The Road to False Positives: Evidence for Specification Search in Studies of Reg SHO

Bernard S. Black
Northwestern University, Pritzker School of Law and Kellogg School of Management

Hemang Desai
Southern Methodist University, Cox School of Business

Kate Litvak
Northwestern University, Pritzker School of Law

Woongsun Yoo
Central Michigan University, College of Business Administration

Jeff Jiewei Yu
University of Arizona, Dhaliwal-Reddy School of Accountancy

(draft January 2026)

Northwestern University, Pritzker School of Law
Law and Economics Research Paper No. 25-07

This paper is available on SSRN
http://ssrn.com/abstract=5024145

The Appendix is available on SSRN
http://ssrn.com/abstract=5024143

The pre-specified analysis plan can be downloaded at:
http://ssrn.com/abstract=3415529

Sincere thanks to Avanidhar Subrahmanyam for helpful discussions and comments and to Haoyan Shi, Daniel Xu, and Francesco Maria Rossi for excellent research assistance. We also thank Rishabh Aggarwal and seminar participants at Canadian Law and Economics Association; UMass – Lowell, Indian School of Business, Tel Aviv University Business School, Hebrew University Business School, and WHU Business School for their comments.

The Road to False Positives: Evidence for Specification Search in Studies of Reg SHO

Bernard Black, Hemang Desai, Kate Litvak, Woongsun Yoo, and Jeff Jiewei Yu

Abstract:

We provide evidence of p-hacking (not just file drawer or publication bias) by reexamining results from 10 studies that exploited an SEC randomized experiment (Reg SHO) and found a wide variety of indirect effects of the experiment, despite no direct effects and no evidence for a causal channel for the indirect effects. Using either our pre-specified sample and design (regression of various outcomes on treatment dummy with year and firm fixed effects) or best-match specifications (which closely follow the sample and specification choices reported in the respective papers) we obtain null results. However, reported results are mostly statistically significant, often highly so, due to both larger coefficients and smaller standard errors. The large gap between reported results and either our specifications or the best-match specifications is inconsistent with the reported results reflecting only lucky specification choice and are instead consistent with specification search or p-hacking. Binomial and Kolmogorov-Smirnov tests confirm that the reported results cannot plausibly reflect different reasonable draws from the same underlying distribution of potential results.

Keywords: natural experiments; p-hacking; specification choice; Regulation SHO; SEC's short-sale experiment

Disclosure regarding our prior related paper

This paper builds on Black, Desai, Litvak, Yoo, and Yu (2024), The SEC's Short-Sale Experiment: Evidence on Causal Channels and on the Importance of Specification Choice in Randomized and Natural Experiments, *Management Science* 70(8), 5131-5456 (*BDLYY 2024*). The focus of the prior paper was on examining the plausible causal channels (short interest, returns and managerial fear channel (media coverage)) and having not found evidence for the causal channels, we examined selected results from four papers to show lack of robustness.

Overall, 11 of the 70 regression results in Tables 1-3 were previously reported in BDLYY 2024. Figures 2-9 are new, as are 7 of the 11 graphs in Figure 1. All other results are new.

The Road to False Positives: Evidence for Specification Search in Studies of Reg SHO

I. Introduction

The issue of specification search (or p-hacking), in which authors search for significant results, is gaining attention in multiple fields, including economics, finance, and accounting (e.g., Harvey 2014, 2017, 2019; Brodeur et al., 2016, 2020, 2022; Ohlson, 2020; Hail, Lang, and Leuz, 2020; Gow, 2023; Menkveld et al., 2024; Gerber and Malhotra, 2008; Mitton, 2022). Several studies provide indirect evidence of p-hacking. For example, Brodeur et al. (2016) analyze the distribution of reported test statistics in 641 papers published in top economics journals from 2005-2011 and find a two-humped camel shape of the distribution of p-values, with under-representation of marginally insignificant statistics (p-values between 0.25 and 0.10) relative to significant statistics (p-values just below 0.05). They interpret the higher incidence of p-values just below 0.05 as evidence of p-hacking. Harvey, Liu, and Zhu (2016) analyze asset-pricing studies published in top finance journals from 1963-2012 and find that almost as many studies reporting *t*-statistics between 2.0-2.57 as between 2.57-3.14, and that very few published papers have *t*-statistics below 2.0.

However, the above studies document patterns of test statistics suggestive of p-hacking based solely on reported results from and hence cannot provide direct evidence of specification search. In this paper, we take a different approach. We closely examine 10 papers that rely on the same shock and compare the reported results to other specifications for the same outcomes and samples. Specifically, we exploit a randomized trial conducted by the SEC -- the Reg SHO experiment that ran from May 2005 to July 2007. In this experiment, the SEC suspended short-sale restrictions (uptick rule for NYSE/AMEX firms; bid test for Nasdaq firms) for one-third of the Russell 3000 (R3000) firms (pilot firms). The remaining R3000 firms served as controls. This experiment has been studied extensively. At last count, over 80 published papers and working papers report a vast array of *indirect* outcomes for pilot (treated) firms and attribute them to the short-sale experiment (Internet Appendix Table IA-1).

The SEC experiment is ideally suited to providing direct evidence of specification search for several reasons. First, early studies of the experiment found no direct effect of removing the short-sale restrictions on short interest, stock returns, volatility, or price efficiency for pilot firms (Diether et al., 2009; Alexander and Peterson, 2008; SEC, 2007). BDLYY (2024) extend these studies by utilizing a larger sample, a longer time period and studying an additional purported

causal channel, managerial fear. They confirm lack of evidence for a plausible causal channel for the reported indirect effects. Put differently, the original shock was too weak to credibly produce the broad array of outcomes reported in the indirect effect studies. This raises the likelihood that the reported results in these studies are false positives.

Second, the random assignment of firms to treatment (pilot) versus control provides covariate balance between the pilot and the control firms (see Internet Appendix Table IA-4). Thus, one can obtain valid inference by comparing univariate differences between pilot and control firms; or running simple DiD regressions of the outcome on a pilot dummy, with firm and year fixed effects (our specification, henceforth). Covariates are not needed. Thus, the results from these simple specifications can proxy for results without specification search and provide a benchmark against which the reported results in the indirect effect studies can be compared.

Third, we know, from the SEC releases, the exact sample of treatment (pilot) firms and control firms. This lets us assess sample selection choices in the indirect effect papers. Many of these studies impose data restrictions that result in a significant loss of observations, but seem unnecessary given the randomized experiment, with covariate balance and no differential attrition between treatment and control firms. For example, Fang, Huang, and Karpoff (2016) require a balanced sample over 2001-2010, resulting in a sample with 388 pilot and 709 control firms, versus the available unbalanced sample of 702 pilot and 1,413 control firms (see Part II.B).

The above three aspects of the SEC experiment let us test whether the significant results reported in the 10 indirect studies we reexamine represent only different, reasonable specification choices made in these studies (relative to our simple specification) or instead reflect a search for specifications that produced significance when other reasonable specifications would not (p-hacking). To test for specification search (p-hacking), we follow the "specification curve" framework proposed by Simonsohn, Simmons and Nelson (2020) and test whether the specifications that result in statistical significance in the predicted direction are extreme, i.e., they result in *t*-statistics that are too large (reflecting larger coefficient magnitudes, smaller standard errors (s.e.'s), or both, when compared to our benchmark specification or to best-match specifications (which closely follow the sample and specification choices reported in the respective papers).

Using a pre-specified research design, including a sample that closely matches the SEC experiment, we assess the robustness of the key reported outcomes attributed to the short-sale

experiment, and related outcomes in 10 indirect effects studies.[1] Overall, we examine 78 specifications involving 63 distinct outcomes. Of these, 37 specifications (32 outcomes) were reported by these studies (reported results, henceforth); we also examine related outcomes and specifications to assess the robustness of the reported results in the re-examined papers.

With our specification, almost none of the results are significant at the conventional 5% significance level. More surprisingly, when we conduct "best match" analyses (where we closely follow the sample selection procedure and research design choices stated in the respective papers), the "best match" results are still almost never statistically significant and are generally close to the results with our specification. In contrast, almost all the outcomes are significant in the re-examined studies. The contrast between the reported results and the results from the best match specifications suggests that these studies made unreported specification choices that produce statistical significance while ours and the best-match specifications did not produce significant results. In this sense, the reported results are extreme and hence suggestive of specification search.

Consider, for example, the likelihood of obtaining a significant coefficient at the 5% level. With our specifications, we obtain a 5%-significant coefficient, with the predicted sign, in three out of 78 specifications. For the best match specifications, five out of 77 specifications produce a significant coefficient at the 5% level. A binomial test cannot reject the null hypothesis that the results from our specification and the best-match specification come from the same underlying pool of reasonable specifications. A Kolmogorov-Smirnov (K-S) test for whether ours and the best-match results could come from the same underlying distribution also fails to reject the null ($p = 0.24$). Moreover, these small numbers of significant coefficients could arise by chance. In contrast, for the reported results, 29 of 37 coefficients are significant at the 5% level. The K-S test strongly rejects the null that the reported results were drawn from the same distribution as either our specification or the best-match specifications ($p < 0.000000001$ for both comparisons).

We also explore the reasons for the differences in results across our specification, the best match specifications, and the reported results. We find large differences between the best match and reported results for both coefficients and s.e.'s. The reported coefficients are larger, sometimes much larger than the coefficients from either ours or the best match specifications. Also, the reported s.e.'s are often smaller, sometimes much smaller than those obtained from our

[1] These 10 papers rely on standard databases such as CRSP and Compustat, making it feasible to conduct detailed re-examination. Some results presented here for four of these studies are drawn from BDLYY (2024).

specification or the best match specifications. The gaps between the best-match specifications and reported results indicate that the re-examined studies made important specification choices that affected coefficient magnitudes and s.e.'s, but these choices were not reported in the paper and hence not captured in our best-match analysis.[2]

Our results have important implications for academia as well as for policy making. A standard hierarchy for the credibility of research designs, as a guide to causal inference, puts randomized experiments at the top, natural experiments next, and both ahead of designs in which people, firms, or other units of observation can self-select into the treatment. Greater confidence in inferences from randomized trials or natural experiments is partly due to exogenous selection of which units are treated and the belief that these research designs limit researcher discretion. Yet we find that even the results for a randomized experiment are highly sensitive to specification choice and thus vulnerable to specification search.

Given that the Reg SHO experiment was conducted to guide SEC rulemaking, the issue of robustness of the results in papers that study this experiment takes on heightened significance, especially as several papers attribute wide-ranging adverse effects to the experiment. Finally, our findings of large differences between the best-match results and reported results underscore the importance of transparent disclosure of methodology and research design in academic papers.

This paper proceeds as follows. Section II summarizes the studies we examine, our sample selection steps, and other specification choices. Section III summarizes our pre-specified research design. Section IV presents our re-examination results, using our specification. Section V presents results for the best match specifications. Section VI presents statistical tests for evidence of specification search. Section VI provides implications and conclusions.

II. Summary of Studies, Sample Selection, and Specification Choices

A. Summary of Re-Examined Studies

We re-examine the following ten studies. The first four were also considered in our prior work (BDLYY, 2024): Fang, Huang, and Karpoff, 2016 (FHK); Grullon, Michenaud, and Weston, 2015 (GMW); Hope, Hu, and Zhao, 2016 (HHZ); and Lin, Liu and Sun, 2019 (LLS). We add six

[2] For three studies (Fang, Huang and Karpoff, 2016; Hope, Hu and Zhao, 2016, and Chen, Fu, and Wang, 2023), we have the authors' exact sample, and for HHZ, their sample selection steps. This lets us examine how the authors' specification choices produced significant results when other reasonable choices would not. See Internet Appendix, Parts III.A-C, F for replication of these studies.

additional studies (Bui, Hasan, Lin, and Nguyen, 2023 (BHLN); Chen, Fu, and Wang, 2023 (CFW); Chen, Zhu, and Chang, 2017 (CZC); Gong, 2020; Kim, Lu, and Peng, 2020 (KLP); and Tsai, Wu, and Xu, 2021 (TWX). We chose these studies for the following non-mutually exclusive reasons: (i) they were published in peer reviewed journals, mostly strong, respected journals;[3] (ii) they rely on different causal channels; (iii) they provide evidence for a variety of indirect effects; and (iv) their core results can be reexamined using data from standard sources, which both eases the task of reexamination and permits replication by others.[4] We cannot, of course, know how representative the ten reexamined papers are for the remaining indirect effects studies, but have no reason to think that they are not representative.

We briefly summarize the core conjectures and key outcomes in the studies below.

FHK conjecture that in response to a greater threat of short selling, pilot firms would reduce their earnings management and thus have lower performance-matched discretionary accruals (PMDA), and lower likelihood of a future restatement (measured using "HF-score").

HHZ conjecture that suspension of price tests would increase short selling at the pilot firms. This would increase auditors' litigation risk. The auditors, anticipating higher litigation risk, would increase audit fees.

GMW conjecture that short sale constraints result in overvaluation of firms, which leads to overinvestment. They conjecture that removing short sale constraints should reduce share prices, which will lead to lower investment and lower capital raising.

LLS argue that relaxing short sale restrictions make stock prices more informative and that with more informative prices to guide managerial actions, shareholders will perceive lower need for direct incentives, so pilot firms will have lower sensitivity of CEO wealth to performance, and investment decisions will become more sensitive to share price.

CFW and TWX study total investment. Both report evidence that pilot firms reduce both overinvestment and underinvestment.

KLP conjecture that pilot firms will choose more conservative tax policies to reduce the risk of being targeted by short sellers.

---

[3] The sole exception is KLP which is still a working paper.

[4] There were no additional papers we reexamined but do not discuss, except Wang (2018), an unpublished working paper studying cash ratio. We found a negative, insignificant coefficient on Pilot*During, versus the positive, significant coefficient reported by the author.

CZC conjecture that the pilot firms, to ward off attacks by short sellers, increase payouts to shareholders. They study dividends and share repurchases.

Gong conjectures that pilot firms will reduce leverage to reduce short selling threats.

BHLN report that pilot banks increased risk-taking during the financial crisis period of 2007-2009, a period with fiscal year-ends almost entirely *after* the experiment ended in July 2007.

B. Sample Selection

We summarize our sample selection steps here; please see our pre-analysis plan (Black et al., 2019) and the Online Appendix for additional details.[5] We exercised great care in constructing a sample that is both true to the SEC experiment design and substantially larger than those in the re-examined studies. The more complete sample provides one reason why results with our specification diverge from those in the re-examined papers.

We start with the R3000 list as of June 30, 2004, merged with CRSP and Compustat; this match preserves all sample firms. The SEC published a list of 986 pilot firms but never published a full list of control firms. To create the pilot and control samples, we follow the SEC's exclusion rules and exclude 32 firms that were not listed on NYSE/AMEX or the Nasdaq national market and 12 firms that began trading after April 30, 2004. This leaves 2,954 firms comprising 985 pilot firms (the SEC's 986 pilot firms less one firm that delisted on June 28, 2004) and 1,969 control firms, as of experiment announcement on July 28, 2004.

We next update this sample to the start of the experiment on May 2, 2005, using the SEC's updated lists on April 13, 2005, of pilot firms and a subset of control firms, principally to exclude firms due to mergers, acquisitions, and bankruptcy. This leaves 943 pilot and 1,891 control firms.

We then exclude financial firms (SIC 6000-6999), utilities (SIC 4900-4999) and three firms that did not file 10-Ks for fiscal 2004, resulting in a "2005 Financial Analysis Sample" of 2,115 firms (702 pilot; 1,413 control). Of the re-examined papers, BHLN study only financial firms; all others except LLS exclude financial and utility firms.

[5] The pre-analysis plan was developed for the first four re-examined papers. For the remainder, we retained the same sample, time periods, and other design choices. The principal changes were the specific outcomes studied and, for some outcomes, deciding for which variables when to replace missing Compustat values with 0, and when to leave them missing. We made these decisions before running any regressions on the outcome variables.

C. Other Specification Choices

We summarize here other specification choices made in our pre-analysis plan.

1. Firm and Year Fixed Effects

It is standard, in DiD analysis of a shock at a single point in time with panel data, to include unit and time fixed effects (FE), in a so-called "two-way fixed effects" (TWFE) model. We use the TWFE design, organize firm-year observations based on fiscal year, and use firm and fiscal year FE. All re-examined papers use firm FE except FHK. Most also use year FE; the exceptions are FHK and HHZ, who use Pre, During, and Post dummies. See details in Internet Appendix Table IA-3. We confirm in the Appendix that results for our and the best-match specifications are similar without firm FE.

2. One-Versus Two-Way Clustering

It is standard to cluster standard errors (s.e.'s) on firm, and we do so. So do HHZ, GMW, LLS, TWX, and BHLN. However, FHK, CZC, Gong, KLP, and CFW cluster on both firm and year. Two-way clustering can produce downward-biased standard errors when the time-series is short and the number of clusters is small (Cameron, Gelbach, and Miller, 2008). Below, we compare results with two-way clustering to the "truth", where p-values are obtained from randomization inference, and show that downward bias can be severe.

3. Use of Covariates

Our research design does not include time-varying covariates. The re-examined papers use a wide range of covariates. Some also report univariate results (FHK, HHZ, GMW, CZC), or results with firm and year FE but without covariates (LLS, CFW, Gong, KLP); but TWX and BHLN only report results with covariates. We prefer results without covariates. First, given the initial randomization, estimates without covariates are unbiased. Second, including covariates will introduce bias if the covariates are intermediate outcomes of treatment. Given the wide range of proposed indirect outcomes of the SEC experiment, we cannot think of important covariates that are not potential intermediate outcomes. Third, not requiring data on covariates preserves sample size. We confirm in the Appendix that results for our and the best-match specifications are similar using the covariates used in each paper.

4. Balanced versus Unbalanced Sample; Requiring R3000 Membership in 2005

We rely on an unbalanced panel. We confirm that sample attrition is balanced between pilot and control firms. Thus, there is no reason to impose survivorship bias and accept significant loss in sample size from using a balanced panel. Of the re-examined papers, only FHK use a balanced panel. We study the FHK outcomes using both unbalanced and balanced panels.

HHZ, GMW, and CFW induce survivorship bias in a different way: they require sample firms to still be included in the R3000 when the experiment starts in May 2005. This excludes small firms whose market value fell by enough, between 2004 and 2005, to cause them to drop out of R3000.

5. Sample Periods and Inclusion of a Post-Experiment Period

We measure firm performance over fiscal years. So do seven of the re-examined papers. FHK and CZC use calendar years, and Gong uses fiscal quarters. Each re-examined paper makes a different choice on how to define the Pre-experiment ("Pre"), during-experiment ("During"), and post-experiment ("Post") periods; whether to include or drop an "interim" period between the SEC announcement and the experiment start; and whether to include a Post period in the study. We believe the logic for studying a Post period is compelling: any treated-minus-control difference during the experiment should reverse after the experiment ends.

We use fiscal years 2001-2010 as the sample period; thus, roughly four pre-experiment years, three experiment years, and three post-experiment years. We use the Compustat convention, under which, for fiscal year ending January-May, fiscal year = (calendar year – 1); otherwise, fiscal year = calendar year. Thus, our sample period includes fiscal year ending June 2001 through May 2011. We also need to map fiscal years to the Pre, During, and Post periods. The experiment was announced on July 28, 2004, and ran from May 2, 2005, to July 6, 2007.

Judgment is needed on how to map fiscal years to Pre, During and Post. We chose to treat firm fiscal years for which half or more of the fiscal year falls within the experiment period as within the During period (thus, fiscal years ending October 2005 through December 2007). Earlier fiscal years are part of the Pre period, and later fiscal years are part of the Post period. Among the re-examined papers, all except GMW, TWX, and KLP exclude an interim period (2004). There is danger in excluding an interim period, however. If treated and control firms show non-parallel pre-treatment trends, presumably for random reasons, removing 2004 could lead to a spurious

finding of a treatment effect. We find this to be the case for several of the re-examined papers (LLS, CFW, TWX).

6.. Handling Variables: Winsorization and Missing Values

We follow the common practice and winsorize continuous variables from Compustat at 1% and 99% (Hribar and Nichols, 2007). This ensures that neither outlier values nor Compustat data errors will influence the results. Eight of the re-examined papers make a similar choice. LLS winsorize their wealth-performance sensitivity (WPS) outcome but not other outcomes or covariates; and Gong is silent on winsorization.[6]

For a few Compustat variables, missing values may reflect true zeros. When this seems likely, we replace missing values with zero. We replace missing with zero, as relevant for the following, in measuring both outcomes and covariates: R&D expenses, acquisitions, sales of PPE (property, plant and equipment), dividends, and share repurchases. Eight of the re-examined papers are silent; we assume that they do not replace missing values with zero (the exceptions are TWX and KLP).[7] Failure to do so, for variables with many true zeros, can produce large loss of sample size, and skewed loss, since firms with missing values will not be a random set of all firms.

III. Research Design

We summarize our research design below. See BDLYY (2024) and our pre-analysis plan for additional details. Given that the SEC experiment was a randomized trial, with a good covariate balance between pilot and control firms, our core approach, a simple DiD analysis with firm and year fixed effects, but without time-varying covariates, is unbiased and reasonable.

A. Simple DiD Specification

We estimate the following DiD model for each outcome over 2001-2010.

$$y_{i,t} = \beta_0 + \gamma_t + f_i + \beta_1 Pilot_i * During_t + \beta_2 Pilot_i * Post_t + \varepsilon_{i,t} \quad (1)$$

[6] CFW are also silent but confirmed in response to our inquiry that they winsorize continuous variables at 1%/99%.

[7] KLP say that they replaced missing R&D with zero; we infer that they did not replace missing with zero for other variables. CFW advised us, in response to our inquiry, that they replace missing values with zero for the variables they use to construct their total investment outcome: capital expenditure (CAPX), R&D expenditure (XRD), acquisition (AQC), and sale of PP&E (SPPE).

Here $y_{i,t}$ is the outcome; $Pilot_i$ = 1 for pilot (treated) firms and 0 for control firms; During is a dummy variable for the experiment period; Post is a dummy variable for the post-experiment period, and the γ_t and f_i are year and firm FE. Non-interacted terms are absorbed by the firm and year FE. The estimated coefficient $\widehat{\beta_1}$ on Pilot*During measures the treatment effect. The estimate $\widehat{\beta_2}$ on Pilot*Post should be close to zero because short-sale restrictions were removed for all firms following the experiment. We test for sign reversal in the Post period, relative to the experiment period, by replacing Pilot*During with Pilot*(During or Post) in equation (1). In the Internet Appendix, as robustness checks, we also use a specification with covariates, in which we add $\boldsymbol{\lambda}^*\mathbf{x}_{i,t}$ to eqn. (5), where $\mathbf{x}_{i,t}$ is a vector of covariates (for each i, t) and $\boldsymbol{\lambda}$ is a coefficient vector.

B. Univariate Annual Differences and Leads-and-Lags Specification

Good DiD practice includes presenting univariate graphs for the treatment and the control groups, leads-and-lags graphs, or both. These graphs can be used both to confirm whether pre-treatment trends appear parallel, and to assess how the treatment effect evolves during the treatment period (e.g., Atanasov and Black, 2016). Univariate graphs showing time trends for the treatment and control groups can also be valuable, especially if, as here, there is a good balance between the two groups.

Examining time trends during the experiment period is especially revealing for the SEC experiment because if there were any adverse effect of the experiment, they should have become apparent over time. In fact, there does not appear to be any evidence to suggest that there were market disruptions caused by the experiment or that there were complaints from the firms.[8] Thus, at a minimum, by 2006, it should have become apparent to the firms that there were no major effects of the experiment. Thus, if the effects show up in 2006 or 2007, it is difficult to meaningfully attribute it to the experiment.

Both to assess whether pre-treatment trends are parallel and to allow for treatment effect to emerge gradually during the experiment period, we use two graphical approaches. First, we use univariate differences in means between pilot and control firms. When the data are available, we extend these plots back to 1998. A longer pre-treatment period is useful in identifying non-parallel

[8] Speaking at a conference at Pace University on October 19, 2005, about six months after the experiment started, SEC Commissioner Annette Nazareth stated that she was not aware of any complaints about the program. Speaking at the same conference, Chief Regulatory Officer of NYSE, Richard Ketchum stated that "From an economic standpoint, we are not seeing market impact to suggest that there are serious problems".

pre-treatment trends, and in assessing the extent of presumably random pre-experiment variation in means between the two groups. Second, we use a "leads-and-lags" specification, in which we estimate a separate "treatment effect" for each year and plot annual coefficients and 95% confidence intervals (CIs) in leads-and-lags graphs. Inference from both is similar, as expected given the randomization. We provide univariate graphs in the text for selected outcomes for each re-examined paper. The Internet Appendix includes additional univariate graphs and leads-and-lags graphs.

C. Triple Difference Specification

For some outcomes, LLS CFW, and TWX use a triple difference specification, with Pilot*During interacted with another variable (Tobin's q for LLS; a measure of overinvestment for CFW and TWX). Let $q_{i,t}$ be this additional variable. In the triple difference specification, we replace eqn. (1) with:

$$y_{i,t} = \beta_0 + \gamma_t + f_i + \beta_1 * q_{i,t} + \boldsymbol{\gamma} * \boldsymbol{DBL_{i.t}} + \beta_2 * q_{i,t} * Pilot_i * During_t + \beta_3 * q_{i,t} * Pilot_i * Post_t + \varepsilon_{i,t} \quad (2)$$

Here q is the other variable; **DBL** is a column vector of the double interactions ($Pilot_i$ * $During_t$; $Pilot_i$ * $Post_t$; $q_{i,t}$ * $Pilot_i$; $q_{i,t}$ * $During_t$; and $q_{i,t}$ * $Post_t$), and **γ** is the corresponding row vector of coefficients. The core coefficients are those on the triple interactions, β_2 and β_3.

IV. Results with Our Specification

We begin with graphical assessment, using univariate graphs, based on our specification (our sample and research design). We study 63 distinct outcomes and 78 different specifications. Given covariate balance between the treatment and the control firms due to randomization, univariate graphs should be informative about existence of a treatment effect. In Figure 1, we plot 11 core outcomes -- one for each study (two for FHK).

None of the 11 graphs show convincing evidence of a treatment effect. Across all the graphs, the outcomes for the pilot and the control firms are very similar during the experiment period. This is not surprising as there is no evidence of any direct effect of the suspension of price tests on the pilot firms as discussed earlier. Some of the graphs show non-parallel trends in the Pre

period.[9] For example, Panel E shows the graph for Wealth Performance Sensitivity (WPS) for LLS. WPS for pilot firms declines from 2003 to 2004 relative to controls: with essentially minimal change in relative WPS from 2004 to 2006. To measure the change in WPS from the Pre to the During period, LLS drop 2004. This choice drives their regression result that WPS declines for pilot firms during the experiment. If 2004 is included in the Pre period, there is no evidence of a decline in relative WPS during the experiment period (see also Table 1, Panel E).[10] For several papers, the graphs show erratic treatment-vs-control gaps in the Pre and/or Post periods.[11] For some, pilot-vs-control gaps are largest in the Pre and/or Post period.[12] These patterns suggest that similar gaps, which appear during the treatment period, could reflect random fluctuations, rather than a true treatment effect. Also, most of the graphs show no reversal of any pilot-versus-control gap in the Post period; even though a true treatment effect should reverse in the Post period. Finally, some graphs show evidence for a treatment effect only for 2007 (the wrong time for a treatment effect to appear).[13]

Next, we turn to regression results, summarized in Table 1. With our specification, 31 coefficients (out of 78) have signs opposite to the predicted sign (shown in red). Four are statistically significant at the 5% level (shown in boldface); three with the predicted sign and one with the opposite sign. Of the three with predicted sign, two are from CFW and TWX, who study the same outcome. The third is for LLS, for which the DiD coefficient is driven by non-parallel pre-treatment trends as discussed above. Even leaving aside the lack of graphical evidence for the treatment effect, these few significant results could easily arise by chance.

Given the randomized nature of the trial and covariate balance, graphical evidence and a simple specification where an outcome is regressed on the treatment dummy with firm and year fixed effects is a reasonable place to start. That we find virtually no evidence for a treatment effect across so many outcomes from 10 studies is remarkable. In contrast, the re-examined studies report 37 specifications (32 outcomes) and find statistical significance for 29 of these specifications. This

[9] LLS, CFW; TWX; KLP.

[10] LLS rely on a price efficiency channel to explain their results. That channel cannot arise before the experiment start in May 2005, so they should have included 2004 in the Pre period.

[11] FHK accruals; FHK HF-score. For example, FHK's PMDA measure shows a relative rise for pilot firms in 2000 (driven by a drop for control firms) and a relative jump in 2008 for pilot firms.

[12] LLS, HHZ; GMW; CZC; Gong.

[13] CZC, Gong.

is initial evidence that many of the reported results may reflect specification search. While our specification differs from the re-examined studies, we started with *their chosen outcomes*, plus related outcomes. Exploring why our results are so far from the reported results is our main goal for the remainder of this paper.

V. Moving from Our Specification to the Best-Match Specifications

A. Overview of Best-Match Results

To understand why results with our specification are so different from the reported results, we conducted a best-match specification analysis for each paper, where we seek to match, as best as we can, the specification and sample in each paper based on the authors' own description in the papers.[14] We report best-match regression results in Table 2 and annual univariate graphs in Internet Appendix Figures IA-5 to IA-13.[15] For papers that use two-way clustering, we report results with both one-way and two-way clustering, to highlight the importance of this choice.

With the best-match specifications, there is a tendency for coefficients to be closer to the reported results than those with our specification. A few more coefficients have the same sign as predicted for best match (53/77) versus for our specification (50/78).[16] However, most of the best-match coefficients remain insignificant and often their magnitudes are far from the reported coefficients. With s.e.'s clustered on firm, the best-match specifications produce weak significance for 2/77 coefficients with the predicted signs (one for GMW, one for LLS). The small number of significant coefficients could easily arise by chance. Moreover, the GMW result is unconvincing given the univariate graphs showing larger treatment-vs-control gaps in the Pre period, and similar gaps in the Post period (Figure 1). Also, the GMW result is driven by Nasdaq firms, where one should not expect an effect as the bid test was weak to begin with. The LLS result is driven by non-parallel pre-treatment trends and their choice to exclude 2004 from the Pre period.

Using two-way clustered s.e.'s for the five studies that use them, three coefficients become significant for FHK HF-score. The best-match specifications now produce 5/77 significant

[14] For FHK, HHZ, and CFW, we later obtained their replication data and code. Our best-match analysis was developed, as for the other papers, based solely on their papers without knowing their actual choices and sample.

[15] The Internet Appendix also includes results from step-by-step moves from our specification to the best-match specifications and leads-and lags graphs for our specification and the best-match specification.

[16] One best-match specification, for BHLN, is mis specified.

coefficients with predicted sign.[17] In the Internet Appendix, we move step-by-step from our specifications to the best-match specifications. These results are also almost always insignificant.

When we began this project, we expected that many results would not be robust using our specification but expected to come much closer to the published results using the best-match specifications. Instead, across studies and outcomes, we often find very different best-match results versus reported results, for coefficients, standard errors and often both. The gap between best-match and reported results is suggestive of specification search, combined with incomplete disclosure of the research design choices in the papers.

B. Comparing Reported Results to our Specifications and Best-Match Specifications

Table 3 has the same format as Tables 1 and 2 but is limited to the specifications for which the papers report results. The picture changes dramatically. Now, 36 of 37 coefficients have the predicted sign; of which 29 are significant and 5 more are marginally significant.

In Figure 2, we summarize the differences in coefficient signs and statistical significance from Tables 1-3. This figure reports *t*-statistics for our specification, the best-match specifications (using one-way clustering) and the reported results, for the 37 reported outcomes across the 10 papers. The figure shows *t*-statistics as positive (the absolute value) except that we report negative *t*-statistics for our specification and the best-match specifications when the coefficient sign is opposite from the reported sign. The reported results are winsorized at 4, to avoid compressing the X-axis. With our specification and the best-match specifications: almost all coefficients are insignificant; many have the opposite sign from the reported results; and only a few coefficients are significant, most with *t*-statistics only slightly above +1.96. Moreover, there is a large, visually apparent difference between the reported results and those with either our specification or the best-match specifications.

In Figure 3, we show the signs of the coefficients and *t*-statistics for our specification and the best-match specifications for the other 41 specifications, which we studied as robustness

[17] To summarize: FHK accruals: 0 significant results across 8 specifications. FHK HF-score: 0/12 (3 are significant with two-way clustered s.e.'s). HHZ audit fees: 0/3. GMW: 1/8 outcomes. LLS: 1/6. CFW: 0/8 using CFW *adjusted* best-match specification. TWX: 0/ 8. KLP: 0/9 specifications (6 outcomes) (treating BTD and BTD_{adj} as two specifications for the same outcome; and similarly for NetR3, $NetR3_{adj}$, and $NetR3_{truncated}$). CZC. 0/6. Gong: 0/2. BHLN: 0/7 (an 8th outcome is mis-specified, book leverage is both outcome and a predictor).

checks for the reported specifications.[18] With both our specification and the best-match specifications, almost all of the coefficients are statistically insignificant; a fair number have the opposite sign from the sign that would be expected based on the reported coefficients. With our specification, only 2/41 coefficients are statistically significant, of which one (Capex for LLS) has the opposite sign from predicted. With the best-match specifications, only one (R&D/assets, for CFW) is significant.

Figure 4 uses a histogram to summarize the statistical significance of all coefficients from Figures 2 and 3. The histogram bars are reversed for our specification to make it easier to present three histograms in one figure. Results from our specifications and the best match specifications are broadly similar, but both are far different from the reported results.

C. Differences in Coefficients versus Differences in Standard Errors

We next explore the sources of the differences in statistical significance. The larger *t*-statistics for the reported results can come from larger reported coefficients, smaller reported s.e.'s, or both. We show in the next three figures that both contribute to larger reported *t*-statistics in the re-examined papers. Figure 5 reports regression coefficients for our specifications, scaled to +1.00 if our coefficient has the same sign as the reported coefficient and to -1.00 if our coefficient has the opposite sign. It also reports coefficients for the best match specifications, scaled relative to those from our specification (scaled coefficient = best-match coefficient/our coefficient) and the reported coefficients (similarly scaled). The best-match and reported results are winsorized at -2 and +6. Many coefficients with our specification have the opposite sign to those reported (also seen in Figure 2); so do a number of best-match coefficients. The best match coefficients tend to be closer to the reported coefficients, relative to our specification**.** But the dominant impression from Figure 5 is that reported coefficients are much larger in magnitude than ours, and often much larger than best-match coefficients.

---

[18] The additional specifications are: For FHK-accruals, we study both balanced and unbalanced panels, and both PMDA and three simpler measures of accruals, operating accruals, total accruals, and abnormal accruals. For FHK HF-score, we study both balanced and unbalanced panels and study both HF-score and the underlying F-score. The papers that study investment, GMW, LLS, CFW, and TWX, use varying measures. Capex/assets; (Capex + R&D)/assets, R&D/assets, and total investment. We study all of these measures, for all four papers. For LLS, who study R&D/assets, we also study R&D/sales. GMW study percent asset growth; we also study *ln*(assets). One KLP measure is book-tax-difference; we also study a measure which has non-missing values for firms without foreign operations. CZC study unscaled dividends and repurchases. We also study payout (dividends + repurchases), and these measures scaled by assets. BHLN study book and market leverage. They say they study debt/market value of *assets*, but likely study debt/market value of *equity*. We study book and market leverage using both denominators.

In Figure 6, we use a similar scaling approach for s.e.'s. The figure reports standard errors for our specification (scaled to 1.00), s.e. for the best-match specifications (scaled to our s.e., with clustering on firm) and reported s.e. (clustering varies). The best match and reported s.e. are winsorized at 2. Best-match s.e.'s tend to be *larger* than those with our specification. However, reported s.e.'s are mostly smaller than s.e.'s using either ours or the best match specification**.** One reason for the smaller reported s.e.'s is that five of the papers use two-way clustering.

Figure 7 explores an additional reason why reported results differ from those with either our specification or the best-match specifications: The reported results often use smaller samples, sometimes much smaller. Panel A shows the number of *firms* with our specification (scaled to 1.00), and scaled numbers for the best match and reported specifications. Six of the 10 studies make explicit sample-selection choices that lead to loss of firms; this leads to best-match sample size being smaller than our sample. More surprisingly, all studies except BHLN lose additional firms in the reported results relative to the best match specifications. In the figure, this leads to reported counts being to the left of the best match counts.[19] The additional loss in sample size likely plays a role in the difference between best match and reported results.

To summarize, using a standard, simple DiD specification with firm and year fixed effects, there is no evidence of a treatment effect for any of the reported outcomes. When we move to best match specifications, almost all the results remain insignificant and for the significant ones, the significance is driven by either non-parallel trends or the year 2007 (wrong time for a treatment effect to emerge), or both. The reported results have substantially higher coefficients, and often lower standard errors, than those obtained with either our specification or the best-match specification. The additional specifications we study as robustness checks confirm lack of evidence for treatment effect. The reported results also rely on fewer observations than either ours or the best match sample. Overall, our analysis provides evidence that these 10 studies made sample selection and specification choices that were critical to obtaining statistically significant results, but these choices were not reported and thus they were not captured by our best match analysis.

[19] The BHLN counts are higher than ours mainly because BHLN include in their sample banks that were not in their sample during their Crisis (treatment) period. This is an error; these banks should have been dropped.

D. Close Analysis of Each Re-examined Study

In the Internet Appendix, we provide close analysis of each re-examined study. This includes annual univariate graphs for all outcomes, with both our specification and the best-match specifications, and intermediate regressions in which we move step-by-step from our specification to the best-match specifications. We also document the loss of sample size relative to our specification and discuss the often-questionable theoretical basis for the reported results.

VI. Evidence of Specification Search

The evidence above is strongly suggestive of specification search. In this section, we seek to provide statistical evidence that the reported results reflect specification search rather than file drawer/publication bias.

To provide direct statistical evidence on specification search, we adopt the following approach based on Simonsohn et al. (2020). We posit that there is an underlying pool of many reasonable specifications from which our specification, the best match specifications, and the reported results could be drawn, resulting in a "specification curve." We assess whether the three sets of results could plausibly have been drawn from the same distribution, as would be expected if all results reflect different specification choices rather than specification search. The idea is that if the reported results come from the same underlying distribution as those from our specification or the best-match specifications, they could simply reflect "lucky" specification choices, that happened to generate statistical significance, rather than affirmative search. Conversely, if the reported results are extreme and could not plausibly come from the same distribution as ours or the best-match specifications, this provides evidence suggests that the reported results reflect specification search, not just lucky specification choice.

A. Binomial Tests for Same Sign as Predicted

Our first approach to assessing whether all three sets of results could come from the same underlying distribution relies solely on whether the coefficients from our specifications or the best match specifications have the same sign as predicted in the re-examined papers. We use a simple binomial test for equal proportions. We assume that our 78 results across the 10 papers, the 77 best-match results (120, if we treat one-way and two-way clustered s.e.'s as separate specifications for papers that use two-way clustering), and the 37 reported results reflect *independent* draws from an underlying binomial distribution, where for each draw, the coefficient has a probability p of

having the predicted sign. The assumption of independent draws is not correct as the outcomes within each paper are related and thus non-independent. We first present the sign test results and then explain why non-independence implies that our results are conservative.

In Figure 8, Panel A, we present box-and whiskers plots for the draws from binomial distributions (1 million draws for each), for whether the observed coefficient sign is the same as predicted. The box for each plot shows the 25th percentile, mean, and the 75th percentile. The inner whiskers show the 2.5 and 97.5 percentiles, and the outer whiskers show the 0.01 and 99.99 percentiles. Grey dots show the most extreme percentages from the 1 million draws. Black triangles show the most extreme percentages from 1 billion draws. For example, for our specification, 47/78 results (60%) have the same sign as those predicted in the underlying papers. We take 1 million draws from a binomial distribution with a 47/78 chance of a "1" outcome and plot the results. Black triangles show the most extreme percentage from 1 *billion* draws.

Panel A shows that the best-match specifications are more likely than our specification to have the same sign as reported, but not strongly so. The 95% CIs overlap, so we cannot rule out that both our and the best-match results were drawn from the same underlying distribution. However, as we discuss below, there is a very low likelihood that the reported results came from the same distribution as either ours or the best match specifications.

We can formally assess the likelihood that each outcome would be observed, if the true proportion came from one of the other distributions. For example, if truth were that 47/78 coefficients have the same-sign-as-predicted (the observed proportion in our specifications), we can assess the probability that one would observe at least 53/77 coefficients with the same-sign-as-predicted (observed for the best-match specifications) or at least 36/37 (observed for the reported results).

We estimate these probabilities by taking 1 million draws from each proportion and provide results in the small table below. The results suggest that if the true base probability was 47/78 coefficients with same sign as predicted, the likelihood that we would observe at least 53 of 77 coefficients with the same sign as predicted is 6.41%. Conversely, if the true base probability were 53/77, there would be a 5.77% likelihood of observing 47/78 or fewer results with the same sign as predicted. Consistent with Figure 8, Panel A, ours and the best-match distributions overlap and thus could have been drawn from the same underlying distribution.

However, there is a near-zero probability that either (a) the reported results and those from our specification come from the same distribution; or (b) the reported results and the best-match results come from the same distribution. This non-overlap suggests that the reported results reflect extreme choices that were not disclosed in the paper and hence reflect specification search. The inference supporting specification search is especially strong when comparing best-match to reported results, because the best-match results already reflect all specification choices that the papers reported. In a world in which researchers search for statistical significance, the best match results should already be tilted, perhaps strongly, toward finding significant results.

Results from 1 million draws using each base probability

Base specification	Consistent signs	Base prob (s.e.).	Target specification	Target outcome	Prob of target outcome
Ours	47/78	0.603 (0.055)	Best-match	≥ 53/77	0.0641
			Reported	≥ 36/37	0
Best-match	53/77	0.688 (0.053)	Ours	≤ 47/78	0.0577
			Reported	≥ 36/37	0.00001
Reported	36/37	0.973 (0.027)	Ours	≤ 47/78	0
			Best-match	≤ 53/77	0

For our specification, the surprising result is not that we obtain the same sign as reported more than half of the time, but that we so often obtain the opposite sign. For the best-match specifications, the surprise is not that the best-match results move toward the reported results, relative to our specification, but that they remain so far from the reported results, including often having the opposite sign.

B. Likelihood of Statistical Significance at Different Significance Levels

In the remaining panels of Figure 8, we proceed similarly, except now the "1" outcome indicates whether the observed coefficient has the same sign as predicted *and* is statistically significant at the α% level or better. We consider α = 5% (Panel B), α = 10% (Panel C), and α = 1% (Panel D). Each panel includes a dashed vertical line at the α level.

For the 5% significance level (Panel B), we get significant coefficient in three out of 78 instances with our specification, and five out of 77 coefficients with the best match specifications. We can further decompose the best match results based on clustering choice. The probability of finding 5% significance are 2/77 with one-way clustering, 5/76 with the actual best-match clustering choice made in each paper; and 5/120 treating one- and two-way clustering as separate specifications for the papers that use two-way clustering. In contrast, 29/37 reported coefficients

are significant at the 5% level. The plots in Panel B show that the results for our specifications and the best-match specifications could plausibly come from the same distribution (the CIs overlap) and that both sets of results could arise by chance (the CIs include 5%). However, the reported results are far outside the distributions for ours and the best match specifications and thus could not plausibly have come from the same underlying distribution. Similar inference can be drawn from Figure 8, panels C and D which report results for significance at the 10% level and 1% level, respectively.

C. Likely Conservative Nature of Differences in Probabilities

The results reported above provide, if anything, conservative measures of the likelihood of specification search, because we treated each outcome as independent of all other outcomes. This is surely false. First, for each paper, we study a number of related outcomes. This makes it more likely that signs and statistical significance will be similar across outcomes. Second, for the reported outcomes, we are studying the same outcome across ours, the best match, and the reported specifications. Both factors should bring the proportions from these three approaches closer together. Thus, the differences between the reported results and the other specifications should understate the extent to which the reported results reflect specification search.

D. Kolmogorov-Smirnov Test for Similar Distributions

In Figure 9, we consider an alternative approach to assessing whether our specification, the best-match specifications, and the reported results could have been drawn from the same underlying pool of reasonable specifications. We compare cumulative distribution functions (CDFs) for the *t*-statistics from our specification, the best-match specifications, ours and best-match combined, and the reported results. As in Figures 2-4 4, t-statistics are reported as positive, except that we use negative *t*-statistics when the coefficient sign is opposite from expected based on the reported results. We also report Kolmogorov-Smirnov test statistics and p-values for the probability that both CDFs were drawn from the same underlying distribution.

Panel A compares CDFs for our specifications and the best match specifications. The two are reasonably similar; the K-S test cannot reject the null that they reflect draws from the same underlying distribution ($p = 0.24$). Panel B compares CDFs for our specification and the reported results. The two distributions are visually very different; the K-S test strongly rejects the null ($p = 0.000000000$). Panel C compares CDFs for the best-match specification versus the reported

results ($p = 0.000000000$); Panel D compares the combined results from our specification and the best-match specifications to the reported results ($p = 0.000000000$).

Thus, the results of the K-S test confirm the inference from the binomial tests that the reported results could not plausibly have come from the same underlying distribution as either our or best match specifications.

E. Are the Specifications in Reported Results Somehow Preferred over Ours or the Best-Match Specifications?

Can it be argued that the specifications that produced significant reported results are preferred over either ours or the best-match specifications? We do not believe so, for a number of reasons. First, any choices that the authors deemed important should have been discussed in the papers and thus captured in the best-match specifications.

Second, in Internet Appendix Tables IA-5 to IA-17, we methodically move step by step from our pre-specified design to each paper's best-match specification. Those steps provide an additional, much larger pool of draws from the unobserved pool of reasonable specifications. The results are usually intermediate between ours and the best-match specification, and rarely far from either. This reinforces the message from Figure 9, that the reported results are extreme outliers, relative to reasonable alternatives.

Third, some choices made in these papers are clearly *not* preferred. Examples include losing sample size, relative to the actual experiment sample due to criteria that are not intuitive or seem unnecessary such as sample selection criteria that introduce survivorship bias, using covariates that are potential outcomes of the treatment, using two-way clustered standard errors for a relatively short panel; etc.

VII. Conclusion

In this paper, we provide evidence for specification search (p-hacking) using the setting of Reg SHO. Reg SHO is ideally suited to test for p-hacking as it was a randomized trial conducted on a large sample of firms with no evidence of a causal channel for the reported indirect treatment effects. That over 80 papers report a wide range of indirect effects in the absence of a causal channel raises the likelihood that the reported results could be false positives. Our re-examination of key findings in 10 papers reporting indirect effects finds no meaningful support for the core results in any of them. Results with both our pre-specified design and with the best-match

specifications are substantially different from the reported results. The large gap between best-match and reported results suggests that the authors of these papers made unreported specification choices, which produced coefficients that are substantially larger than what we obtain with our or the best-match specifications and/or smaller s.e.'s that are smaller than those obtained with our specification or the best-match specification.

Our analysis has important implications for the credibility of the profession. Many natural experiments result from regulatory change and/or involve policy implications. Several studies that have examined the Reg SHO experiment, draw policy implications (directly or indirectly) from their results. Given the fragility of these results and evidence suggesting that many of these results are false positives, it is particularly important that papers inferring policy implications receive closer scrutiny. For example, journals can demand posting of statistical codes, of the underlying data where possible, and greater transparency in the reporting of sample selection criteria and research design choices. Journals' openness to publishing carefully conducted replication studies can also help (Bartos et al., 2025).

Our assessment of whether ours, the best match, and the reported results could all come from the same underlying specification pool provides a generalizable approach for assessing specification search, and distinguishing specification search from file-drawer/publication bias. The key is to develop a set of reasonable but unreported specifications, that one can compare to the reported results. The randomized nature of the SEC's short sale experiment allowed us an opportunity to do just that.

References

Adams, John, Darren Hayunga, Sattar Mansi., David Reeb, and Vincenzo Verardi (2019), Identifying and Treating Outliers in Finance, *Financial Management* 48, 345-384.

Alexander, Gordon J. and Mark A. Peterson (2008), The Effect of Price Tests on Trader Behavior and Market Quality: An Analysis of Regulation SHO, *Journal of Financial Markets* 11, 84-111.

Angrist, Joshua D. and Jörn-Steffen Pischke (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton university press.

Atanasov, Vladimir, and Bernard Black (2016), Shock-Based Causal Inference in Corporate Finance and Accounting Research, *Critical Finance Review* 5, 207-304.

Bai, John (Jianqui), Eunju Lee, and Chi Zhang (2020), Capital Market Frictions and Human Capital Investment: Evidence from Workplace Safety Around Regulation SHO, *Financial Review* 55, 339-360.

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H.C. Sant'Anna (2025), Difference-in-Differences Designs: A Practitioner's Guide.

Bartos, Frantisek, Alexandra Sarafoglou, Balazs Aczel, Suzanne Hoogeveen, Christopher D. Chambers, et al. (2025), Introducing the Journal of Robustness Reports, at doi:10.21468/JRobustRep.0-Editorial.

Biddle, Gary C., Gilles Hilary, and Rodrigo S. Verdi (2009), How does financial reporting quality relate to investment efficiency?, *Journal of Accounting and Economics* 48(2–3), 112–131.

Black, Bernard, Hemang Desai, Kate Litvak, Woongsun Yoo, and Jeff Jiewei Yu (2019), Pre-Analysis Plan for the Reg SHO Reanalysis Project, working paper, at http://ssrn.com/abstract=3415529.

Black, Bernard, Hemang Desai, Kate Litvak, Woongsun Yoo, and Jeff Jiewei Yu (2024), The SEC's Short-Sale Experiment: Evidence on Causal Channels and on the Importance of Specification Choice in Randomized and Natural Experiments, *Management Science* 70(8), 5131-5456.

Black, Bernard, Alex Hollingsworth, Leticia Nunes and Kosali Simon. (2022) Simulated Power Analysis for Observational Studies: Application to the Affordable Care Act Medicaid Expansion, *Journal of Public Economics* 213, 104713.

Brodeur, Abel, Mathias Le, Marc Sangnier, and Yanos Zylberberg (2016), *American Economic Journal: Applied Economics* 8(1), 1-32.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2020), Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics, *American Economic Review* 110(11), 3634-3660.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2022), Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Reply, *American Economic Review,* 3137-3139.

Bui, Dien Giau, Iftekar Hasan, Chih-Yung Lin, and Hong Thoa Nguyen (2023), Short-Selling Threats and Bank Risk-Taking: Evidence from the Financial Crisis. *Journal of Banking and Finance* 150, 106834.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008), Bootstrap-Based Improvements for Inference with Clustered Errors, *Review of Economics and Statistics* 90, 414-427.

Catan, Emiliano M., and Michael Klausner (2017), Declassification and Firm Value: Hae Shareholders and Boards Really Destroyed Billions in Value?, working paper, at http://ssrn.com/abstract=2994559.

Chen, Hang, Yushu Zhu, and Liang Chang (2019), Short-Sale Constraints and Corporate Payout Policy, *Accounting and Finance* 59, 2273-2305.

Chen, Wei, Paul Hribar, and Samuel Melessa (2018), Incorrect inferences when using residuals as dependent variables, *Journal of Accounting Research* 56(3), 751–796. https://doi.org/10.1111/1475-679X.12195.

Chen, Zhihong, Siwen Fu, and Ke Wang (2023), Short-Sale Constraints and Firm Investment Efficiency: Evidence from a Natural Experiment, *Journal of Accounting and Public Policy* 42, 107149.

Cunningham, Scott (2025), Causal Inference: The Mixtape (2d ed., forthcoming).

De Angelis, David, Gustavo Grullon, and Sebastien Michenaud (2017), The Effects of Short-Selling Threats on Incentive Contracts: Evidence from an Experiment, *Review of Financial Studies* 30, 1627-1659.

Dechow, Patricia M., Weili Ge, Chad R. Larson, and Richard G. Sloan (2011), Predicting Material Accounting Misstatements, *Contemporary Accounting Research* 28, 17-82.

Desai, Hemang, Srinivasan Krishnamurthy and Kumar Venkataraman, Do Short Sellers Target Firms with Poor Earnings Quality: Evidence from Earnings Restatements, *Review of Accounting Studies* 11, 71-90.

Diether, Karl, Kuan-Hui Lee, and Ingrid Werner (2009), It's SHO Time! Short-Sale Price-Tests and Market Quality, *Journal of Finance* 64, 37-73.

Dreber, Anna, and Magnus Johannesson (2024), A framework for evaluating reproducibility and replicability in economics, *Economic Inquiry*, 1-19.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wuthrich (2022), Detecting p-Hacking, *Econometrica* 90(2), 887-906.

Fang, Vivien W., Allen Huang, and Jonathan Karpoff (2016), Short Selling and Earnings Management: A Controlled Experiment, *Journal of Finance* 71, 1251-1293.

Fang, Vivien W., Allen Huang, and Jonathan Karpoff (2019), Reply to "The Reg SHO Reanalysis Project: Reconsidering Fang, Huang and Karpoff (2016) on Reg SHO and Earnings Management" by Black et al. (2010), at http://ssrn.com/abstract=3507033.

Gelman, Andrew, and Eric Loken (2014), The Statistical Crisis in Science, *American Scientist* 102, 460-465.

Gerber, A.S., and N. Malhotra (2008), Publication Bias in Empirical Sociological Research, *Sociological Methods and Research* 37(1), 3-30.

Gong, Rong (2020), Short Selling Threat and Corporate Financing Decisions, *Journal of Banking and Finance* 118, 105853.

Gow, Ian D. (2023), The Elephant in the Room: p-Hacking and Accounting Research, working paper, at http://ssrn.com/abstract=4460192.

Gruber, Jonathan (1984), The Incidence of Mandated Maternity Benefits, *American Economic Review* 84, 622-641.

Grullon, Gustavo, Sebastien Michenaud, and James Weston (2015), The Real Effects of Short-Selling Constraints, *Review of Financial Studies* 28, 1737-1767.

Hail, Luzi, Mark Lang, and Christian Leuz (2020), Reproducibility in Accounting Research: Views of the Research Community, *Journal of Accounting Research* 58, 519-543.

Harvey, Campbell R. (2014), Reflections on Editing the Journal of Finance, in M. Szenberg and L. Ramrattan eds., *Secrets of Economics Editors* 67-82 (MIT Press).

Harvey, Campbell R., Yan Liu, and Heqing Zhu (2016), . . . and the Cross-Section of Expected Returns, *Review of Financial Studies* 29, 5-68

Harvey, Campbell R. (2017), The Scientific Outlook in Financial Economics, *Journal of Finance* 72, 1399-1440.

Harvey, Campbell R. (2019), Editorial: Replication in Financial Economics, *Critical Finance Review* 8, 1-9.

He, Jie (Jack), and Xuan Tian (2016), Do Short Sellers Exacerbate or Mitigate Managerial Myopia? Evidence from Patenting Activities, working paper, at http://ssrn.com/abstract=2380352.

Healy, Paul M. (1985), The Effect of Bonus Schemes on Accounting Decisions, *Journal of Accounting and Economics* 7, 85-107.

Heath, Davidson, Matthew C. Ringgenberg, Mehrdad Samadi, and Ingrid M. Werner (2023), Reusing Natural Experiments, *Journal of Finance* 78, 2329-2364.

Hope, Ole-Kristian, Danqi Hu, and Wuyang Zhao (2017), Third-Party Consequences of Short-Selling Threats: The Case of Auditor Behavior, *Journal of Accounting and Economics* 63, 479-498.

Hribar, Paul, and D. Craig Nichols (2007), The Use of Unsigned Earnings Quality Measures in Tests of Earnings Management, *Journal of Accounting Research* 45, 1017-1053.

Imai, Taisuke, Severine Toussaert, Aurelien Baillon, Anna Dreber, Seda Erfac, et al. (2025), Pre-Registration and Pre-Analysis Plans in Experimental Economics, working paper, at http://sssn.com/abstract=5219070.

Imbens, Guido W. and Donald B. Rubin (2015), *Causal Inference in Statistics, Social, and Biomedical* Sciences.

Jackson, Andrew B. (2018), Discretionary Accruals: Earnings Management or Not? *Abacus* 54, 136-153.

Karpoff, Jonathan M., and Xiaoxia Lou, Short Sellers and Financial Management, *Journal of Finance* 65, 1879-1913.

Kezdi, Gabor (2004), Robust Standard Error Estimation in Fixed-Effects Panel Models, *Hungarian Statistical Review* 9, 95-116.

Kim, E. Han, Yao Lu, and Zhang Peng (2022), Monitoring from Capital Market and Corporate Tax Avoidance: Evidence from Short Selling Pilot Program, working paper, at http://ssrn.com/abstract=3564782.

Kothari, Sagar P., Andrew J. Leone, and Charles E. Wasley (2005), Performance Matched Discretionary Accrual Measures, *Journal of Accounting and Economics* 39, 163-197.

Kranz, Sebastian, and Peter Putz (2022), Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment, *American Economic Review* 112(9), 3124-3136.

Larson, Chad R., Richard Sloan, and Jenny Zha Giedt (2018), Defining, Measuring, and Modeling Accruals: A Guide for Researchers *Review of Accounting Studies* 23, 827-871.

Leone, Andrew J., Miguel Minutti-Mexa, and Charles E. Wasley (2019), Influential Observations and Inference in Accounting Research, *The Accounting Review* 94(6), 337-364.

Lin, Tse-Chun, Qi Liu, and Bo Sun (2019), Contractual Managerial Incentives with Stock Price Feedback, *American Economic Review* 109(7), 2446-2468.

Litvak, Kate, Bernard Black, and Woongsun Yoo (2020), The SEC's Short-Sale Experiment: What Can and Cannot Be Learned, working paper, at http://ssrn.com/abstract=2647418.

Massa, Massimo, Bohui Zhang, and Hong Zhang (2015). The Invisible Hand of Short Selling: Does Short Selling Discipline Earnings Management? *Review of Financial Studies* 28, 1701-1736.

Menkveld, Albert, et al. (2024), Non-Standard Errors, *Journal of Finance* 79(3), 2339-2390.

Mitton, Todd (2022), Methodological Variation in Empirical Corporate Finance, *Review of Financial Studies* 35, 527-575.

Ohlson, James (2022), Researchers' Data Analysis Choices: An Excess of False Positives?, *Review of Accounting Studies* 27, 649-667.

Perignon, Christophe, Olivier Akmansoy, Christophe Hurlin, Anna Dreber, et al. (2004), Computational Reproducibility in Finance: Evidence from 1,000 Tests, *Review of Financial Studies* 37 (11), 3558-3593.

Richardson, Scott A. (2003), Earnings quality and short sellers. *Accounting Horizons* 17 (Supp.), 49-61.

Romano, Joseph P., and Michael Wolf (2005), Stepwise multiple testing as formalized data snooping, *Econometrica*, 73(4), 1237-1282.

Romano, Joseph P., and Michael Wolf (2016), Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics and Probability Letters*, 113, 38-40.

Rosenbaum, Paul R. (2020), *Design of Observational Studies* (2d ed.).

SEC, Office of Economic Analysis (2007), Economic Analysis of the Short Sale Price Restrictions under the Regulation SHO Pilot, at https://www.sec.gov/news/studies/2007/regshopilot020607.pdf.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson (2020), Specification Curve Analysis, *Nature Human Behavior* 4, 1208-1214.

Sloan, Richard G. (1996), Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings? *The Accounting Review* 71, 289-315.

Tsai, Hsin-Jun Stephie, Yuliang Wu, and Bin Xu (2021), Does capital market drive corporate investment efficiency? Evidence from equity lending supply, *Journal of Corporate Finance* 69, 102042.

Wang, Zexi (2018), Short sellers, institutional investors, and corporate cash holdings, working paper, at https://ssrn.com/abstract=2410239.

Welch, Ivo (2019), Reproducing, Extending, Updating, Replicating, Reexamining, and Reconciling, *Critical Finance Review* 8, 301-304.

Table 1. DiD Regression Results with Our Specification

Table summarizes results with our specification. S.e.'s with firm clusters in parentheses. Red = opposite sign from predicted. *, **, *** indicates significance at the 10%, 5%, and 1% level, respectively. Significant results, at 5% level or better, in boldface. Table omits, for space reasons, one KLP outcome, $NetR3_{truncated}$ (coeff. = -0.0003 s.e. = 0.005).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. FHK Accruals	Operating	Total	AA	PMDA	Operating	Total	AA	PMDA
Panel	Unbal.	Unbal.	Unbal.	Unbal.	Balanced	Balanced	Balanced	Balanced
Pilot*During	-0.0024	-0.0000	-0.0020	0.0029	-0.0037	0.0035	-0.0030	0.0075
s.e.	(0.0035)	(0.0076)	(0.0050)	(0.0080)	(0.0036)	(0.0079)	(0.0054)	(0.0087)
B1. FHK HF-Score	Unbal.	Unbal.	Unbal.		Balanced	Balanced	Balanced	
Outcome	HF-1	HF-2	HF-3		HF-1	HF-2	HF-3	
Pilot*During	0.0144	0.0424	-0.0255		-0.0873	-0.0354	0.1566	
s.e.	(0.1282)	(0.1281)	(0.1386)		(0.1381)	(0.1436)	(0.1650)	
B2. FHK F-score: Outcome	F-1	F-2	F-3		F-1	F-2	F-3	
Pilot*During	0.0096	0.0106	0.0047		0.0020	0.0071	0.0240	
s.e.	(0.0167]	(0.0185]	(0.0251]		(0.0182]	(0.0205]	(0.0235]	
C. HHZ Audit Fees	No covars.	Limited covars.	Full covars.					
Pilot*During	-0.0046	-0.0086	-0.0132					
s.e.	(0.0192)	(0.0179)	(0.0176)					
D. GMW Investment	Capex	R&D/Sales	(Capex + R&D)	Ttl. Invest	Asset growth	*ln*(assets)	Equity Issues	Debt Issues
Pilot*During	-0.496*Mae	4.443	-0.743	-0.5139	-2.423	-0.0006	-0.723	-1.616
s.e.	(0.2952)	(9.6587)	(0.6880)	(1.0488)	(2.6626)	(0.0300)	(1.5717)	(1.4052)
E. LLS Investment, WPS	Capex	R&D/Sales	R&D/Assets	(Capex+R&D)	Ttl. Invest.	WPS		
Pilot*During(*Q exc. WPS)	**-0.0046*** **	0.0338	-0.0007	-0.0061*	-0.5690	-0.0748		
s.e.	**(0.0017)**	(0.0300)	(0.0025)	(0.0033)	(0.4720)	(0.0514)		
F. CFW Underinvestment	Ttl. Invest.	Capex/assets	R&D/assets	(Capex+R&D)				
Pilot*During	**2.2736****	0.2955	0.3882	0.8462*				
s.e.	(**1.0706**)	(0.3204)	(0.2377)	(0.5122)				
Overinvest. F-stat (p value)	3.75 (p=0.053)*	2.11 (p = 0.147)	0.30 (p = 0.586)	1.92 (p = 0.167)				
G. TWX Underinvestment								
Pilot*During	**1.5791****	0.2383	0.2641	0.6276				
s.e.	(**0.8041**)	(0.2413)	(0.1869)	(0.3844)				
Overinvest. F-stat (p value)	2.61 (p = 0.107)	**4.43** (p=0.035)**	0.33 (p = 0.564)	2.92 (p=0.088)*				
H. KLP Tax Sheltering	Book-Tax Diff	$BTD_{alt.}$	NetR3	BTD_{adj}	$NetR3_{adj}$	NCETR3	Perm	Residuals
Pilot*During	-0.0019	-0.0024	-0.0165	0.0006	-0.0123	-0.0147	-0.0038	-0.0022
s.e.	(0.005)	(0.005)	(0.016)	(0.005)	(0.015)	(0.017)	(0.005)	(0.005)
I. CZC Dividends	Dividends	Repurchases	Payout	Div./Assets	Rep./Assets	Pay./Assets		
Pilot*During	0.2840	-1.5473	-1.2276	0.0007	-0.0042	-0.0040		
s.e.	(0.3302)	(1.3455)	(1.4790)	(0.0009)	(0.0048)	(0.0050)		
J. Gong Leverage	Book leverage	Mkt. leverage						
Pilot*During	-0.0067	-0.0082						
s.e.	(0.0064)	(0.0099)						
K. BHLN Bank Risk	Book Lev. I	Mkt Lev I	Book Lev. II	Mkt Lev II	NPL/Loans	NPL/Equity	z-score	Exp. Default
Pilot*Crisis	-0.0017	-0.0023	-0.3246	-2.0329	-0.0023	-0.0398	0.0754	0.0037
s.e.	(0.0028)	(0.0058)	(0.4109)	(2.6401)	(0.0027)	(0.0362)	(0.44)	(0.14)

Table 2. Results with Best-Match Specifications

All panels. Table summarizes results with best-match specifications. See Internet Appendix for regression details. S.e.'s with firm clusters in parentheses. For papers that use 2-way clustering on firm and time, s.e.'s using 2-way clusters in brackets. Red = opposite sign from predicted. *, **, *** indicates significance at the 10%, 5%, and 1% level, respectively. Significant results, at 5% level or better, in boldface. Table omits one KLP outcome, $NetR3_{truncated}$ (coeff. = -0.0129 s.e. = 0.010 w firm-clusters; 0.009 w. firm-year clusters). For CFW we use adjusted best-match specifications.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. FHK Accruals	Operating	Total	AA	PMDA	Operating	Total	AA	PMDA
Panel	Unbal.	Unbal.	Unbal.	Unbal.	Balanced	Balanced	Balanced	Balanced
Pilot*During	-0.0047	-0.0033	-0.0030	0.0038	-0.0068	0.0027	-0.0073	-0.0022
s.e. (firm clusters)	(0.0039)	(0.0077)	(0.0054)	(0.0085)	(0.0042)	(0.0082)	(0.0062)	(0.0096)
s.e. (2-way clusters)	[0.0034]	[0.0064]	[0.0045]	[.]	[0.0048]	[0.0094]	[0.0064]	[0.0030]
B1. FHK HF-Score	Unbal.	Unbal.	Unbal.		Balanced	Balanced	Balanced	
Outcome	HF-1	HF-2	HF-3		HF-1	HF-2	HF-3	
Pilot*During	-0.087	-0.162	0.011		-0.207	-0.194	0.052	
s.e. (firm clusters)	(0.1376)	(0.1349)	(0.1405)		(0.1652)	(0.1660)	(0.1695)	
s.e. (2-way clusters)	[0.0972]	**[0.0674]****	[0.1056]		**[0.0399]*****	**[0.0938]****	[0.1198]	
B2. FHK F-Score Outcome	F-1	F-2	F-3		F-1	F-2	F-3	
Pilot*During	-0.005	-0.001	-0.001		0.007	0.011	0.022	
s.e. (firm clusters)	(0.0199)	(0.0215)	(0.0254)		(0.0212)	(0.0230)	(0.0269)	
s.e. (2-way clusters)	[0.0145]	[0.0160]	[0.0177]		[0.0165]	[0.0192]	[0.0245]	
C. HHZ Audit Fees								
Covariates	None	Limited	Full					
Pilot*During	0.0151	0.0135	0.0116					
s.e.	(0.0244)	(0.0218)	(0.0215)					
D. GMW Investment	Capex	R&D/Sales	(Capex + R&D)	Ttl. Invest	Asset growth	*ln*(assets)	Equity Issues	Debt Issues
Pilot*During	**-0.644****	-0.270	-0.644	0.0189	-0.532	-0.0235	0.505	-1.56
s.e.	**(0.3126)**	(9.0000)	(0.6440)	(0.9450)	(2.6600)	(0.0309)	(1.3649)	(1.3929)
E. LLS Investment, WPS	Capex	R&D/Sales	R&D/Assets	(Capex+R&D)	Ttl. Invest.	WPS		
Pilot*During(*Q exc. WPS)	-0.0006	0.0049*	0.0024	0.0022	0.3852	**-0.1292****		
s.e.	(0.0016)	(0.0027)	(0.0026)	(0.0029)	(0.4674)	**(0.0627)**		
F. CFW Underinvestment	Ttl. Invest.	Capex/assets	R&D/assets	(Capex+R&D)				
Pilot*During	0.8357	-0.5160	0.9808*	-0.0254				
s.e. (firm clusters)	(2.7816)	(0.4752)	(0.5281)	(0.8624)				
s.e. (2-way clusters)	[3.0951]	[0.4914]	[0.4087]	[0.8467]				
Overinvest. F-stat (p value)	1.85 (p = 0.174)	1.52 (p = 0.218)	0.70 (p = 0.403)	0.81 (p = 0.368)				
(2-way clustered s.e.'s)	1.93 (p = 0.207)	1.00 (p = 0.350)	1.00 (p = 0.351)	0.85 (p = 0.387)				
G. TWX Underinvestment								
Pilot*During	1.2033	0.1173	0.1698	0.3947				
s.e.	(0.8182)	(0.2464)	(0.1937)	(0.3936)				
Overinvest. F-stat (p value)	1.46 (p = 0.228)	2.97 (p=0.085)*	0.10 (p = 0.749)	1.71 (p = 0.191)				

H. KLP Tax Sheltering	Book-Tax Diff	$BTD_{alt.}$	NetR3	BTD_{adj}	$NetR3_{adj}$	NCETR3	Perm	Residuals
Pilot*During	-0.0063	-0.0073	-0.0284	-0.0039	-0.0238	0.0087	-0.0061	-0.0061
s.e. (firm clusters)	(0.006)	(0.006)	(0.033)	(0.005)	(0.029)	(0.035)	(0.006)	(0.006)
s.e. (2-way clusters)	[0.006]	[0.008]	[0.023]	[0.005]	[0.022]	[0.041]	[0.008]	[0.006]
I. CZC Dividends	Dividends	Repurchases	Payout	Div./Assets	Rep./Assets	Pay./Assets		
Pilot*During	0.4507	-1.8682	-1.4559	0.0015	-0.0017	-0.0006		
s.e. (firm clusters)	(0.4060)	(1.7625)	(1.9412)	(0.0012)	(0.0057)	(0.006)		
s.e. (2-way clusters)	(0.3852)	(1.7460)	(1.8665)	(0.0011)	(0.0044)	(0.0043)		
J. Gong Leverage	Book leverage	Mkt. leverage						
Pilot*During	0.0009	-0.0378						
s.e. (firm clusters)	(0.0069)	(0.0461)						
s.e. (2-way clusters)	[0.0069]	[0.0473]						
K. BHLN Bank Risk	Book Lev. I	Mkt Lev I	Book Lev. II	Mkt Lev II	NPL/Loans	NPL/Equity	z-score	Exp. Default
Pilot*Crisis	-0.0004	-0.0031	n.m.	-1.7797	-0.0021	-0.0237	0.0780	-0.0029
s.e.	(0.0013)	(0.0041)		(1.4352)	(0.0018)	(0.0196)	(0.82)	(0.19)

Table 3. Reported Results

All panels. Table summarizes 37 reported results in the respective papers. S.e.'s with indicated clusters in parentheses. *, **, *** indicates significance at the 10%, 5%, and 1% level, respectively. Significant results, at 5% level or better, in boldface. Table omits one KLP outcome, $NetR3_{truncated}$ (coeff. = -0.010*; 2-way cluster s.e. = 0.005).

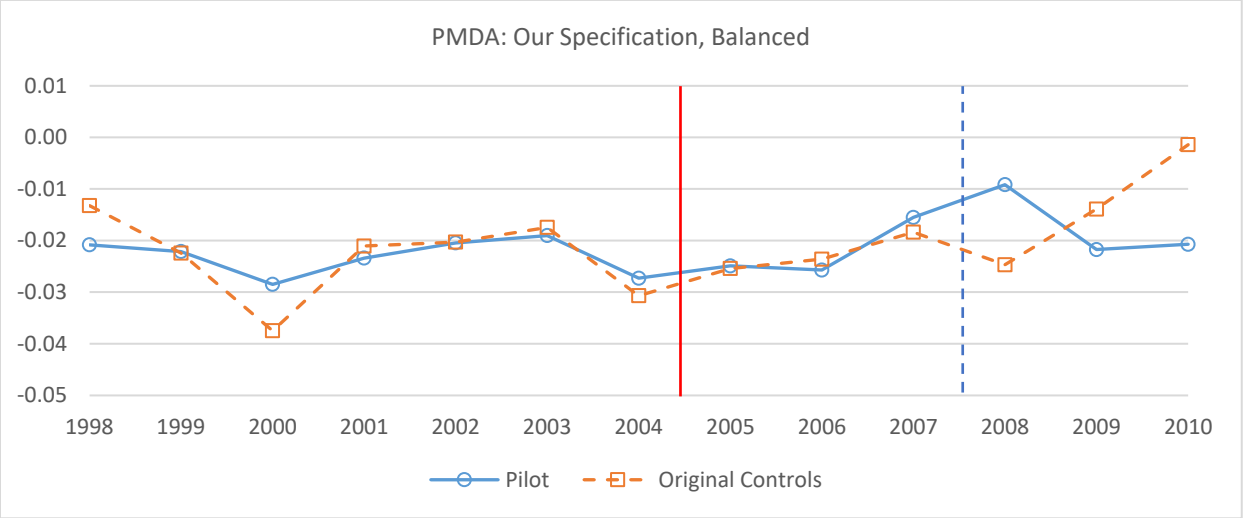
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. FHK Accruals								PMDA
Panel								Balanced
Pilot*During								**-0.010****
s.e. (2-way clusters)								**[0.004]**
B. FHK HF-Score					Balanced	Balanced	Balanced	
Outcome					HF-1	HF-2	HF-3	
Pilot*During					**-0.178****	**-0.189****	**-0.200****	
s.e. (2-way clusters)					**[0.080]**	**[0.079]**	**[0.080]**	
C. HHZ Audit Fees								
Covariates	None	Limited	Full					
Pilot*During	0.048	**0.0476**	**0.0465**					
s.e. (firm clusters)	(0.0440)	**(0.0239)****	**(0.0235)****					
D. GMW Investment	Capex		(Capex + R&D)		Asset growth		Equity Issues	Debt Issues
Pilot*During	**-0.97**		-1.05		**-6.12**		-1.61	-1.80
s.e. (firm clusters)	**(0.3368)*****		(0.5801)*		**(2.4000)****		(0.8846)*	(1.184)
E. LLS Investment, WPS			R&D/Assets	(Capex+R&D)		WPS		
Pilot*During(*Q exc. WPS)			**0.004*****	**0.004****		**-0.174****		
s.e. (firm clusters)			**[0.001]**	**[0.002]**		**(0.077)**		
F. CFW Underinvestment	Ttl. Invest.							
Pilot*During	**4.758*****							
s.e. (2-way clusters)	**[0.7298]**							
Overinvest. F-stat (p value)	**n.a. (p =0.006)**							
G. TWX Overinvestment								
Pilot*During	**2.482****							
s.e. (firm clusters)	**(1.089)**							
Overinvest. F-stat (p value)	**4.9 (p < 0.05)**							
H. KLP Tax Sheltering	Book-Tax Diff	.	NetR3	BTD_{adj}	$NetR3_{adj}$	NCETR3	Perm	Residuals
Pilot*During	**-0.011****	-	**-0.039****	-0.028*	**-0.010****	**-0.035****	-0.008*	**-0.012*****
s.e. (2-way clusters)	**[0.004]**	-	**[0.012]**	[0.013]	**[0.004]**	**[0.015]**	[0.004]	**[0.003]**
I. CZC Dividends	Dividends	Repurchases						
Pilot*During	**2.686*****	-0.529						
s.e. (2-way clusters)	**[0.7831]**	[1.9593]						
J. Gong Leverage	Book leverage	Mkt. leverage						
Pilot*During	**-0.007*****	**-0.005****						
s.e. (2-way clusters)	**[0.0023]**	**[0.0022]**						
K. BHLN Bank Risk			Book Lev. II	Mkt Lev II	NPL/Loans	NPL/Equity	z-score	Exp. default rate
Pilot*Crisis			**0.6904*****	**3.3889****	**0.0094*****	**0.0857*****	**-0.518*****	**0.140*****
s.e. (firm clusters)			**(0.3124)**	**(1.4995)**	**(0.0021)**	**(0.0237)**	**(0.160)**	**(0.035)**

Figure 1. Annual Means for Selected Outcomes: Our Specification

Figures show annual univariate means, separately for pilot and control firms, for indicated papers and outcomes, using our specification. **Panel A**. FHK HF-Score. **Panel B**. LLS (Capex + R&D)/Assets (coefficient on triple difference). **Panel C**. CFW Investment (coefficient on triple difference. **Panel D**. TWX Investment (coefficient on triple difference). **Panel E**. KLP Book-Tax Difference. **Panel F**. CZC Dividends (unscaled). **Panel G**. Gong leverage. **Panel H**. BHLN Non-performing loans/total loans. Solid and dashed vertical lines separate Pre, During, and Post periods. See BDLYY (2024) for similar graphs for FHK Accruals; HHZ Audit Fees; GMW Capex/assets, and LLS wealth-performance sensitivity. See Internet Appendix for similar graphs using best-match specifications, and graphs for additional outcomes.

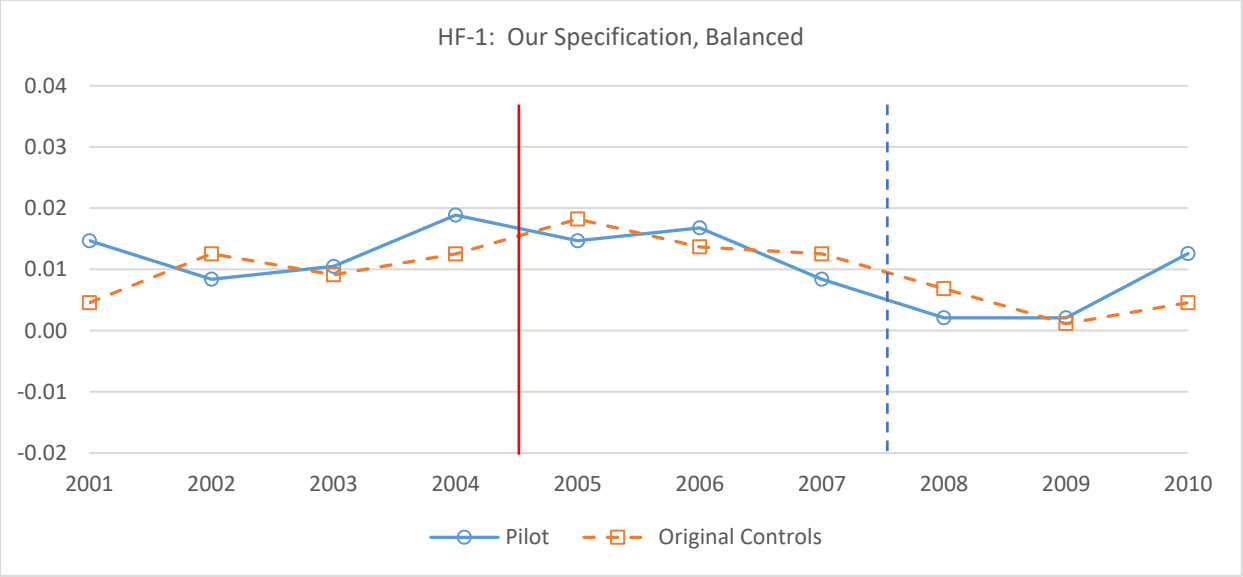
Panel A. FHK Accruals (PMDA, Balanced Panel)

Interpretation: No evidence for a treatment effect. Erratic treatment-vs-control gaps in Pre and Post periods.



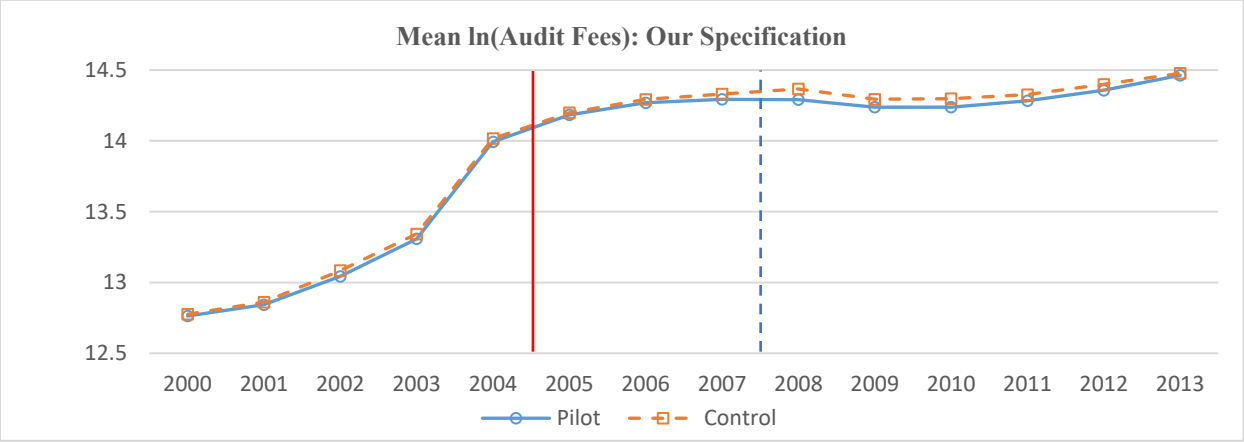
Panel B. FHK HF-Score (HF-1, Balanced Panel).

Interpretation: No evidence for a treatment effect. Erratic treatment-vs-control gaps in Pre and Post periods



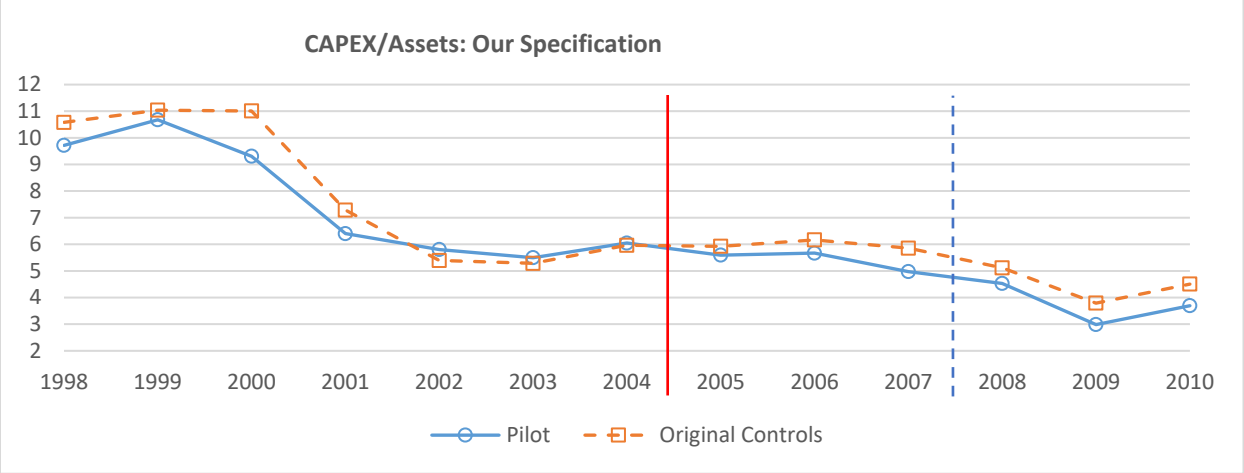
Panel C. HHZ Audit Fees

Interpretation: No evidence for a treatment effect. Larger treatment-vs-control gaps in Post periods.



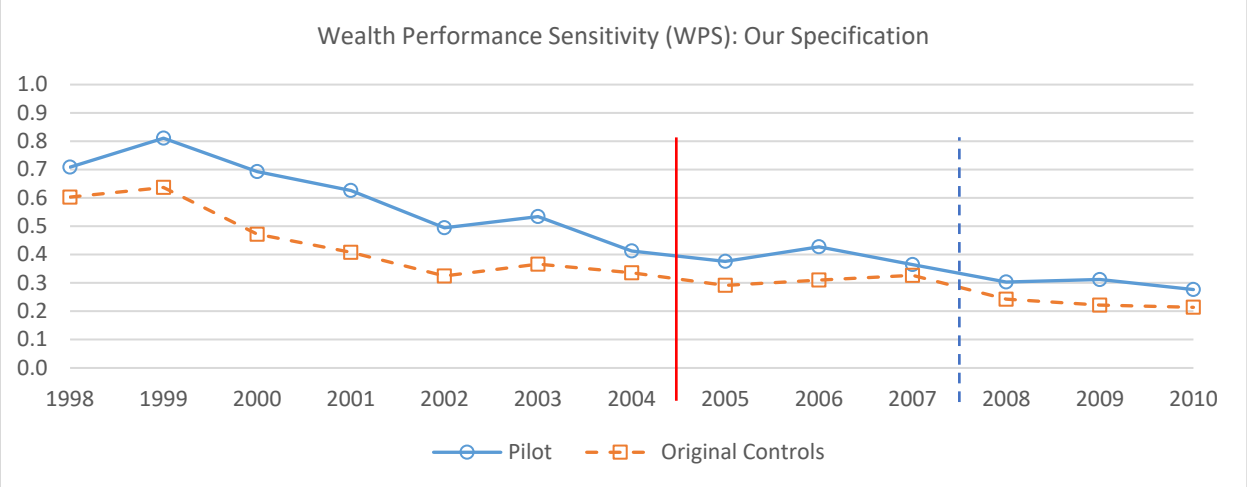
Panel D. GMW Capex/Assets

Interpretation: No evidence for a treatment effect. Similar or larger treatment-vs-control gaps in Pre and Post periods.



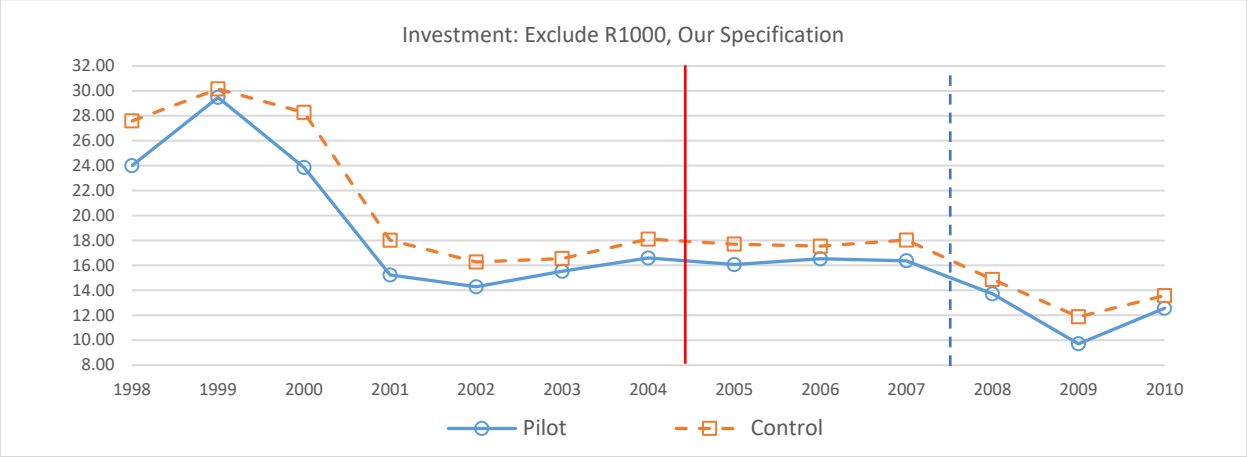
Panel E. LLS CEO Wealth Performance Sensitivity

Interpretation: No evidence for a treatment effect. Non-parallel trends in Pre period (2001-2004), lead to biased DiD estimates if 2004 is excluded from the Pre Period.



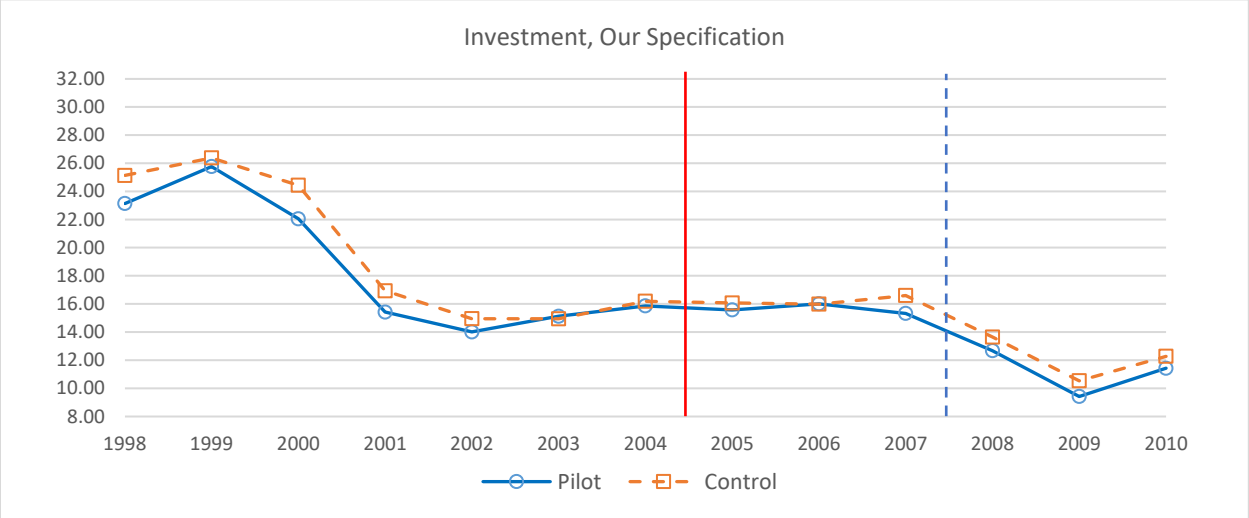
Panel F. CFW Under investment

Interpretation: No evidence for an overall effect on investment. Non-parallel trends in Pre period. No reversal in Post period.



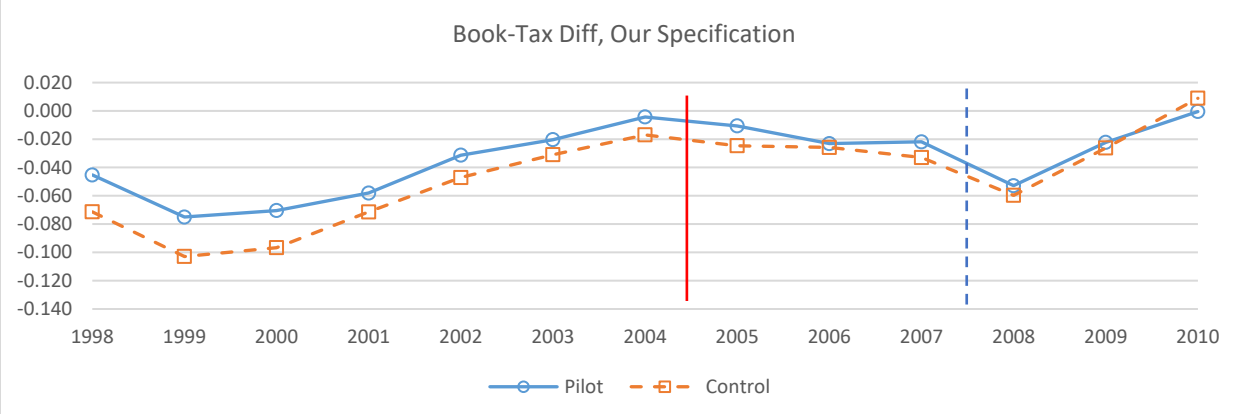
Panel G. TWX Underinvestment

Interpretation: Similar to CFW. No evidence for an overall effect on investment. Non-parallel trends in Pre period (2001-2004). No reversal in Post period.



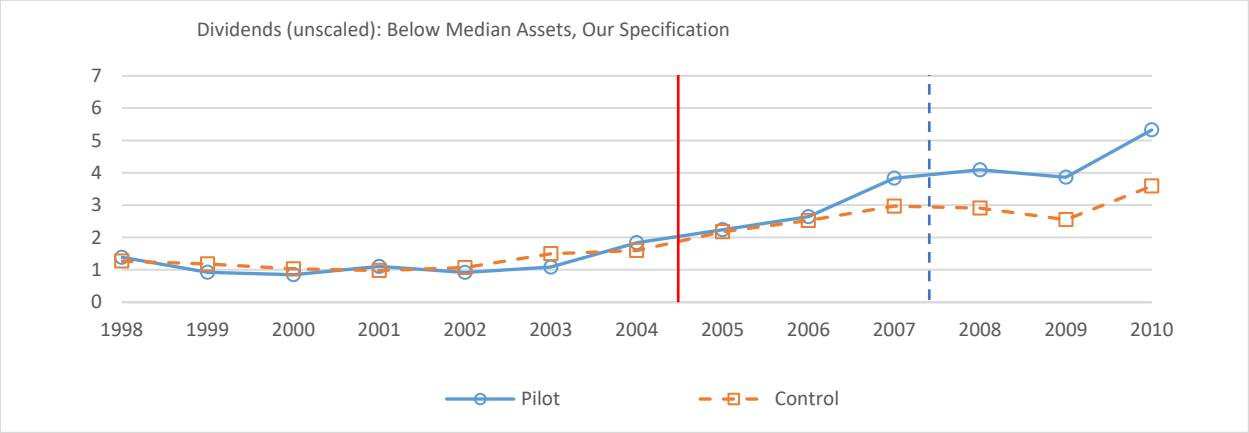
Panel H. KLP Book-Tax Difference

Interpretation: No evidence for a treatment effect. Non-parallel trends in Pre period.



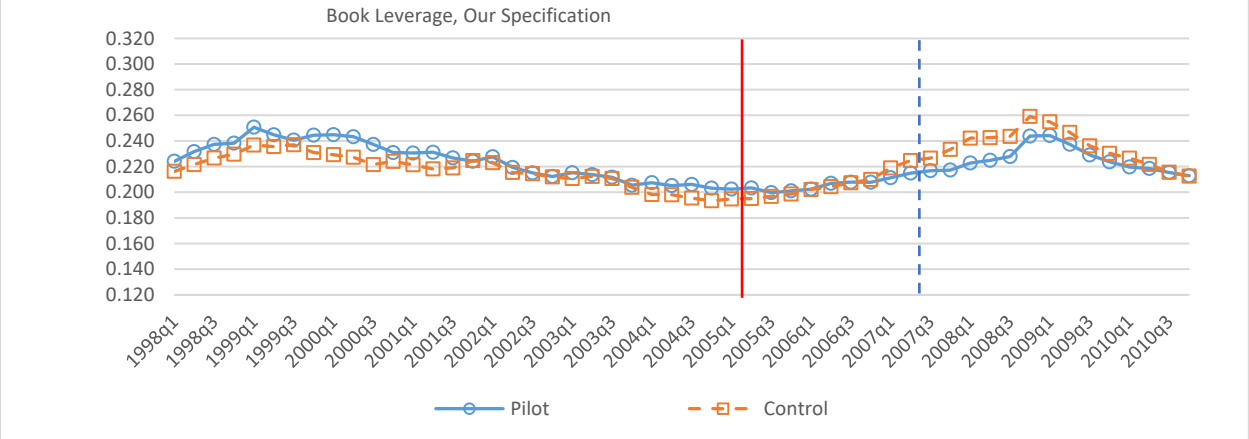
Panel I. CZC Dividends (unscaled)

Interpretation: Evidence for a treatment effect only in 2007. Larger pilot-vs-control gap in Post period.



Panel J. Gong (Book Leverage)

Interpretation: No evidence for a treatment effect. Larger treatment-vs-control gap in Post period.



Panel K. BHLN (Nonperforming Loans/Assets)

Interpretation: No evidence for a treatment effect.

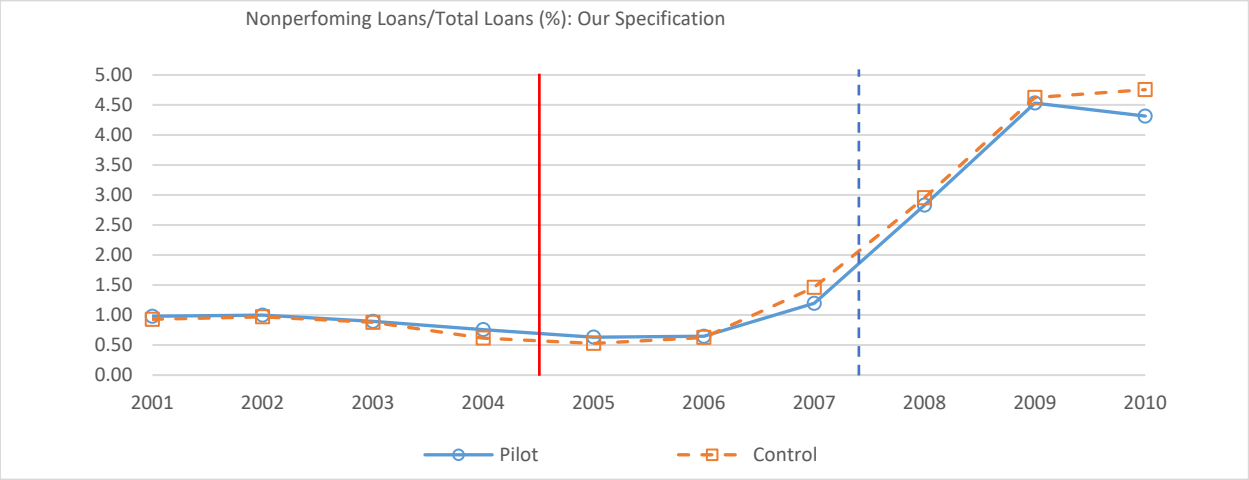


Figure 2. Statistical Significance Across Specifications: Our Specification, Best-Match Specifications and Reported Results

Figure reports *t*-statistics for our specification, best-match specifications (using one-way clustering) and reported results, for all 37 re-examined outcomes. *t*-statistics are reported as positive except that we report negative *t*-statistics for our specification and best-match specifications when the coefficient sign is opposite from the reported sign. For CFW overinvestment, we report a *t*-statistic with the same p-value as the reported p value. For TWX overinvestment, we assume $t = F^{0.5}$. For BHLN book leverage II, we report the coefficient without covariates. For CZC repurchase, reported sign is opposite from expected. Vertical lines at ± 1.96 indicate 5% significance.

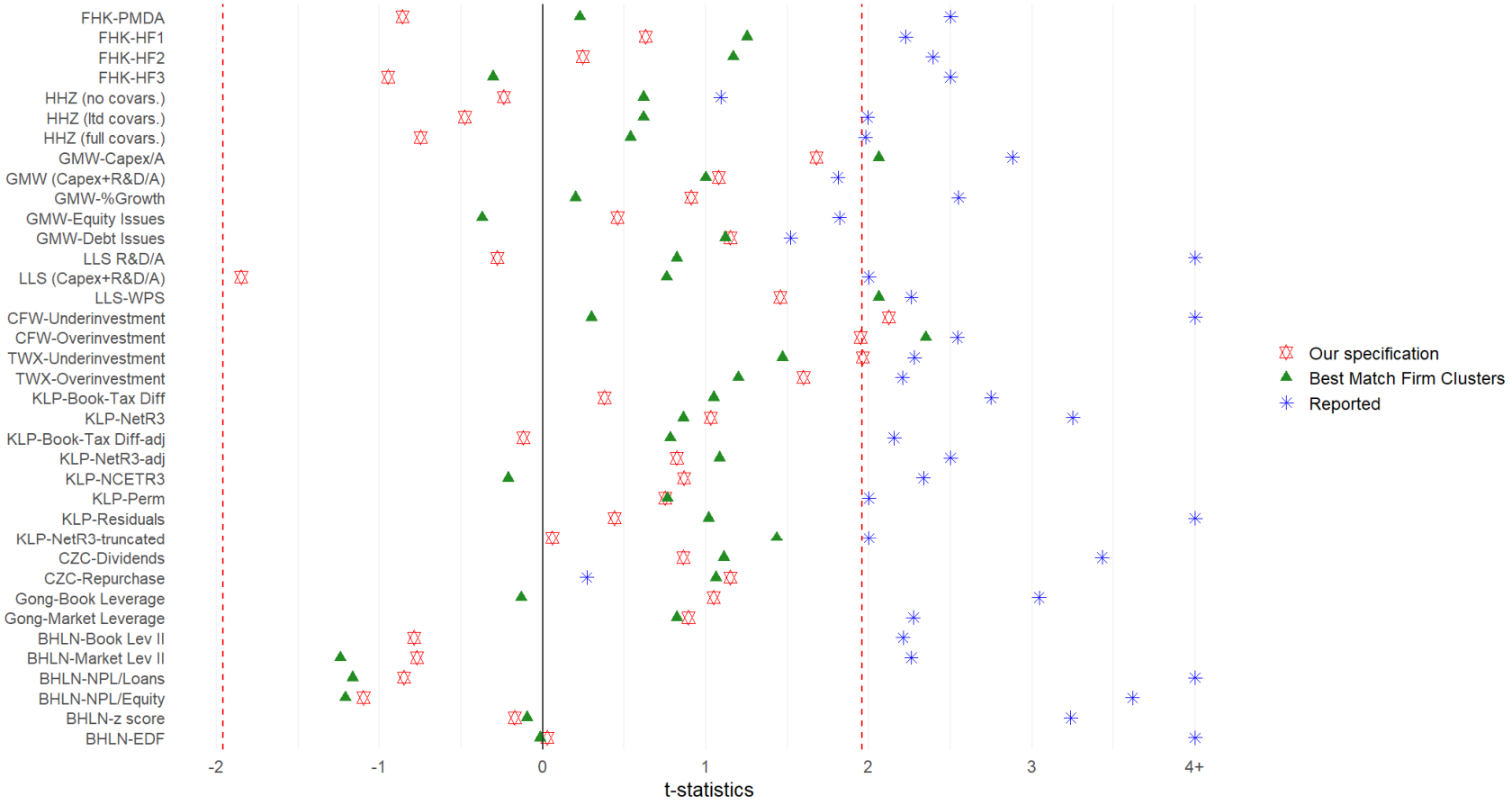


Figure 3. Statistical Significance for Additional Specifications: Our specification and Best-Match specifications

Figure reports *t*-statistics for our specification and best-match specifications (using one-way clustering), for 41 additional specifications, studied as robustness checks for the reported outcomes. *t*-statistics are reported as positive except that we report negative *t*-statistics when the coefficient sign is opposite from that expected based on the reported specifications. For BHLN book leverage II, we report the coefficient without covariates. Vertical lines at ± 1.96 indicate 5% significance.

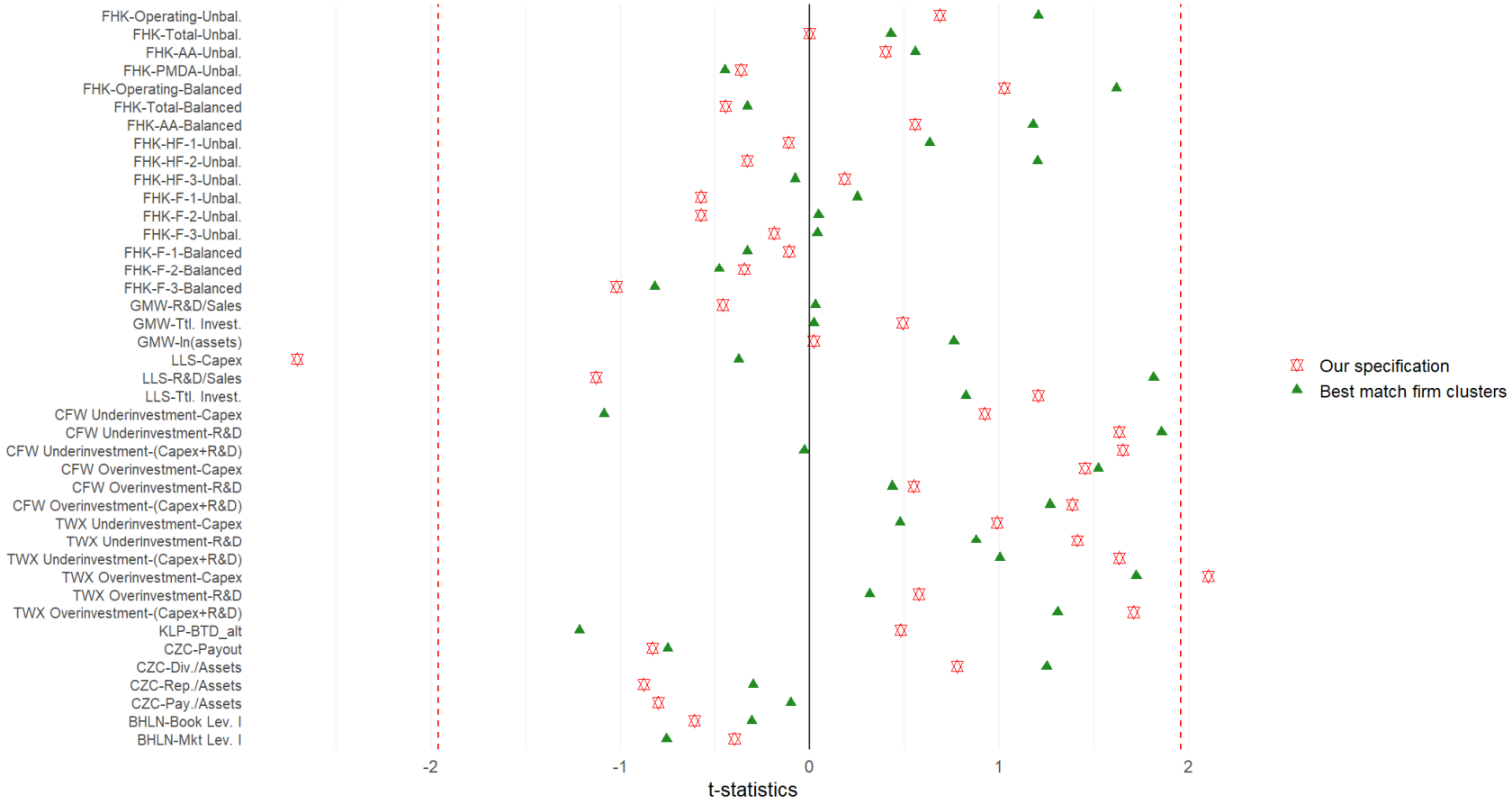


Figure 4. Statistical significance across specifications: Our specification, best match specifications and reported results

Figure provides histogram of number of outcomes with *t*-statistics in indicated ranges for our specification, best-match specifications (using one-way clustering) and reported results, for all specifications included in Figure 2 and Figure 3. Counts for best-match specification are shown below the "0" line, inverted for visual presentation. *t*-statistics are reported as positive except that we report negative *t*-statistics for our specification and best-match specifications when the coefficient sign is opposite from reported specifications. Dashed vertical lines at ± 1.645 and ± 1.96 indicate 10% and 5% significance**.**

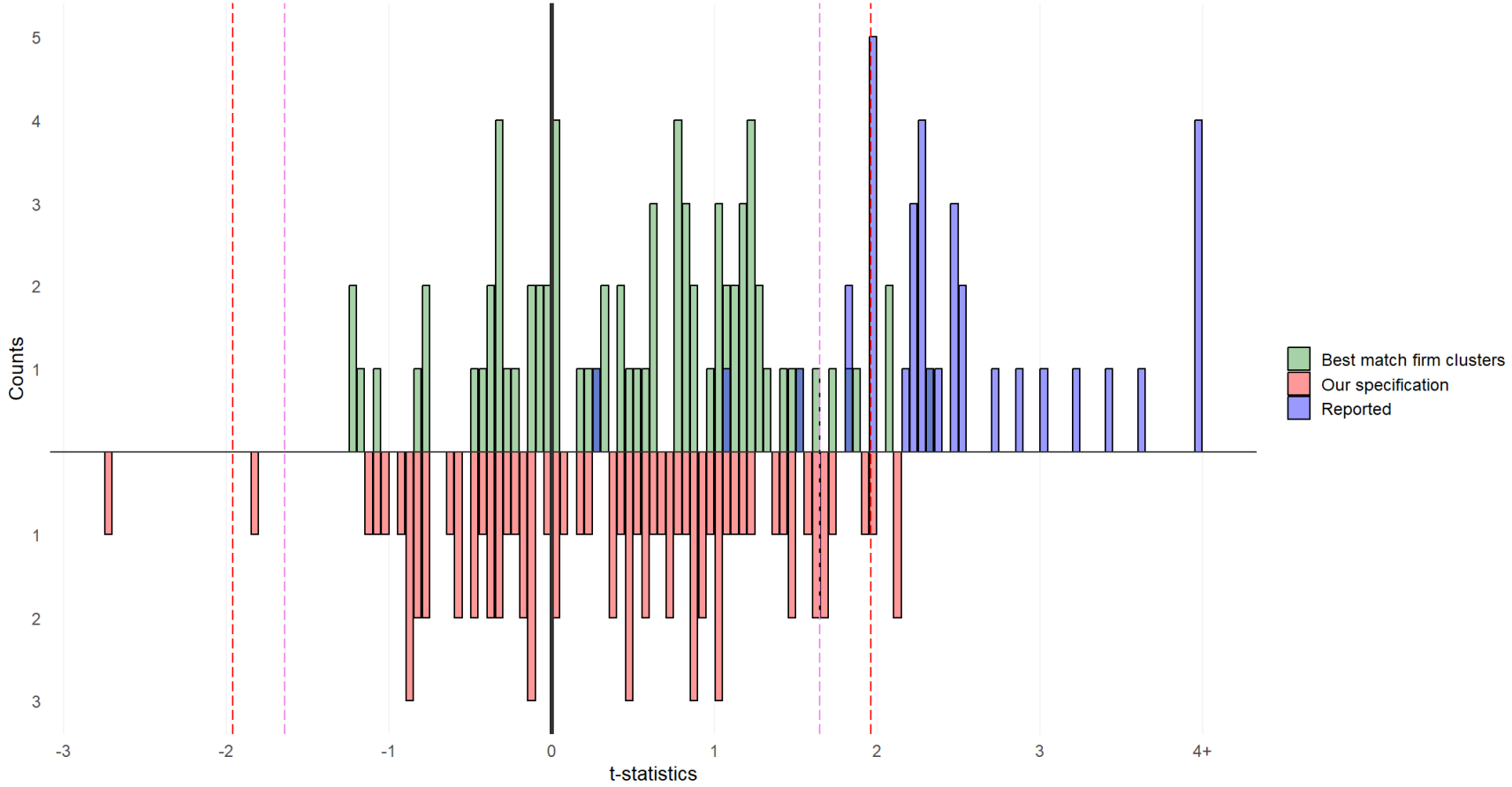


Figure 5. Coefficient Ratios Across Specifications

Figure shows, for indicated outcomes: (i) regression coefficients for our specifications, scaled to 1.00 if our coefficient has same sign as the reported coefficient, -1.00 if we find a sign that is opposite to reported sign; and (ii) coefficients for best-match specifications and reported results. Coefficient magnitudes are scaled to magnitudes for our specification. For CFW overinvestment and TWX overinvestment, we report the sum of the coefficients on Pilot*During and Pilot*During*Overtend. Scaled coefficients are winsorized at 6.0 and -2.0.

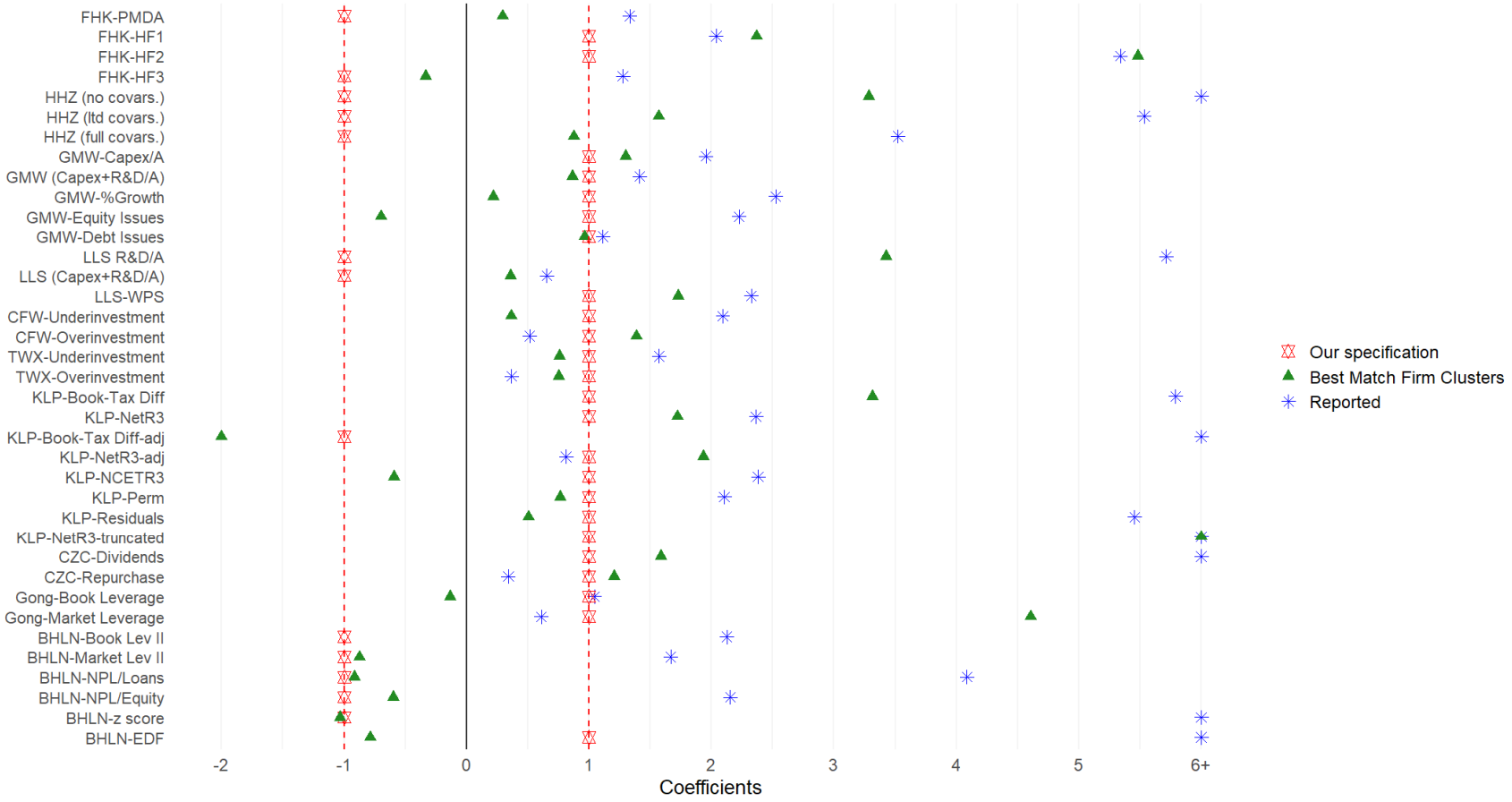


Figure 6. Standard Errors Across Specifications

Figure shows, for indicated reported outcomes, s.e.'s for our specifications, scaled to 1.00, and scaled s.e.'s for best-match specifications (using one-way clustering on firm) and reported results (clustering varies). Best-match and reported s.e.'s are top-winsorized at 2.0 (scaled). For reported results, two-way clustering is on firm and year for all studies except Gong (who uses quarterly data). For Gong, s.e.'s for our and best-match specifications are indistinguishable.

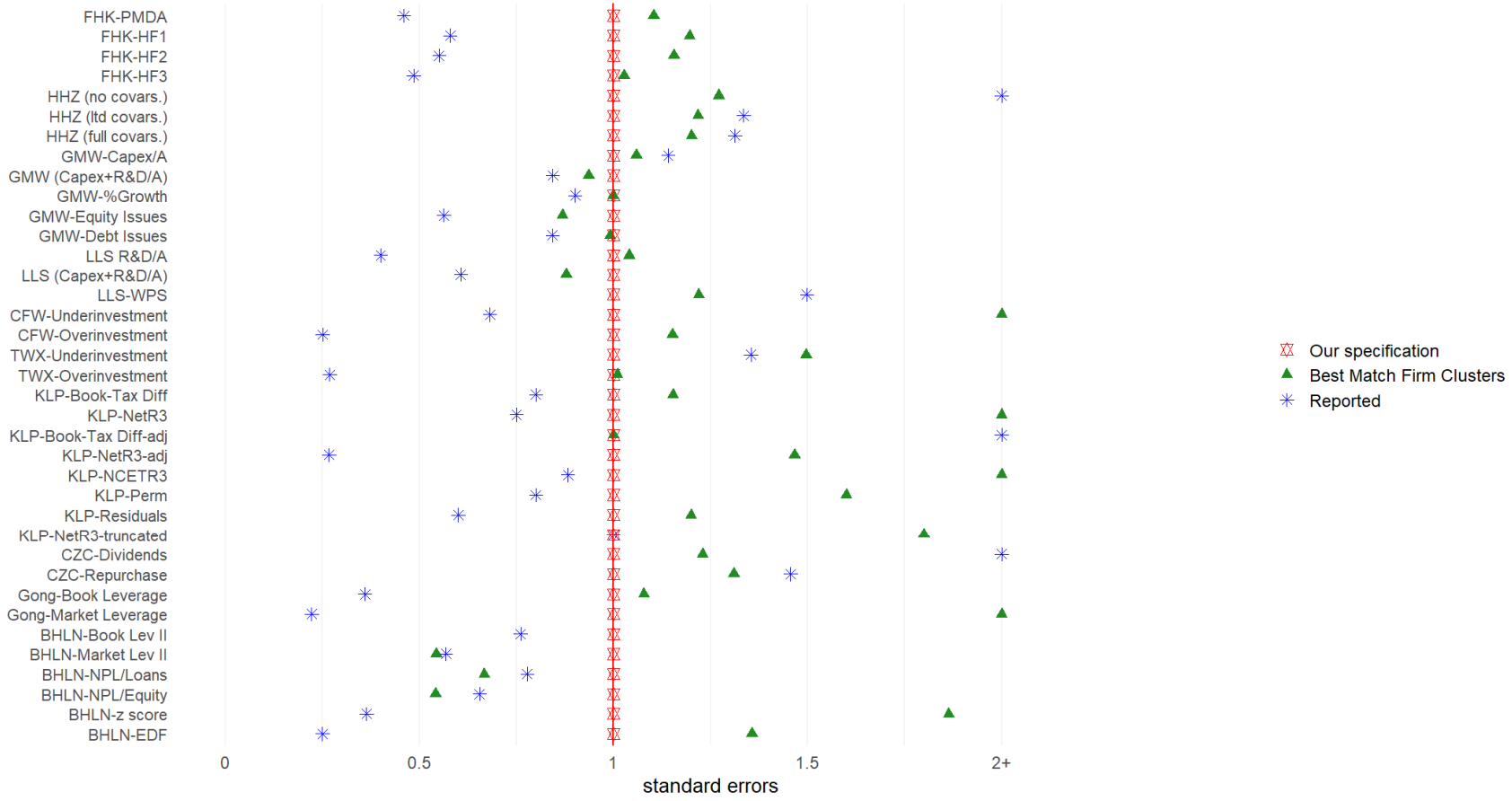


Figure 7. Comparing Sample Sizes

Figure shows, for each reexamined paper, number of firms (Panel A) and number of firm-year observations (Panel B), for our specification (scaled to 1.00; best match specification, and reported results. Some re-examined papers do not report number of observations or firms. For FHK HF-1, we assume the number of firms is same as for FHK PMDA. For Gong, we use quarterly observations.

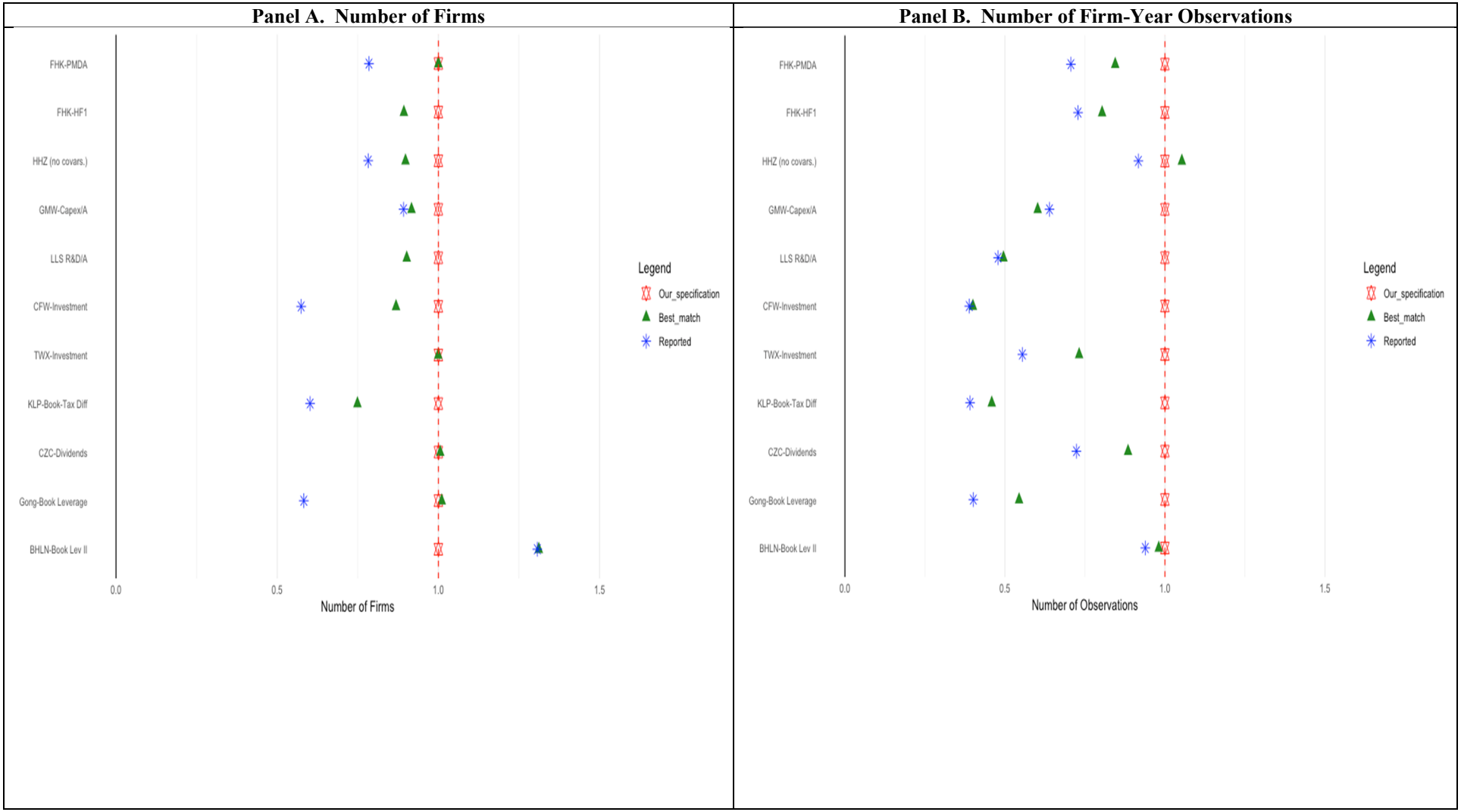
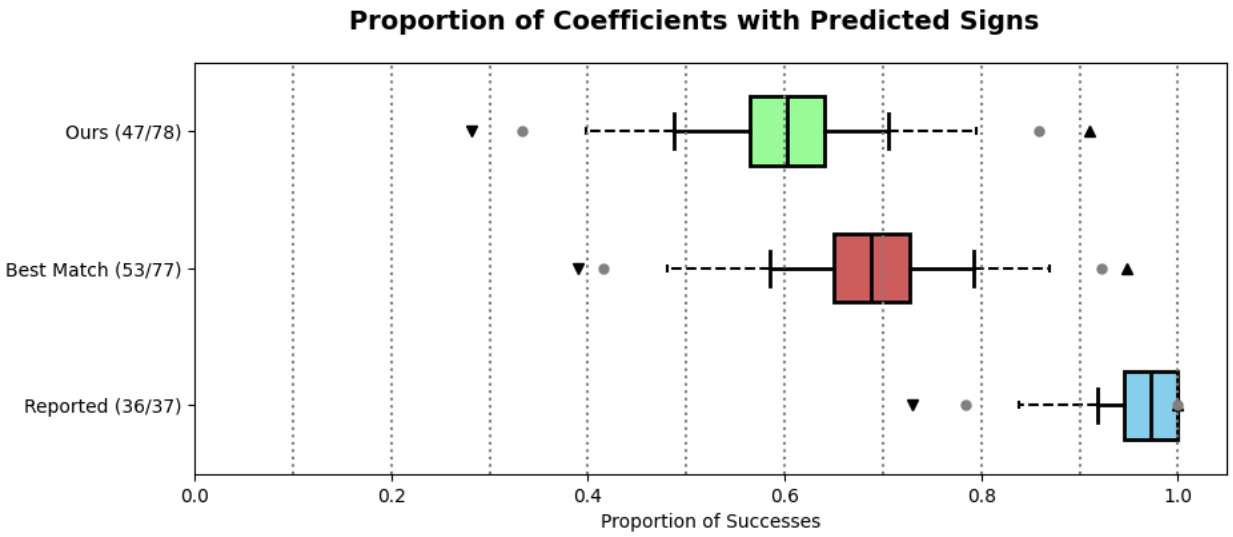


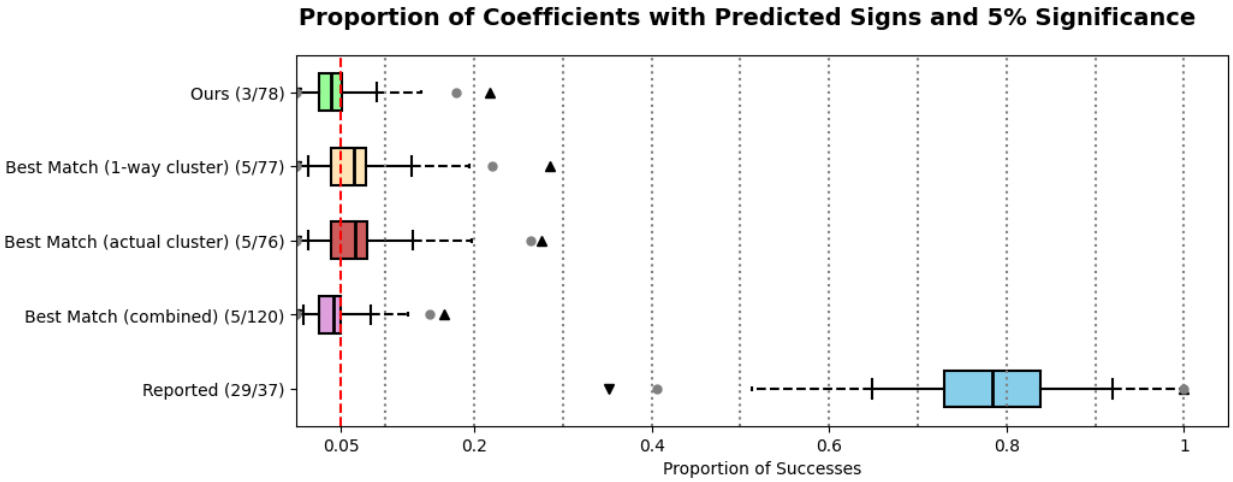
Figure 8. Binomial Distributions

Figure presents box plots for 1,000,000 draws from binomial distributions with indicated probabilities of "1" outcomes. **Panel A**. Same sign as predicted outcome. **Panel B**. Same sign as predicted outcome and statistically significant at 5% level (2-sided test). **Panel C**. Same sign as predicted outcome and statistically significant at 10% level. **Panel D**. Same sign as predicted outcome and statistically significant at 1% level. **All panels**. All outcomes are assumed to be independent. Box plots show 25th percentile, mean, and 75th percentile. Inner whiskers show 2.5 and 97.5 percentiles; outer whiskers show 0.01 and 99.99 percentiles. Triangles show minima and maxima across simulated samples for 1 billion draws. Grey dots show minima and maxima for separate draw of 1 million samples. No minima or maxima are shown for confidence intervals obtained analytically (Panel D). **Panels B-D**. Red vertical dashed line shows significance level (5%, 10%, or 1%).

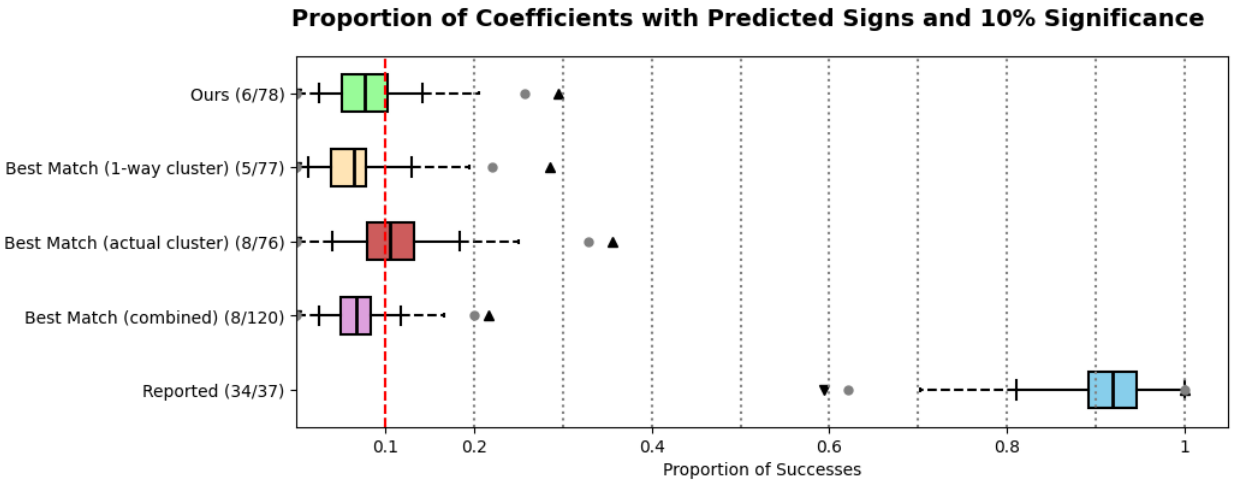
Panel A. Sign Same as Predicted



Panel B. Same Sign and Statistical Significance at 5% level



Panel C. Same Sign and Statistical Significance at 10% level



Panel D. Same Sign and Statistical Significance at 1% level

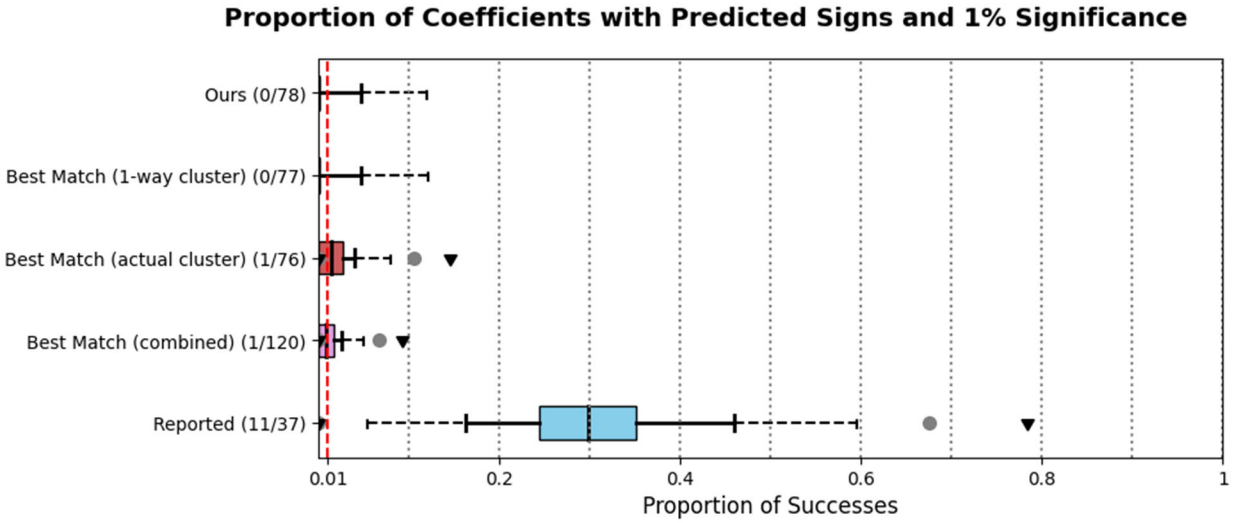
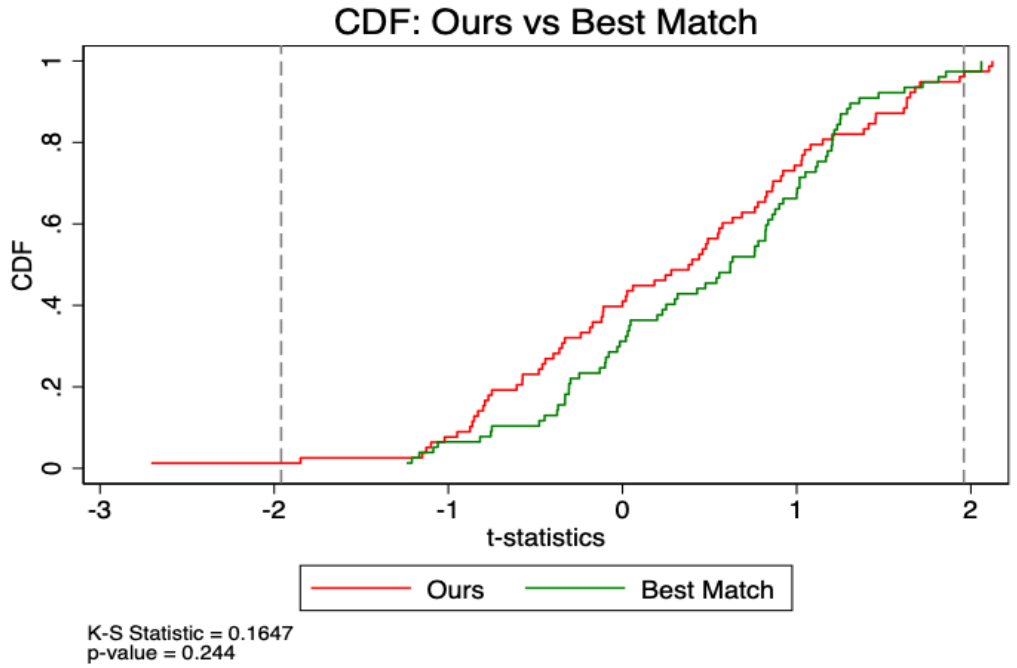


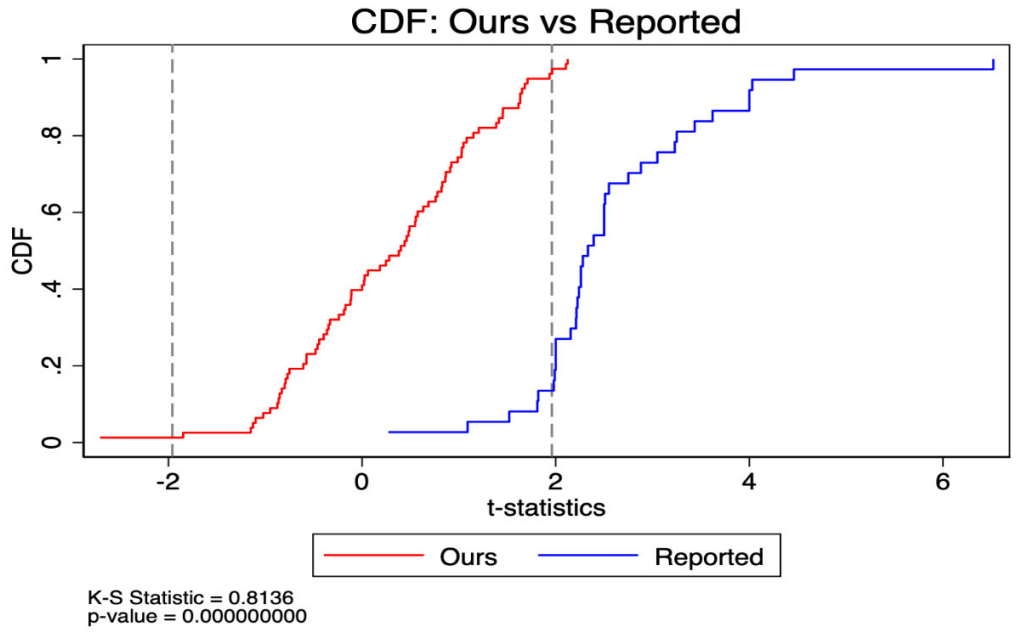
Figure 9. Cumulative Distribution Functions and Kolmogorov-Smirnov Tests

Figure presents cumulative distribution functions (CDFs) for *t*-statistics for our specification, best-match specifications, and reported results. Figures also report Kolmogorov-Smirnov (K-S) test statistic and p-value for comparison of CDFs. **Panel A**. Our versus best-match specifications. **Panel B.** Our specification vs. reported results. Panel C. Best-match specifications vs. reported results. **Panel D**. Ours plus best match specifications (120 combined specifications) vs. reported results. **All panels**. *t*-statistics are same as in Figure 4 (positive except that we use negative *t*-statistics when coefficient sign is opposite from expected). Dashed vertical lines at ± 1.96 indicate 5% significance**.**

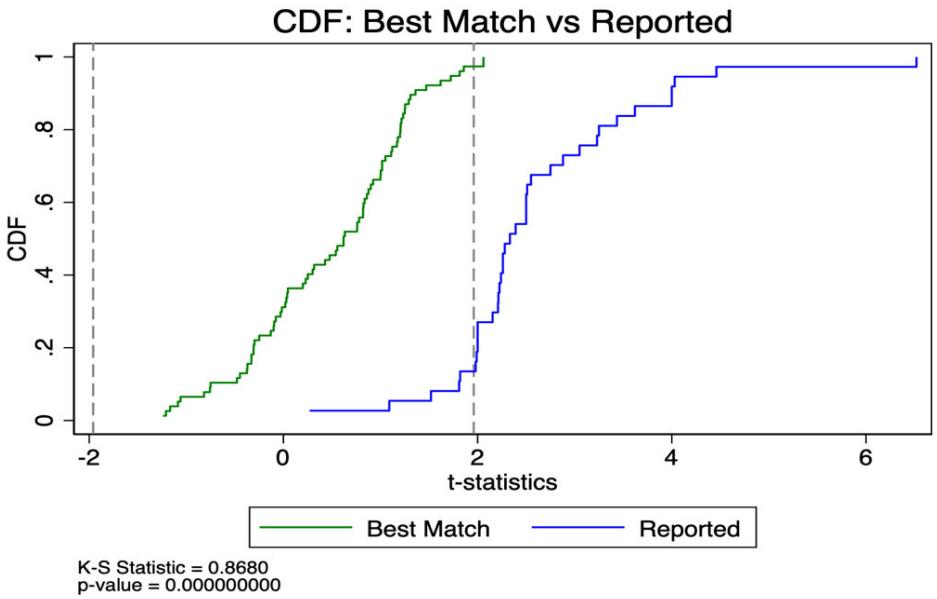
Panel A. Our Specification vs. Best-Match Specifications



Panel B. Our Specification vs. Reported Results



Panel C. Best-Match Specifications vs. Reported Results



Panel D. Combined (Our and Best-Match) Specifications vs. Reported Results

