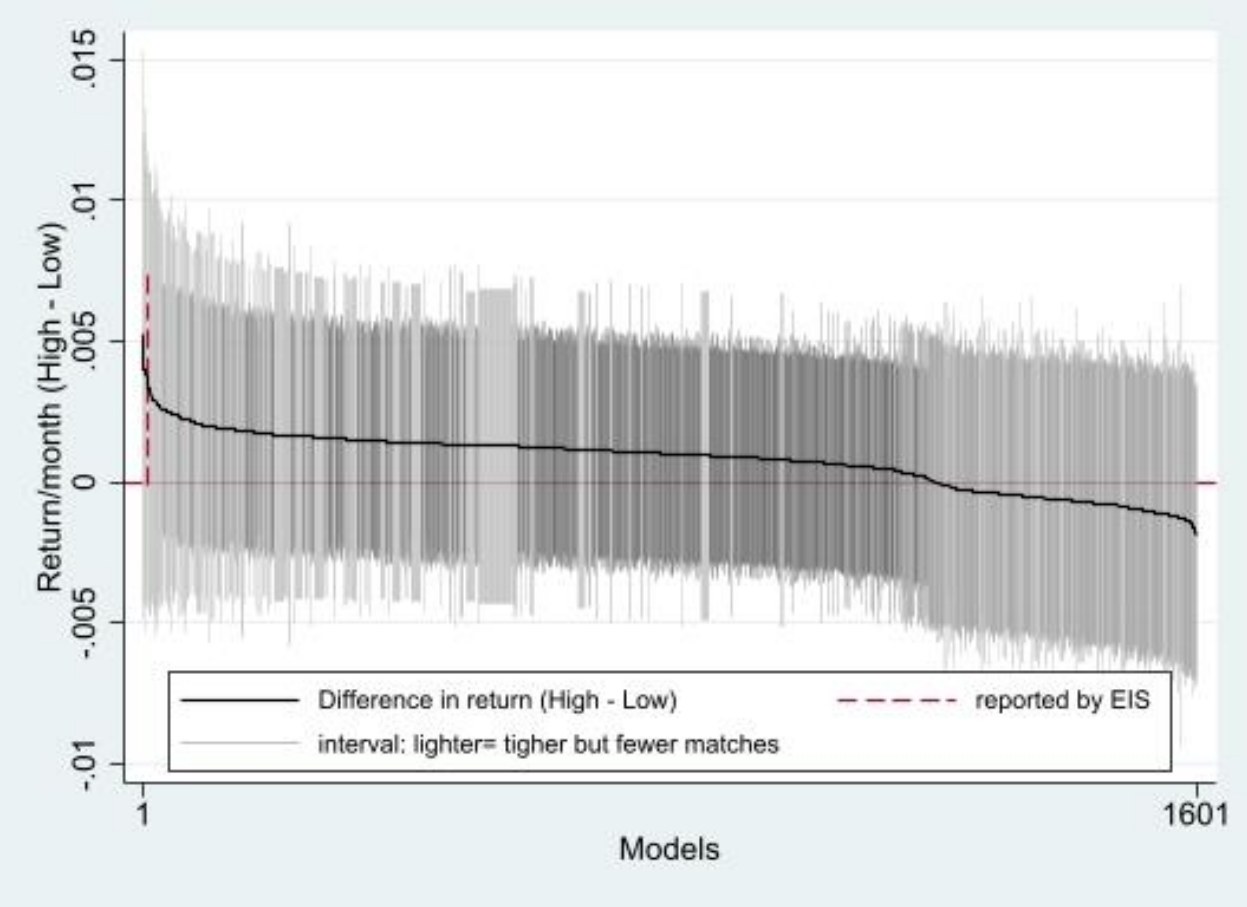
Figure 3a: Equal-weighted portfolios



Positive	Positive CI ni 0	Negative	Negative CI ni 0
71.5%	0%	28.5%	0%

Note: 1600 portfolios plus the one reported by EIS equals 1601.
The difference in returns is shown by the dark thick line. Confidence intervals are shown with grey spikes. The dashed line is the confidence interval for the estimate reported by EIS.

Figure 3b: Value-weighted portfolios

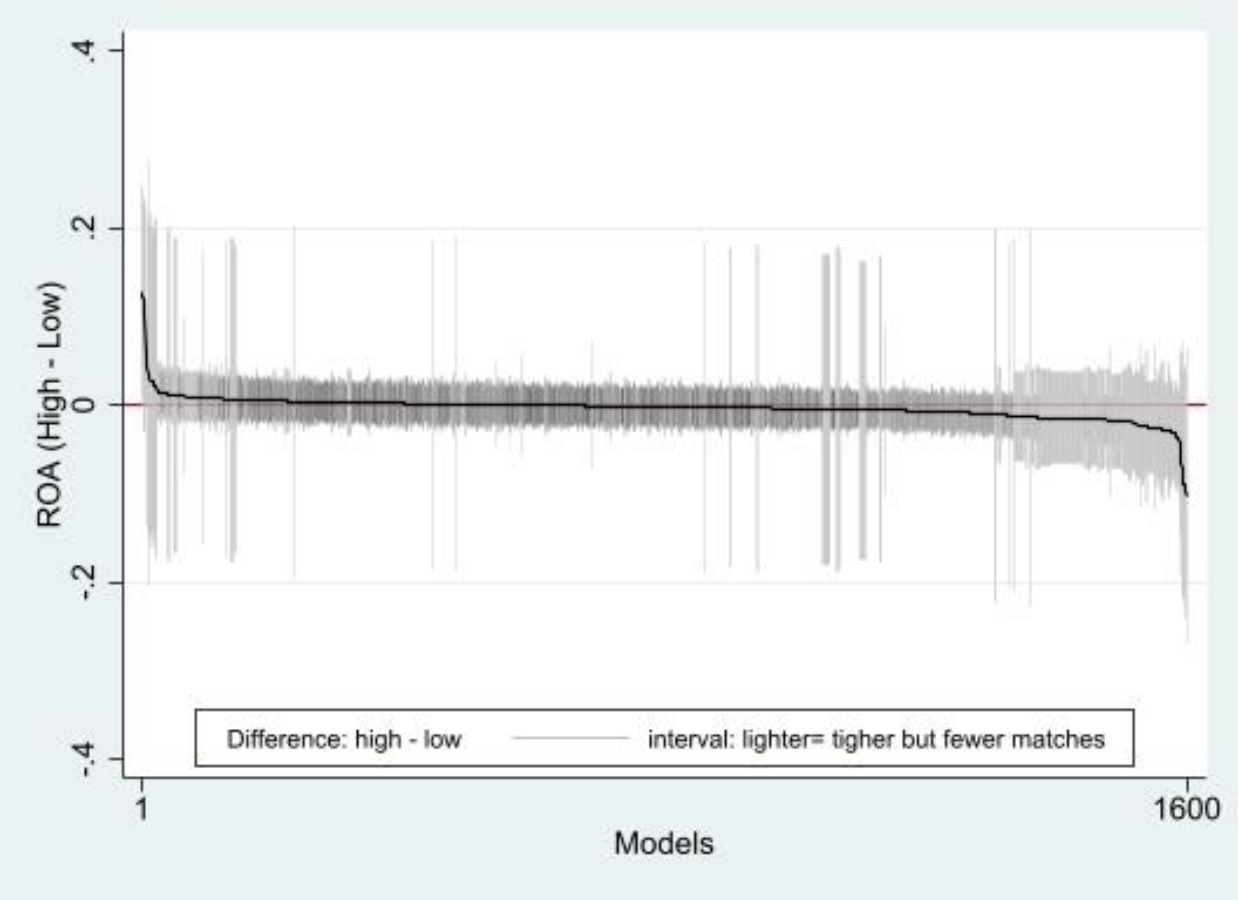


Positive	Positive CI ni 0	Negative	Negative CI ni 0
75.25%	0%	24.75%	0%

Note: 1600 portfolios plus the one reported by EIS equals 1601.
The difference in returns is shown by the dark thick line. Confidence intervals are shown with grey spikes. The dashed line is the difference in performance reported by EIS.

Figure 4: Accounting returns for high vs low sustainability portfolios

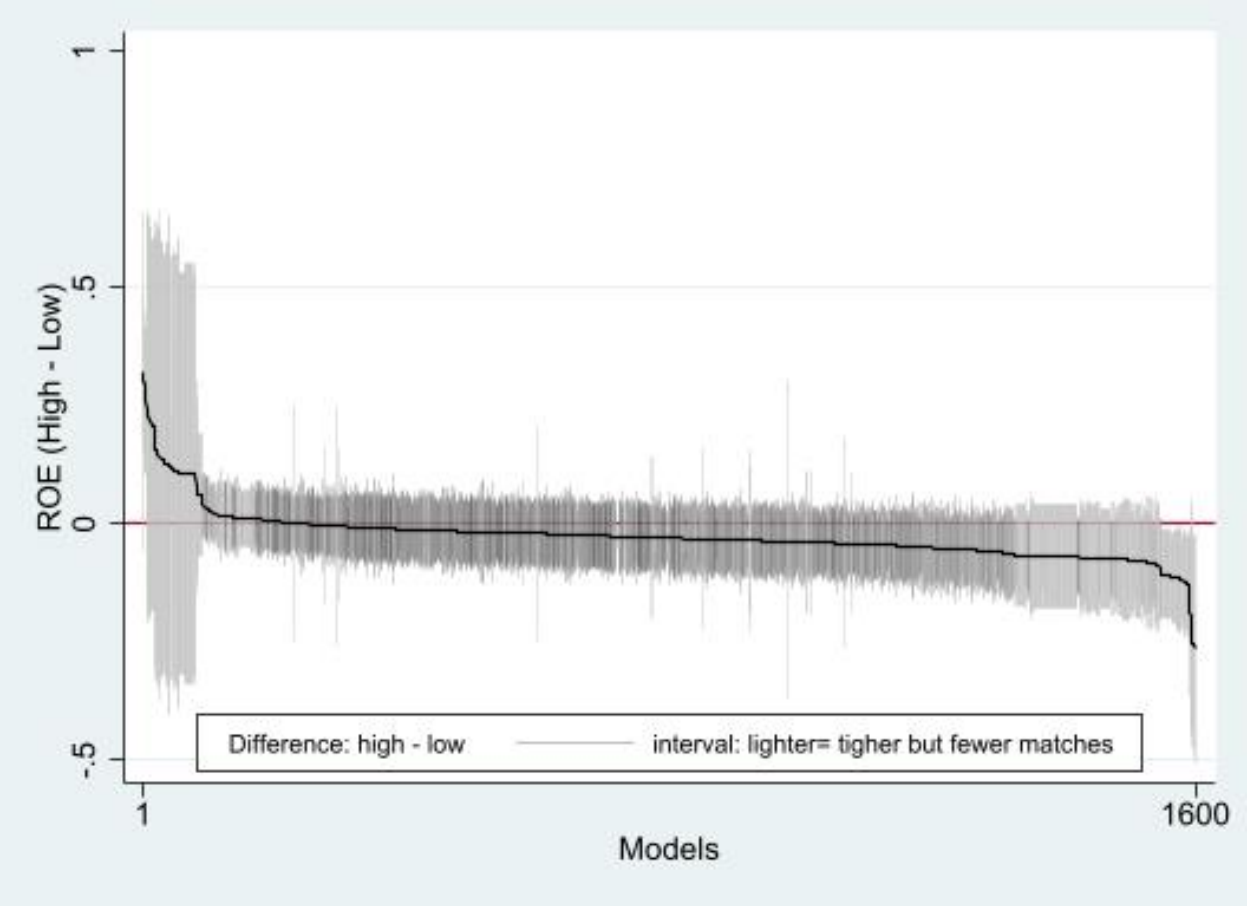
Figure 4a: Differences in ROA – Equal Weighted



Positive	Positive CI ni 0	Negative	Negative CI ni 0
37%	0.2%	65%	0%

Note: 1600 possible portfolio comparisons. The difference in returns is shown by the dark thick line. Confidence intervals are shown with grey spikes.

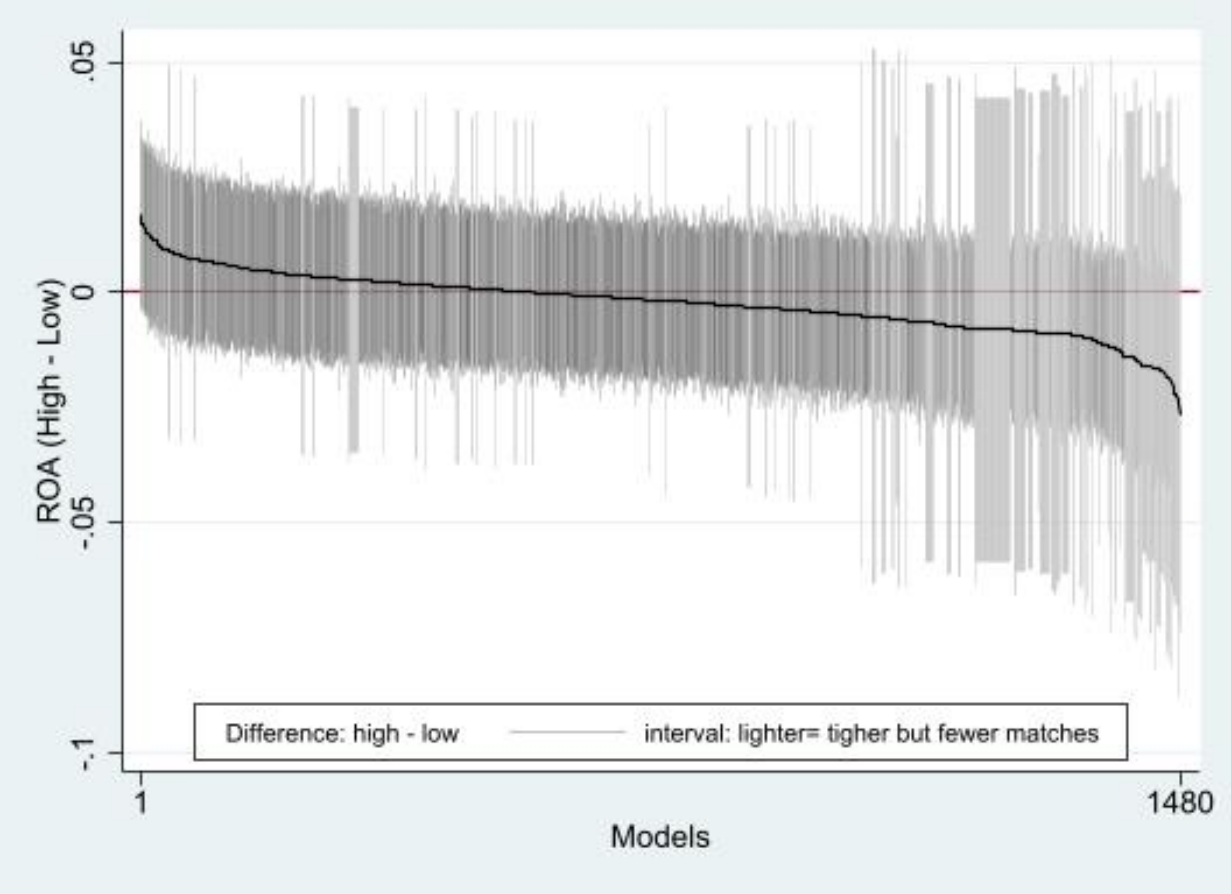
Figure 4b: Differences in ROE – Equal Weighted



Positive	Positive CI ni 0	Negative	Negative CI ni 0
16%	0.4%	84%	4%

Note: 1600 possible portfolio comparisons. The difference in returns is shown by the dark thick line. Confidence intervals are shown with grey spikes.

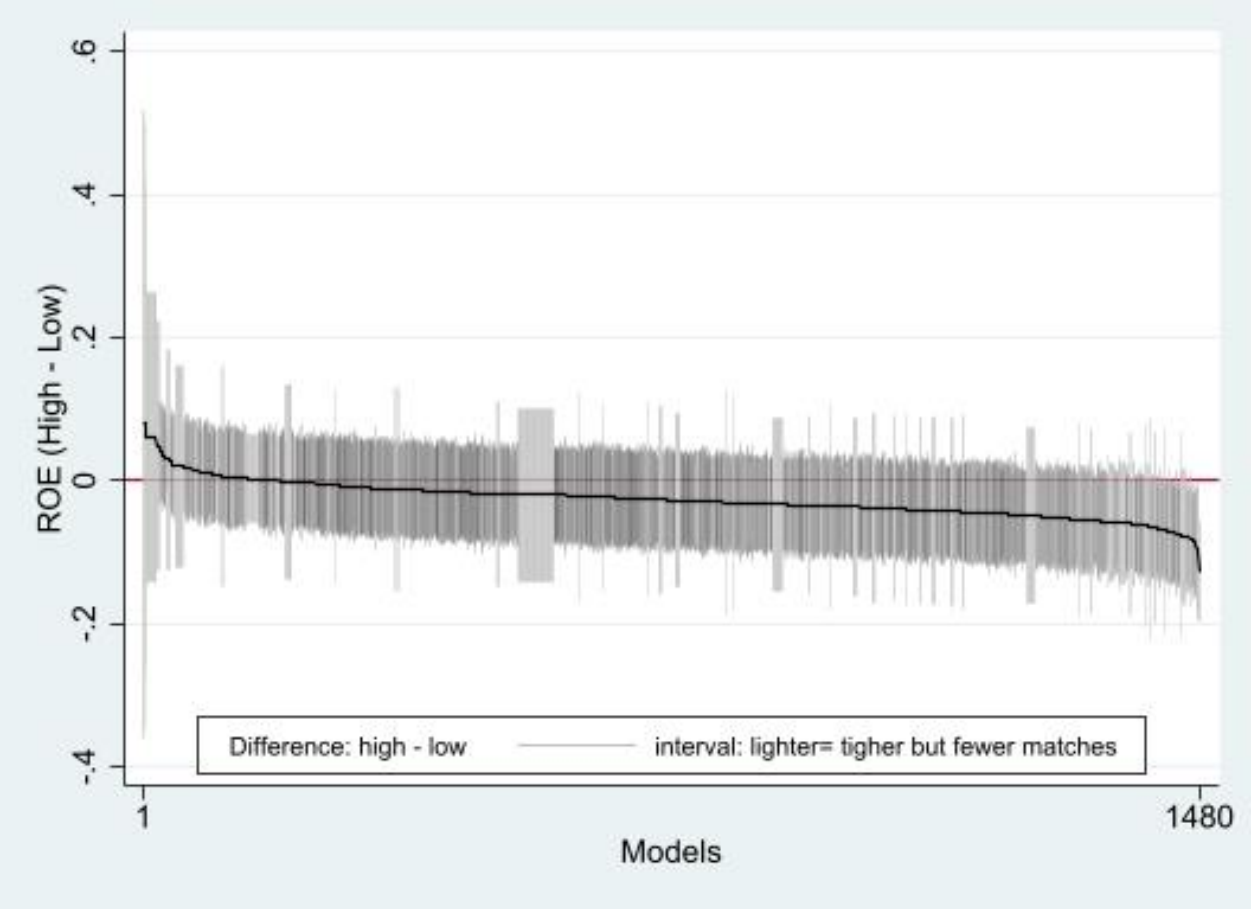
Figure 4c: Differences in ROA – Value Weighted



Positive	Positive CI ni 0	Negative	Negative CI ni 0
37%	0%	63%	0%

Note: we only have 1480 estimates, because for 120 models the estimate could not be calculated. This happens only when there are few matches, causing the sample size to be very small.

Figure 4d: Differences in ROE – Value Weighted



Positive	Positive CI ni 0	Negative	Negative CI ni 0
12%	0%	88%	2.7%

Note: we only have 1480 estimates, because for 120 models the estimate could not be calculated. This happens only when there are few matches, causing the sample size to be very small.

APPENDIX A – REFINITIV POLICIES USED TO SEPARATE HIGH AND LOW SUSTAINABILITY COMPANIES

Policy Name	Description
Bonus Plan for Employees/Employees	Does the company provide a bonus plan to most employees?
Community/Policy I	Does the company have a policy to strive to be a good corporate citizen or endorse the Global Sullivan Principles?
Community/Policy II	Does the company have a policy to respect business ethics or has the company signed the UN Global Compact or follow the OECD guidelines?
Diversity and Opportunity/Policy	Does the company have a diversity and equal opportunity policy?
Emission Reduction Policy Elements/Emissions	Does the company have a policy to reduce emissions?
Emission Reduction/CO2 Reduction	Does the company shows an initiative to reduce, reuse, recycle, substitute, phased out or compensate CO2 equivalents in the production process?
Emission Reduction/Transportation Impact Reduction	Does the company have initiatives to reduce the environmental impact of transportation of its products or its staff?
Employee welfare	Does the company have a work-life balance policy?
Employment Quality/Policy I	Does the company have a competitive employee benefits policy or ensuring good employee relations within its supply chain?
Employment Quality/Policy II	Does the company have a policy for maintaining long term employment growth and stability?
Environmental Supply Chain Management	Does the company use environmental criteria (ISO 14000, energy consumption, etc.) in the selection process of its suppliers or sourcing partners?
Generous Fringe Benefits	Does the company claim to provide its employees with a pension fund, health care or other insurances?
Health & Safety /Policy	Does the company have a policy to improve employee health & safety within the company and its supply chain?
Human Rights Contractor	Does the company show to use human rights criteria in the selection or monitoring process of its suppliers or sourcing partners?
Human Rights/Policy I	Does the company have a policy to guarantee the freedom of association universally applied independent of local laws?
Human Rights/Policy II	Does the company have a policy for the exclusion of child, forced or compulsory labor?
Internal Promotion	Does the company claim to favor promotion from within?
Management Training	Does the company claim to provide regular staff and business management training for its managers?
Positive Discrimination	Does the company promote positive discrimination?
Product Impact Minimization	Does the company design product features and applications/services that promote responsible, efficient, cost-effective and environmentally preferable use?
Product Innovation	Does the company have take-back procedures and recycling programs to reduce the potential risks of products entering the environment?
Product Responsibility/Policy I	Does the company have a policy to protect customer health & safety?
Product Responsibility/Policy II	Does the company have a products and services quality policy?
Resource Efficiency/Energy Efficiency Policy	Does the company have a policy to improve its energy efficiency?
Resource Efficiency/Water Efficiency Policy	Does the company have a policy to improve its water efficiency?
Training and Development/Policy	Does the company have a policy to support the skills training or career development of its employees?
Waste Reduction Total	Does the company have initiatives to recycle, reduce, reuse, substitute, treat or phase out total waste?

APPENDIX B: THE MEANING AND FEASIBILITY OF CUMULATIVE ROA OR ROE.

EIS seem to apply a compound formula to calculate cumulative ROA and ROE[7] 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:

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

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

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

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

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

Profits from one year do not directly influence assets in subsequent years and assets change for other reasons, so the product continues to be a complex combination of attributes that is difficult to interpret or connect to any construct.

Practically, 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.

For the above reasons, compounded ROA or ROE is not accepted 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.

[7] The unpublished penultimate draft of the paper includes graphs showing an evident compounding effect in the measure of cumulative ROA and ROE

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

To bound the feasibility of EIS's reported matching success, I decided to simulate matching under "best case" conditions (see Figure C1). 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. To aid the chance of matches, I used the same beta distribution for the pscores for both *HS* and *LS* firms. In a final model, I allow *HS* and *LS* to have two different beta distributions that approximate those observed in the real data.

Table C1 shows that for EIS's reported caliper, I estimate one could expect to match 88 about once in between 10^{64} samples or more (up to 10^{71}). Even at 10 times the reported caliper (0.1), 88 matches would occur in fewer than one in a billion samples. Only with a caliper at 0.5 does matching 88 become commonplace, but at this level, matching provides no benefit.

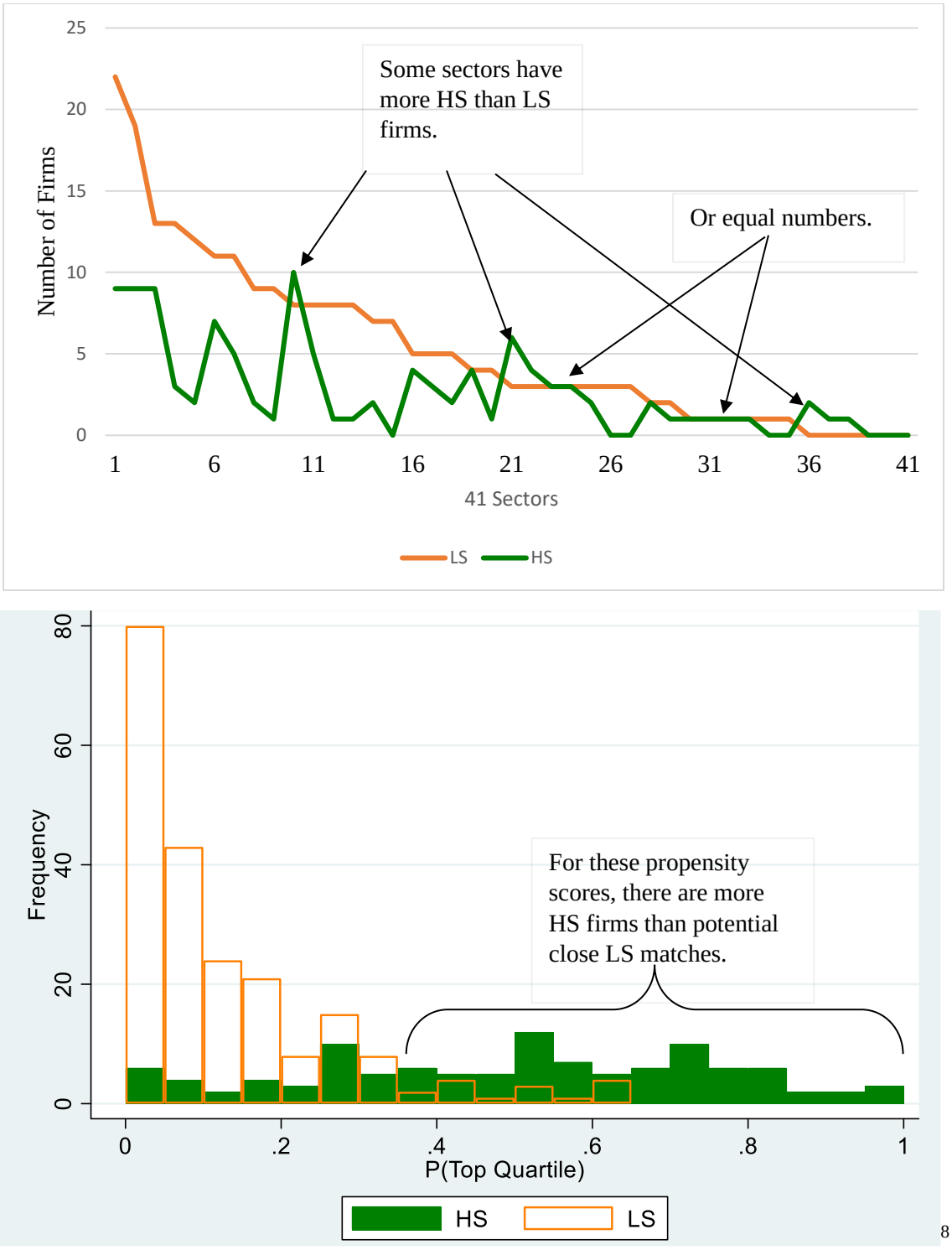
My simulation overstates the probability of matching because the actual conditions are less ideal. As shown in Figure C1, for my sample data the *HS* and *LS* firms are not distributed equally across the 41 sectors, and there are some sectors where *HS* firms outnumber *LS* ones – thereby preventing matches for all *HS* firms. The distributions of propensity scores are also not the same for the *HS* and *LS* firms – as would be expected given the construction.

While my simulation does not conform well with EIS's reported success, it 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 (10), but some loss of performance is to be expected when one moves away from ideal conditions.

Table C1: Expected matches for different conditions

Industry distribution	Propensity distribution	Caliper	m(matches)	sd(matches)	n for E(1|n,p)
Uniform	H:Beta(2,2) L: Beta(2,2)	0.01	12.7	3.2	2.5E+71
		0.1	63.1	4.3	9.7E+10
		0.25	81.6	3.0	126.7
		0.5	87.8	1.8	1.6
beta(2,2)	H:Beta(2,2) L: Beta(2,2)	0.01	14.8	3.5	3.0E+65
		0.1	66.0	4.2	2.5E+09
		0.25	82.2	2.9	74.6
		0.5	87.5	1.8	1.8
beta(1,3)	H:Beta(2,2) L: Beta(2,2)	0.01	20.5	3.9	1.5E+53
		0.1	71.6	3.9	2.9E+06
		0.25	83.7	2.7	22.5
		0.5	87.5	1.8	1.9

Figure C1: Practical barriers to matching



[8]

[8] As expected, HS firms have higher propensity scores and LS firms lower ones. This is common in such analysis and thus scholars often limit their matches to those with shared support.

Figure C2: Stata program for monte carlo estimation

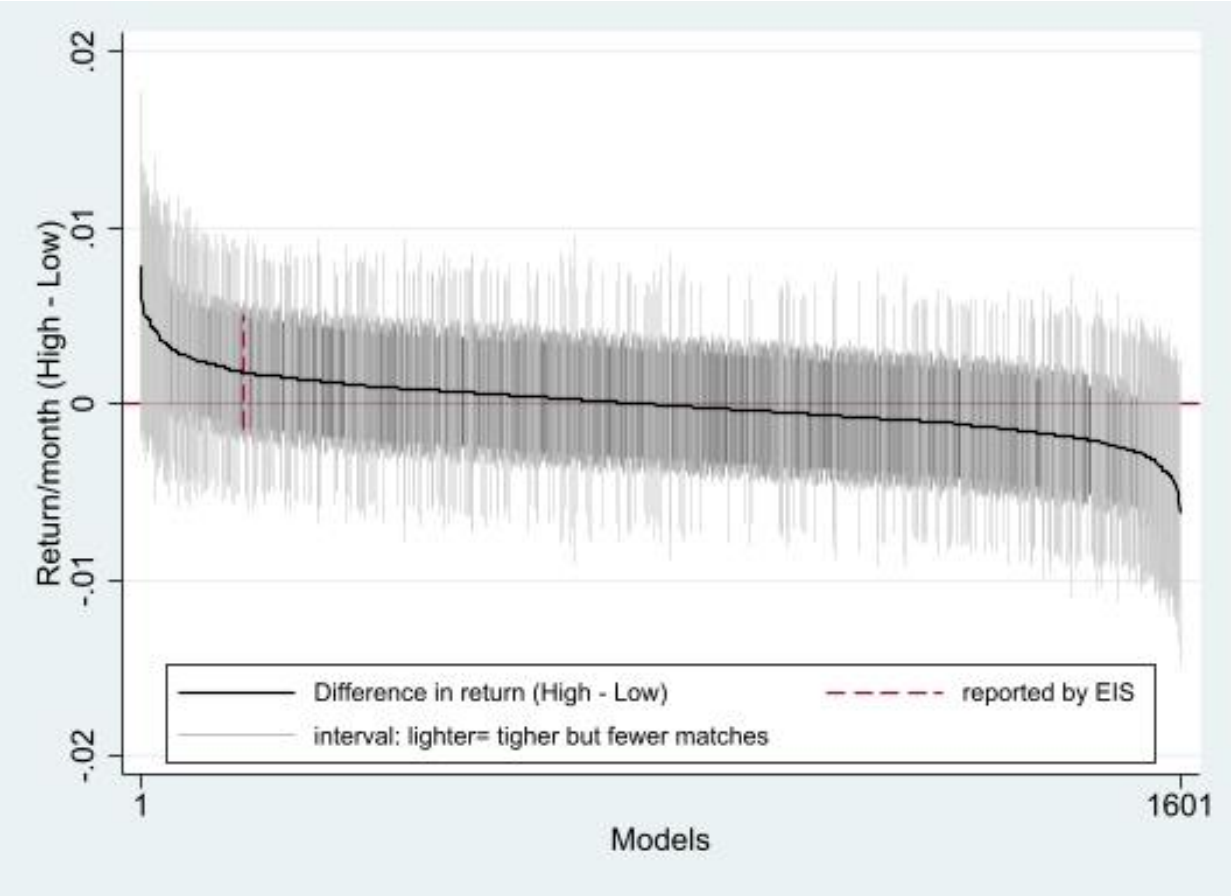
```
/*********************************************************************
**
**  This stata program simulates a matching process where 90 treated firms
**  and 269 possible control firms are distributed across 41 sectors.
**  It varies the caliper distributions of the allocation
**  to sectors and the 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)3{

                  **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(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
```

APPENDIX D: ANALYSIS OF SHUFFLED PORTFOLIOS

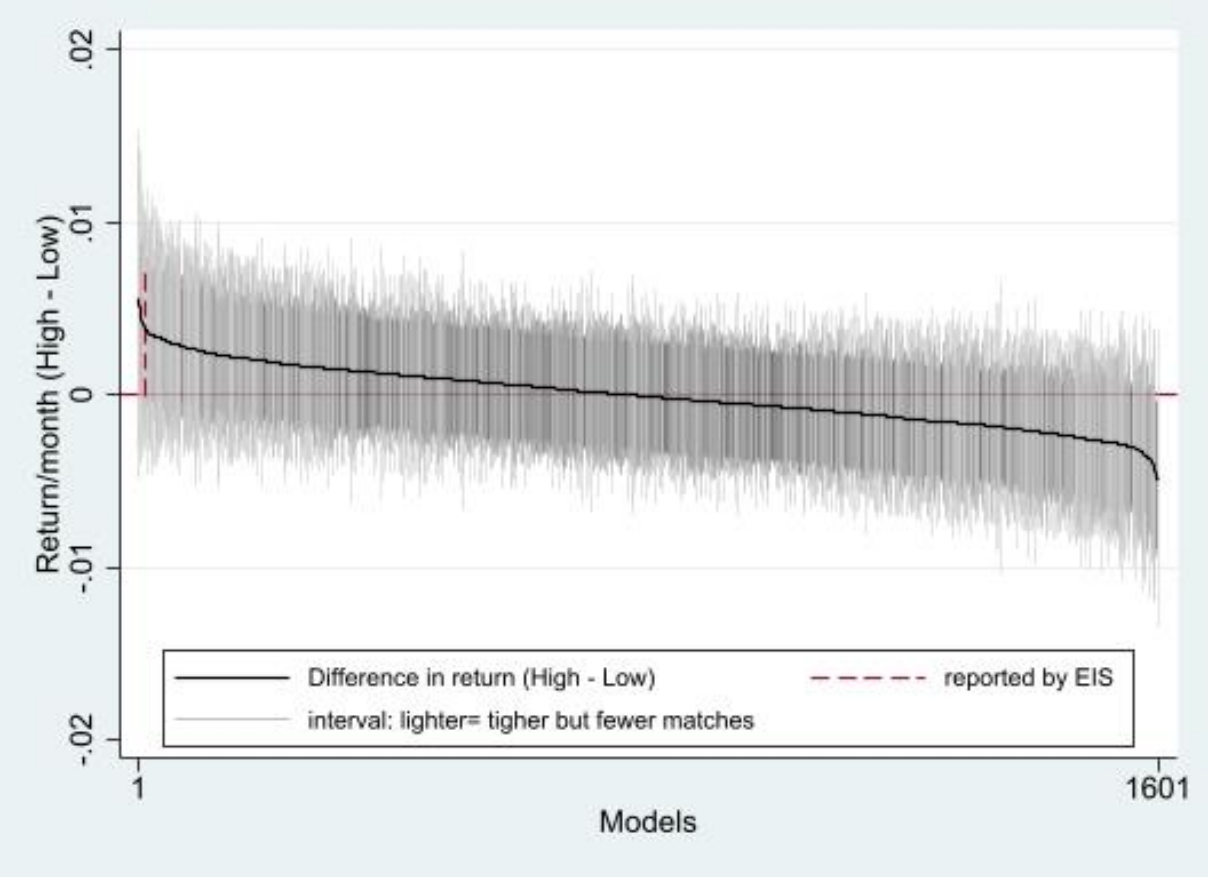
Figure D1: Equal Weighted Shuffled Portfolios (Fama French Analysis)



Positive	Positive CI ni 0	Negative	Negative CI ni 0
51%	0%	49%	0%

Note: The difference in returns is shown by the dark thick line. Confidence intervals are shown with grey spikes.

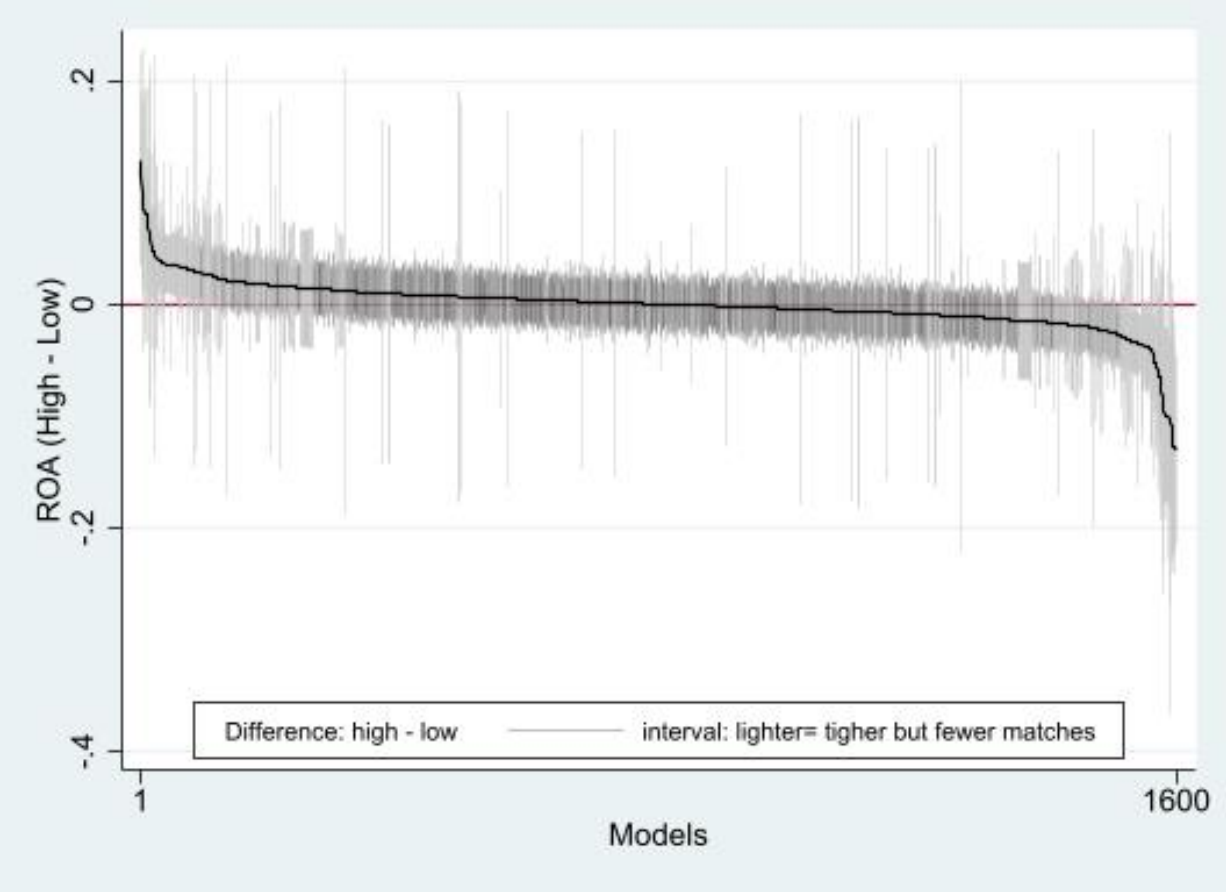
Figure D2: Value Weighted Shuffled Portfolios (Fama French Analysis)



Positive	Positive CI ni 0	Negative	Negative CI ni 0
50%	0%	50%	0%

Note: The difference in returns is shown by the dark thick line. Confidence intervals are shown with grey spikes.

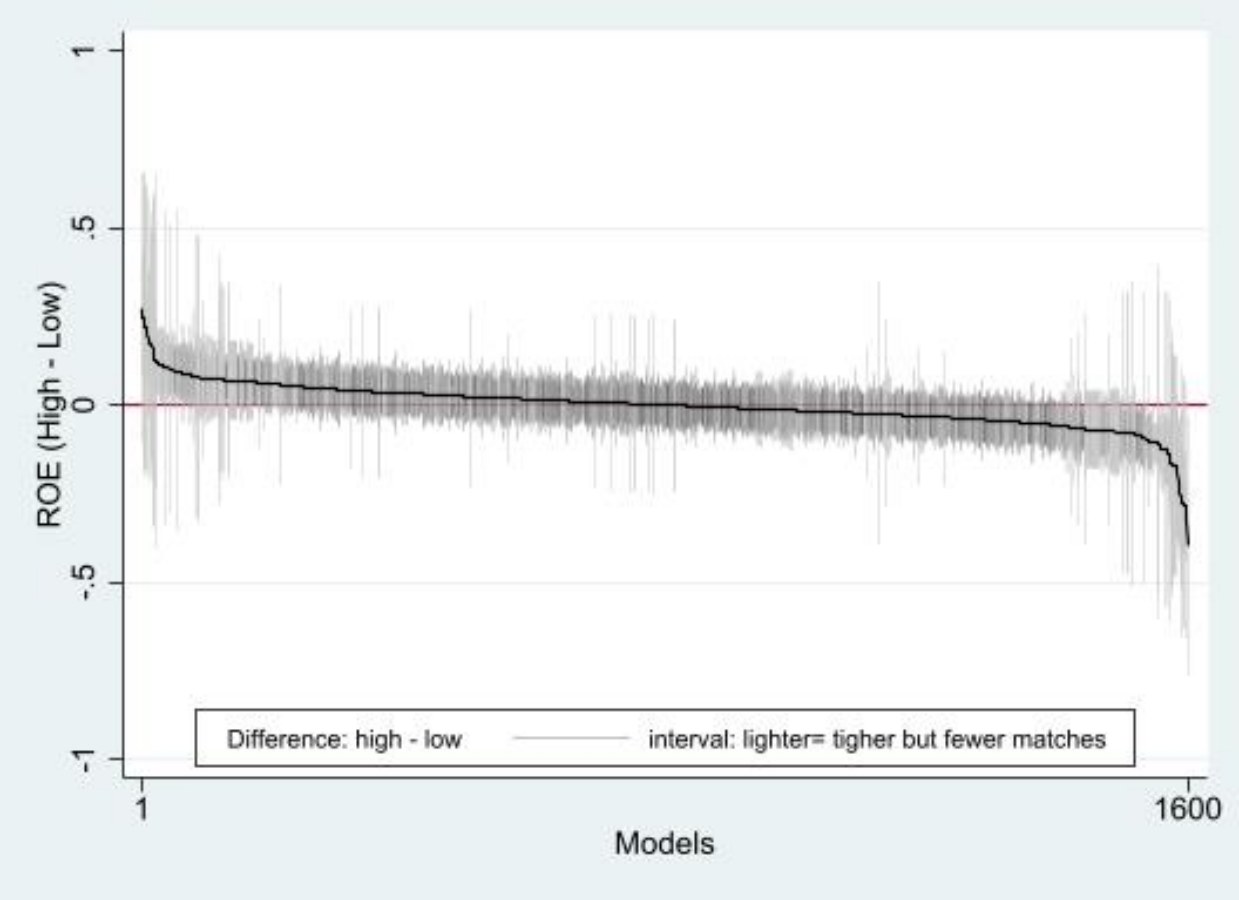
Figure D3: Differences in ROA (Shuffled Portfolios, Firms Equal Weighted)



Positive	Positive CI ni 0	Negative	Negative CI ni 0
49.5%	3.2%	50.5%	2.7%

Note: The difference in returns is shown by the dark thick line. Confidence intervals are shown with grey spikes.

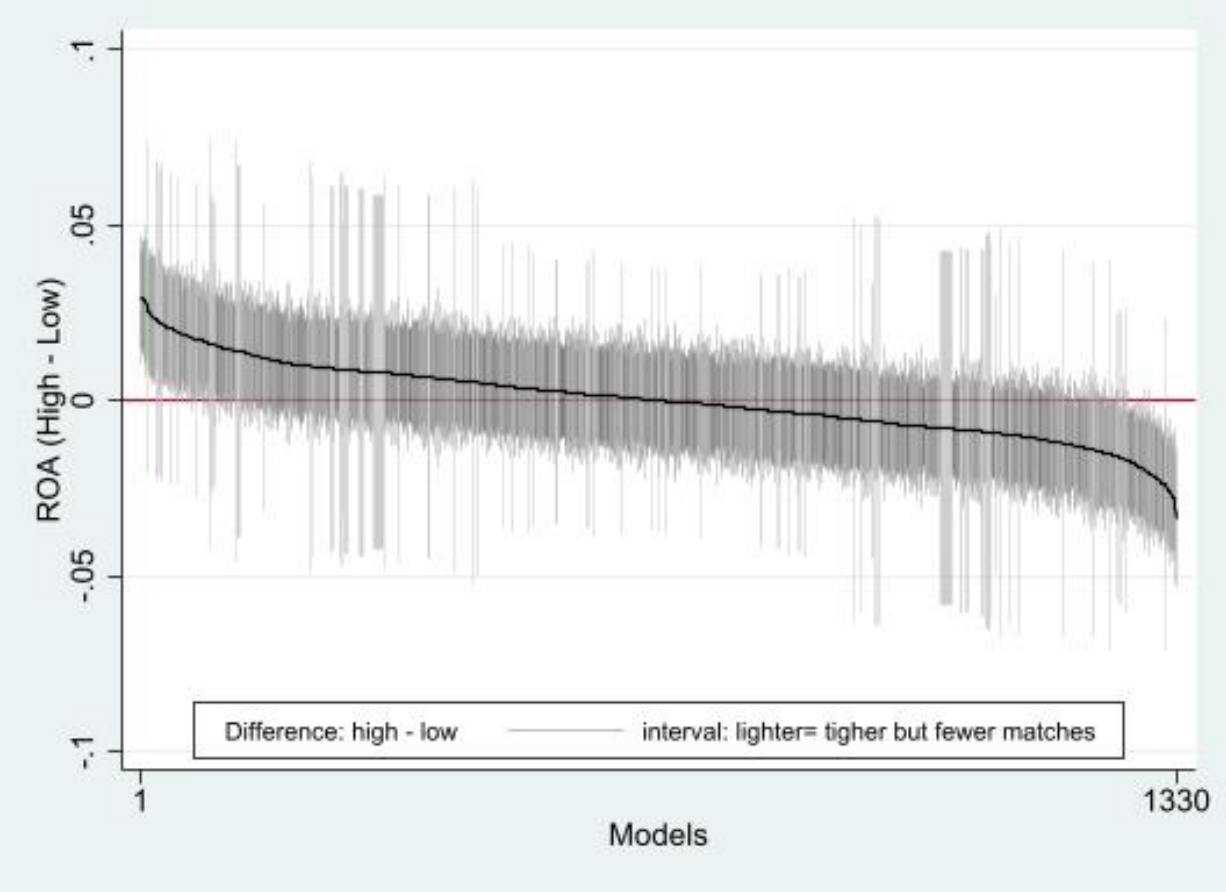
Figure D4: Differences in ROE – (Shuffled Portfolios, Firms Equal Weighted)



Positive	Positive CI ni 0	Negative	Negative CI ni 0
51%	3.3%	49%	2.8%

Note: The difference in returns is shown by the dark thick line. Confidence intervals are shown with grey spikes.

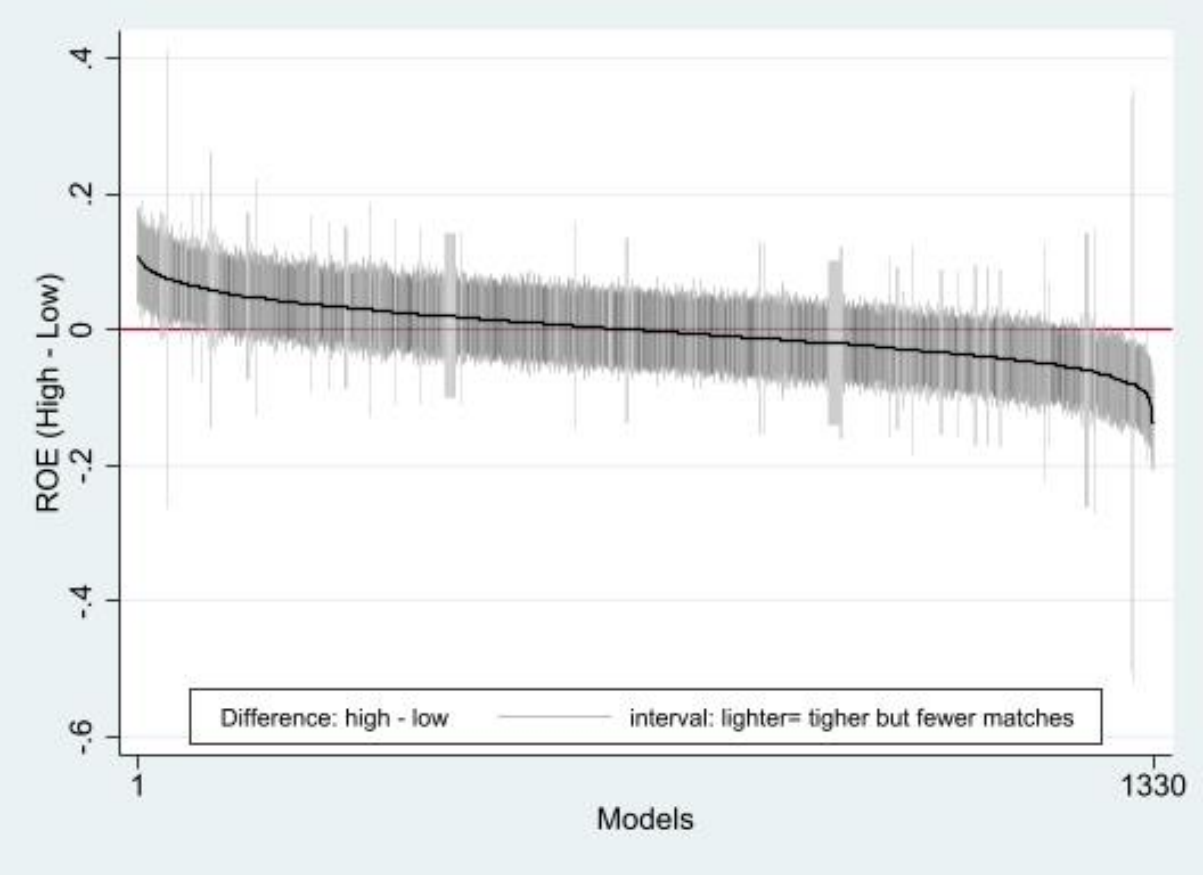
Figure D5: Differences in ROA – (Shuffled Portfolios, Firms Value Weighted)



Positive	Positive CI ni 0	Negative	Negative CI ni 0
50%	5.3%	50%	5.7%

Note: we only have 1330 estimates, because for 270 models the estimate could not be calculated. This happens only when there are few matches, causing the sample size to be small.

Figure D6: Differences in ROE – (Shuffled Portfolios, Firms Value Weighted)



Positive	Positive CI ni 0	Negative	Negative CI ni 0
50%	3.3%	50%	2.8%

Note: we only have 1330 estimates, because for 270 models the estimate could not be calculated. This happens only when there are few matches, causing the sample size to be very small.