The Road to False Positives: Sample Selection and Specification Choice in Randomized and Natural Experiments

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, School of Accountancy

(draft October 2024)

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

Preliminary Draft: Please do not cite

The Road to False Positives: Sample Selection and Specification Choice in Randomized and Natural Experiments

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

Abstract: During 2005-2007, the SEC conducted a randomized trial in which it removed short-sale restrictions from one-third of the Russell 3000 firms. Early studies found no effect of removing the restrictions on short interest, share returns, volatility, or price efficiency. In prior work, we confirm the lack of evidence for a natural causal channel that could explain these results (Black et al., 2024). Yet over 80 studies report evidence for a wide range of indirect effects on firms from the experiment. Given the lack of a causal channel, many of these results are likely to be false positives. We confirm that suspicion by closely reexamining the principal results from 10 of these studies using a simple specification with firm and year fixed effects. None of the results survive. We then examine best-match specifications that closely follow the sample selection, methodology and specification reported in the respective papers. Again, we mostly obtain null results. Our analysis finds that the gap between reported results and best match results is due to either reported coefficients being much higher or reported standard errors being much smaller than what we obtain using best-match specification. The vast gap between best-match results and reported results is surprising. These papers have apparently made choices that are not evident from the explanation of their research design and these choices seem to have a significant impact on the reported results. Our results suggest that researchers retain extensive discretion over the sample and model specification, even (as here) for a true randomized experiment. The choices in these studies produced statistically significant results when other reasonable choices would not. We draw lessons from our analysis for empirical practice and for ways in which researchers can find false positives results.

Keywords: natural experiments; causal channels; specification choice; Regulation SHO; SEC experiment; short-sale experiment

Conflict-of-interest disclosure

Bernard Black

I have nothing to disclose

Hemang Desai

I have nothing to disclose

Kate Litvak

I have nothing to disclose

Woongsun Yoo

I have nothing to disclose

Jeff Jiewei Yu

I have nothing to disclose

Disclosure regarding related paper

This paper builds on and substantially expands 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. In a few instances, specific results for four of the 10 papers studied here, were previously reported in BDLYY.

Overall, 11 of the 70 regression results in Tables 1-3 were reported previously, as were four of the 11 graphs in Figure 1. None of the other figures in this paper were reported in BDLYY. Some additional results for these four papers were reported in the BDLYY (2024) online appendix.

Contents

The Road to False Positives: Sample Selection and Specification Choice in Randomized and Natural Experiments

I. Introduction

The issue of p-hacking or specification search, in which authors search for significant results, is gaining attention in accounting and finance research (e.g., Harvey 2014, 2017, 2019, Ohlson 2020, Hail, Lang, and Leuz 2020, Gow 2023, Menkveld et al. 2024). So is the non-replicability of published results (e.g., Dreber and Johannesson 2024; Perignon et al. 2024).

This paper contributes to the literature on specification search and false positives by providing direct, detailed evidence on how sample selection and research design choices produce false positives. We re-examine the core results from ten studies that report significant, indirect effects attributable to the SEC's Regulation SHO (Reg SHO) experiment. In this experiment, the SEC conducted a randomized trial (May 2005 to July 2007) in which it suspended short-sale restrictions for one-third of the Russell 3000 (R3000) firms (pilot firms). The remaining R3000 firms served as the control group.

The SEC experiment is ideally suited to provide evidence of specification search and false positives for a number of reasons. First, early studies of the experiment found no effect of removing the short-sale restrictions on short interest, stock returns, volatility, or price efficiency for the pilot firms (Diether et al., 2009; Alexander and Peterson, 2008; and the SEC's own study in 2007). Yet an array of (to date) 80 published studies and working papers report a vast array of *indirect* effects and attribute them to the short-sale experiment (Internet Appendix Table A1). In our prior work, Black et al. (BDLYY 2024), we reexamined three primary posited causal channels (short interest, share returns, and managerial fear) for these indirect effects using a more comprehensive sample and a longer period. We found no evidence to support any of these channels. Other studies find no support for a price efficiency channel (see studied cited above, and De Angelis, Grullon and Michenaud, 2017). Thus, studies that report indirect effects from the experiment lack a causal channel that can produce those effects. This raises the likelihood that the reported results are false positives.

Second, we can identify from the SEC releases the pilot and control firms. This facilitates re-examination as does our decision to re-examine outcomes that can be easily computed from standard databases such as CRSP and Compustat.

Third, the random assignment of firms to treatment (pilot) and control provides covariate balance between pilot and control firms (see Internet Appendix Table IA-4). Thus, one can obtain valid inference by comparing univariate differences between the pilot and control firms or running simple regressions of the outcome on a treatment dummy, with firm and year fixed effects, but without covariates. If these simple approaches do not produce statistically significant results, this calls into question the validity of results found using more complex models.

We assess the robustness of the key reported indirect effects attributed to the short-sale experiment by re-examining core outcomes in 10 studies. Overall, we examine 70 specifications involving 58 distinct outcomes. Of these, 31 specifications (29 outcomes) were studied by the authors; we study related outcomes and related specifications to assess robustness.[1]

Almost none of the results replicate with our pre-specified research design, which uses a sample that closely matches the actual SEC experiment and with firm and year fixed effects. Moreover, when we conduct "best match" analyses, where we follow each paper's sample selection, variable definition and methodology as closely as we can, "best match" results are still almost never statistically significant.

We find large differences between the best-match results and the reported results for both the coefficients and s.e.'s. The reported coefficients are larger, sometimes much larger than those obtained from the best-match analysis. Also, the reported s.e.'s are often smaller, sometimes much smaller. These gaps between best-match and reported results suggest that each paper makes important, unreported specification choices, that our best-match analysis did not capture.

For two studies (Fang, Huang and Karpoff, 2016 and Hope, Hu and Zhao, 2016) we have the authors' exact sample and code though not their sample selection steps. This allows us to

[1] Some results presented here for the first four studies are drawn from BDLYY (2024) either in the text or from the internet appendix.

explicitly illustrate how certain specification choices made by the authors produced significant results while several other reasonable choices did not. We illustrate how specification choice results in statistical significance for the core result in Fang, Haung and Karpoff (FHK) in section VI. The detailed analyses of our attempt at replication of both these studies is reported in the Internet Appendix. [2]

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 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 results exploiting a randomized experiment are highly sensitive to specification choice, with a strong tendency for reporting false positives.[3]

This paper proceeds as follows. Section II summarizes the studies we examine, our sample selection steps, and other sample 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 illustrates how specification choices produce significant results using Fang, Huang, Karpoff (2016) as an example. Section VII discusses implications from our project. Section VIII concludes.

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] In response to this project, FHK publicly posted their code and sample for their PMDA result, but not for their HF-score results. However, we can use their accruals sample to conduct what should be a near-exact replication of their HF-score results. Hope, Hu and Zhao (2016) provided their sample and code to us privately; we thank them for doing so.

[3] We will post the Stata code and data sets needed to generate the results in this paper and the SAS code we used to generate starting data sets on Professor Black's website.

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 that they: (i) were published in peer reviewed journals;[4] (ii) rely on different causal channels; and (iii) provide evidence for a variety of indirect effects. We reexamined only outcomes that rely on data from standard sources, which both eases the task of reexamination and permits replication by others.[5]

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, which they measure using performance-matched discretionary accruals (PMDA). They report that the pilot firms have lower PMDA than control firms during the experiment period (2005-2007).

HHZ conjecture that suspension of price tests would increase short selling at the pilot firms. This would increase the risk that these firms would face securities litigation alleging auditing errors, and thus 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 of pilot firms, which will lead to lower investment and lower capital raising.

LLS argue that relaxing short sale restrictions makes stock prices more informative and conjecture 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. They also hypothesize that managers will rely more heavily on the more

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

[5] 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.

informative prices to guide business decisions, so 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.

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.[6] 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 delisted on June 28, 2004) and 1,969 control firms,

[6] 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.

as of experiment announcement on July 28, 2004. Of the re-examined papers, only Gong excludes these 44 firms; many lose large numbers of firms in matching to Compustat.

We next update this sample to the experiment start on May 2, 2005, using the SEC's updated lists from 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. None of the re-examined papers conducts this update.

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.

C. Other Sample Specification Choices

We summarize here the principal additional specification choices made in our pre-specified research design.

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 TWFF 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. See details in Internet Appendix Table IA-3.

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). We show below that downward bias can be severe.

3.. Use of Covariates

Our research design does not include time-varying covariates. The re-examined papers use widely varying 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 (TWX, BHLN). 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.

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

In our research design, we rely on an unbalanced panel. We confirm, in the pre-analysis plan, 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 create 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 the R3000.

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

We measure firm performance over fiscal years. So do seven of the reexamined 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 month-ends from January-May, fiscal year = (calendar year – 1); and otherwise fiscal year = calendar year. Thus, our sample period includes fiscal year ends from 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 drive 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.[7]

For a number of Compustat variables, missing values may reflect true zeros. When this

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

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 expense, acquisitions; sales of property, plant and equipment (PPE), dividends, and share repurchases. Eight of the re-examined papers are silent; we infer that they do not replace missing values with zero (the exceptions are TWX and KLP). Failure to do so, for variables with many true zeros, can produce large loss of sample size, and skewed loss, since the firms with missing values will not be a random set of all firms.

D. Graphical Evidence: Univariate and Leads-and Lags Graphs and When Should Any Results Be Expected?

Good DiD practice includes presenting univariate graphs for the treatment and control groups, leads-and-lags graphs, or both. These graphs can be used to confirm whether pre-treatment trends appear parallel (central for DiD credibility), 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 good balance between the two groups.

Examining time trends during the experiment period is especially revealing for the SEC experiment because: (i) the SEC proposed in December 2006 to repeal the short-sale price test for all firms;[8] (ii) the SEC in fact repealed the short-sale price tests for all firms as of July 3, 2007; and (iii) by 2007, market participants should have realized that there were no known, major effects on firms from the experiment. For all of these reasons, one should expect any treatment effect to be stronger early in the experiment period, weaken or disappear in 2007, and disappear in the Post period. Indeed, the same anticipation concerns that led many of the re-examined papers to exclude 2004 from the Pre period as a transition/anticipation period could well call for excluding 2007 from the During period.

[8] SEC Release 34-54891 (Dec. 13, 2006).

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.

B. Difference-in-Differences (DiD) Specification

1. Simple DiD Specification

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

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

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, 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.

2. Annual Differences and Leads-and-Lags Specification

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 data are available, we extend these plots back to 1998. A longer pre-treatment period is useful in identifying non-parallel pre-treatment trends, and in assessing 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.

3. 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 $\boldsymbol{\gamma}$ 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 based on our specification (that is, our sample and research design as described in section II and III). We apply our specification to study 58 distinct outcomes and a total of 70 different model variations. Given covariate balance between treatment and control firms, from the randomization, univariate graphs should be informative about whether there is any treatment effect. In Figure 1, we illustrate the value of these graphs by plotting 11 core outcomes -- one for each study (two for FHK).

None of the 11 graphs shows convincing evidence of a treatment effect. However, several graphs show non-parallel trends in the Pre period, which could drive DiD results.[9] For example, the LLS graph for CEO wealth-performance sensitivity (WPS) shows that WPS fell for pilot firms

[9] LLS, CFW; TWX; KLP.

relative to control firms in 2004, before the experiment started, and no relative change over 2005-2007. LLS's DiD regression result, a significant decline in WPS for pilot firms during the experiment, is driven by a Pre-period gap, which closed in 2004, and by their decision to drop 2004 from their analysis, for which there is no valid reason.[10] If 2004 is included in the pre-period, there is no evidence for a relative decline in WPS. In addition, several other graphs show erratic treatment-vs-control gaps in the Pre and/or Post periods.[11] 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 the pilot firms. For some others, treatment-vs-control gaps are largest in the Pre and/or Post period.[12] Most show no evidence of reversal of any pilot-versus-control gap 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, 25 coefficients (out of 70) have signs opposite from predicted (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. These studies show evidence of non-parallel trends in the pre-period. The third is LLS, which also has non-parallel pre-treatment trends. Even leaving aside the issue of lack of evidence from graphs, so few significant results could easily arise by chance.

That we find virtually no evidence for a treatment effect across so many outcomes from ten studies is remarkable. In contrast, the re-examined studies report 31 specifications (29 outcomes) and find statistical significance for 26 of these specifications. While our specification differs from the re-examined studies, we started with *their chosen outcomes*, plus related

[10] Recall that LLS argue that suspension of price tests increased the price informativeness of pilot firms' stock prices. They conjecture that with more informative prices to guide decisions, pilot firms had less need to provide direct incentives to CEOs, so WPS would decline. However, price informativeness could not improve until the experiment actually began, in May 2005.

[11] FHK accruals; FHK HF-score.

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

[13] CZC, Gong.

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 we can, the specification and sample in each paper based on the authors' own description.[14] We report best-match regression results in Table 2 and annual univariate graphs in Internet Appendix Figures IA-6 to IA-18.[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. There is a tendency for the best-match coefficients to be closer to the reported results than those with our specification. However, most of the best-match coefficients remain insignificant and often far from the reported coefficients. Across 70 specifications, 19 coefficients have opposite-from-predicted sign. With s.e.'s clustered on firm, we find weak significance for five coefficients with the predicted signs. However, one for GMW is unconvincing given the univariate graphs, one for LLS is driven by non-parallel pre-treatment trends; and three for CFW are driven by a combination of 2007 (wrong time for treatment effect to appear) and non-parallel pre-treatment trends. This number of significant coefficients could plausibly arise by chance.

Using two-way clustered s.e.'s, for the five studies that use them, five more coefficients become significant (3 of these for FHK HF-score; two for KLP) but the two CFW results lose significance. To summarize the results from best-match specifications:

> FHK accruals: No significant results across 8 specifications, with either balanced or unbalanced panel.

[14] For FHK and HHZ, our best-match analysis was developed without knowing their actual choices and sample.

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

FHK F-score: No significant results across 12 specifications (3 are significant with two-way clustered s.e.'s).

HHZ auditor fees: No significant results across 3 specifications.

GMW: One mildly significant result out of 8 outcomes.

LLS: One mildly significant result (driven by non-parallel pre-treatment trends).

CFW: Three significant coefficients across 8 outcomes with s.e.'s clustered on firm (driven by 2007 and non-parallel pre-treatment trends), of which two lose significance with 2-way clustered s.e.'s.[16]

TWX: No significant results across 8 outcomes.

KLP: No significant results across three outcomes (two are significant with 2-way clustering).

CZC. No significant results across 6 outcomes.

Gong: Not significant for two outcomes.

BHLN: Insignificant results for 5 outcomes (the sixth is mis-specified[17]).

In the Internet Appendix, we present intermediate results, in which we move step-by-step from our specification to the best-match specifications. These are also almost always insignificant.

When we began this project, we expected that many results would not be robust, using our pre-specified design, 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, s.e.'s, and often both. We view the gap between best-match results and reported results as strong evidence of both the importance of specification choice, and the importance of choices that the authors deemed not worth stating in their papers.

[16] Based on the CFW *corrected* best-match specification; see Internet Appendix Table IA-23-23.

[17] For this outcome, book leverage is both the outcome and a covariate. See Internet Appendix for details.

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 31 specifications for which the authors report results.[18] The picture changes dramatically. Now, 30 of 31 coefficients have the predicted sign; of which 26 are significant and two more are marginally significant.

In Figure 2, we summarize the differences in 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 all 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 *sign* is opposite from the reported sign. With our specification and the best-match specifications: almost all coefficients are insignificant; many have the opposite sign from reported; and within the very few significant coefficients, most are only slightly above +1.96. Moreover, there is a dramatic difference between the reported results and those with our specification or the best-match specifications. The vast bulk of the reported coefficients are statistically significant, sometimes highly so.

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 sources contribute to larger reported *t*-statistics in the re-examined papers. Figure 3 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 our specification (scaled coefficient = best-match coefficient/our coefficient) and the reported coefficients (similarly scaled). The reported results are winsorized at 6, to avoid compressing the x-axis. Many coefficients with our specification have the opposite sign to those reported (also seen in Figure 3); so do a number of best-match coefficients. The best-match

[18] We do not replicate two specifications for BHLN, see explanation below.

coefficients tend to be closer to the reported coefficients, relative to our specification**.** But the dominant impression from Figure 3 is that reported coefficients are much larger in magnitude than ours, and often much larger than best-match coefficients.

In Figure 4, we use a similar scaling approach for s.e.'s. The figure reports s.e.'s for our specifications (scaled to 1.00), s.e.'s for the best-match specifications (scaled to our s.e.'s, with clustering on firm) and reported s.e.'s (clustering varies). The best-match s.e.'s tend to be *larger* than with our specification. However, the reported s.e.'s are mostly smaller than s.e.'s obtained with either our specifications or the best match specifications**.**

Figure 5 explores one reason for smaller reported s.e.'s. This figure is limited to the 5 studies which use two-way clustering. It compares best-match s.e.'s clustered on firm (scaled to 1.00) to best match s.e.'s with two-way clustering (scaled to s.e.'s with only firm clusters). As the figure shows, the two-way clustered s.e.'s are mostly smaller, sometimes much smaller, than with one-way clustering.[19]

Figure 6 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 *disclosed* sample-selection choices that lead to substantial loss of firms; this leads to best-match sample size being smaller than our sample. More surprisingly, all except BHLN lose additional firms in the reported results for unclear reasons, sometimes substantial. In the figure, this leads to reported counts being to the left of the best-match counts.[20]

Panel B is similar but shows the number of *observations*. Our specification has more observations than the best-match specifications for all but HHZ.[21] The best-match specifications

[19] Gong is an exception; both sets of s.e.'s are similar. This makes sense: she uses quarterly data, so has more time clusters, which avoids the tendency for clustered s.e.'s to be downward biased with a small number of clusters.

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

[21] HHZ have more best-match observations, despite fewer firms, because their Post period includes 2011-2013,

lose observations for a variety of reasons, including dropping a transition year, not including a Post period, requiring a balanced panel or continued membership in the R3000, and requiring data on covariates, even for results without covariates. More surprisingly, all lose additional observations, for unclear reasons, in moving from the best-match specification to the reported results. This additional sample loss likely plays an important role in the differences between the best-match and reported coefficients. Yet this cannot be a full explanation for the large differences we found which exist even for studies with similar best-match and reported sample sizes.

Overall, using a standard DiD specification with firm and year fixed effects, there is no evidence of a treatment effect for any of the studied outcomes. When we move to best match specifications, almost all 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 s.e.'s using either our specification or the best-match specification. The reported results also rely on fewer observations than our best match sample. Overall, our analysis suggests that the authors of these 10 studies have made some choices that were not clearly stated in the research design.

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 70 model variations, 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, document the loss of sample size relative to our specification and discuss the often questionable theoretical basis for the reported results. See the Internet Appendix for similar graphs for the remaining specifications.

---

while ours ends in 2010.

VI. Illustration of Importance of Specification Choice: FHK

In this section, using the sample and code posted by FHK, we illustrate how the authors' specification choices produced statistical significance when other reasonable choices would not. Moreover, several important specification choices became known only from their posted code but were not mentioned in the paper. We examine two FHK outcomes, PMDA and HF-score. FHK posted data and code only for PMDA. However, we have FHK's exact sample and can use their sample to evaluate their HF-score results. We have the final FHK sample but not their sample selection code or sample selection steps.

A. FHK Results for PMDA

In Table 4 we report results for four different accrual measures, for both a balanced sample (used by FHK) and the corresponding unbalanced sample. Operating and total accruals are calculated following Richardson et al. (2005). Abnormal accruals (AA) and PMDA are calculated following FHK. Note that PMDA is calculated as AA of a sample firm minus AA of a matching firm. Panel A of Table 4 reports results using our specification. We then we move from our specification to the FHK Best-Match Specification and then to FHK's actual specification by making one specification change at a time.

As mentioned before, using our specification, the results are insignificant for all outcomes. However, examination of Table 4 reveals several choices made by FHK that produced significance, where other reasonable choices did not. We discuss here four key choices.

First, FHK chose to use PMDA as the only measure of earnings management. This is an unusual choice. PMDA is less powerful than the other measures in Table 4. Its use is recommended only when extreme or correlated performance needs to be controlled for (Dechow et al., 2011). Since the SEC experiment was a randomized trial with good covariate balance between pilot and control firms, use of PMDA was not called for. One can see that even using FHK's exact sample and specification, results are insignificant using other measures, including AA. Thus, FHK's significant result is driven by AA of the matching firms.

Second, FHK chose matching firms for computed PMDA based on lag ROA. This is another unusual choice since Kothari et al (2005) show that choosing matching firms based on current-year ROA produces a better-specified model.[22] The choice to select matching firms based on lag ROA drives their PMDA results (Table 4, Panel H). If we FHK were to select matching firms based on concurrent ROA as recommended by Kothari et al (2005), the PMDA coefficient would be small (-00007) and insignificant (Internet Appendix Table IA-13, Panel F).

Third, FHK reported results only using a balanced panel that requires a firm to be present from 2001-2010. This is an unusual choice. Other re-examined papers all use unbalanced panels. Use of a balanced sample substantially reduces sample size, introduces survivorship bias, and is not needed since the SEC experiment was a randomized trial, with no evidence of differential attrition between pilot and control firms (as we confirmed in our pre-analysis plan). FHK state in their paper that all results are "similar" using an unbalanced panel (FHK at 1262). However, results using an unbalanced panel are never significant even using matching firms selected based on lagged ROA (Table 4, col. (4)).

Fourth, FHK chose both to cluster s.e.'s on firm and year, despite the risk of downward bias from clustering on year for a short panel, and to cluster on fiscal year for a dataset organized by calendar year. Their PMDA results are insignificant when clustering on firm only, or with clustering on firm and calendar year instead of on firm and fiscal year.[23]

This example illustrates how a series of specification choices produce significant results while many other reasonable choices would not. If one were to change even one of these choices, their result becomes insignificant.

---

[22] We learn about this choice only from their code. The FHK paper states they calculate PMDA following Kothari et al (2005), For the FHK Best-Match Specification, we inferred that they matched on current-year ROA 's approach but fail to mention that their PMDA is computed using matching firms matched on lag ROA.

[23] This choice, too, we learn about only from their code. The FHK paper states they cluster s.e.'s on firm and "year." For the FHK Best-Match Specification, we inferred that they clustered on firm and calendar year.

B. FHK Results for HF-score

FHK did not post code or data for their HF-score results, but we know their exact sample. Dechow et al. (2011) specify how to compute F-score; and computing HF-score from F-score is mechanical. So we should be able to generate near-exact replication for the HF-score results.

The replication effort fails. See Internet Appendix Table IA-18. Our HF-1 coefficient is -0.090, roughly half of their -0.178 in magnitude, and insignificant. For HF-2 and HF-3, our coefficients have the opposite signs (HF-2: +0.040 vs. their -0.189; HF-3: +0.041, vs. their -0.200). With an unbalanced sample, all HF-score coefficients are positive (opposite from predicted). The very different coefficients across our specification, the FHK Best-Match Specification, the near-exact replication, and the reported results confirm the HF-score estimates are highly sensitive to specification.

VII. Discussion

Below we discuss several important takeaways from our study, which can inform future research as well as the editorial and review process.

A. General Takeaways

1. An important takeaway from our paper is the importance of confirming a causal channel for hypothesized results. There is evidence from multiple studies cited earlier that the Reg SHO experiment had no meaningful impact on short interest, stock returns, share volatility, or price efficiency. A manager-fear channel, relied on by FHK, GMW in part, and other studies, cannot be directly ruled out, but is not supported as there was minimal pushback from firms or controversy when the experiment was announced or when the SEC suspended the price tests for all firms (see BDLYY, 2024 for a detailed examination of the fear channel). Lack of a causal channel does not preclude a true positive, but raises the risk of false positives, and should call for closer scrutiny of results, including whether they are robust to alternate specifications and to alternate but related

outcomes. Many of the reexamined results depend directly on an assumed causal channel that is not tested, or which is not supported by other studies that do test the channel.[24]

2. Second, an underlying theory is important, even though not all results without a sound underlying theory are wrong. So is assessing whether results are of plausible economic magnitude. Most of the papers as we have discussed above lack a sound economic intuition and have weak theoretical grounding, even if a causal channel existed.[25]

3. Perhaps the most surprising finding (for us) was that the best-match results remained mostly insignificant and often far from the reported results. This suggests that each of these papers made important specification choices that were not discussed or disclosed in the paper. These specific choices produced significant results when a wide variety of specifications that we tried did not, including the best-match specifications.

One plausible response to this is for authors to post replication datasets and code, including sample selection code. Finance and accounting journals are moving in this direction. Even if researchers cannot post data from commercial datasets, they can post full code; and (ii) make data available to researchers who have their own access to the commercial datasets, or provide firm identifiers (Compustat GVKEY, CRSP PERMNO, etc.). Posting code will ensure that important specification choices are knowable, in a replication effort.[26]

---

[24] For example, LLS posit that relaxing short-sale constraints makes share prices more informative. However, as discussed earlier, there is no evidence that the experiment affected price informativeness. HHZ posit a channel running from "short sales driv[ing] prices down" to increased litigation risk and thus to auditor fear of liability (HHZ at 480). However, they offer no evidence of share price declines and other studies do not find this evidence. They also do not assess whether there was an actual effect of the experiment on litigation risk.

[25] Two more examples: He and Tian (2016, abstract) who report a drop in patenting activity, supposedly reflecting firms' exposure to "patenting-related litigation initiated by short sellers." It seems doubtful that firms would change multi-year patenting strategies in response to a planned one-year experiment. Moreover, we have never heard of short-sellers conducting patent-related litigation.Bai, Lee, and Zhang (2020), who report a 17% increase in workplace accidents at pilot firms. The posited mechanism is reduced workplace safety investments "to meet short-term [earnings] targets." The magnitude is implausible. Safety investments likely do not dissipate right away, even if new investment is cut. Another sign of implausibility: Pilot firms had vastly fewer accidents during the pre-treatment period (coeff. = -3.69, p = .005); relative to sample mean of 8.44. This is inconsistent with random assignment.

[26] For example, Adams et al. (2019) and Leone, Minutti-Mexa, and Wasley (2019) discuss the importance of winsorization. Our re-examinations offer confirming examples. For CFW, not winsorizing nearly doubles the

4. Another surprising finding was how much specification choices can affect s.e.'s and coefficients, including choices not discussed or disclosed in the paper. We found that "standard errors" were not at all standard. One source of variation was using two-way clustering on firm and year; and for FHK, clustering on firm and fiscal year for a dataset organized by calendar year (Figure 5). Two-way clustering on firm and year can produce s.e.'s that are severely downward biased for some specifications and some outcomes. Clustering on firm, in contrast, appears to produce s.e.'s that are close to the exact s.d.'s from randomization inference.[27] Beyond that, for 8 of the 10 re-examined papers, reported s.e.'s are well below those from our specification, for some or all outcomes, including 7 for which reported s.e.'s are well below those from best-match specifications (Figure 4).

5. Our re-examination highlights the importance of preserving sample size, and the need to be explicit about how research design choices affect sample size. Our own sample is generally larger, sometimes much larger, than the reported samples and this seems to matter.[28] HHZ provide a model of how to specify sample selection steps. They specify each step they took and how it affected sample size. This made it easy for us to compare the effect of each step that we found to theirs, and understand how they lost sample size, relative to our pre-specified design and the HHZ Best Match. Conversely, LLS, TWX, and CZC do not even report the number of firms in their final samples. Several papers either did not report results without covariates or limited the sample, including univariate results, to firms with data on all covariates. Sample size can also be lost through lack of care in matching across datasets, which may affect the treatment and control groups differently.[29]

---

coefficient on Pilot*During. For Gong, not winsorizing would change the sign on Pilot*During. For LLS, not winsorizing would reduce the best-match coefficient for R&D/assets from 0.0036 (t = 2.12) to 0.0024 (t = 0.92).

[27] In the Internet Appendix, we provide both s.d.'s and s.e's with firm clusters for the papers that use two-way clustering. See Internet Appendix Tables IA-9, IA-12, |IA-14, IA-16, IA-18, IA-23, IA-26, IA-27.

[28] Above, we confirm the importance of preserving sample for HHZ, using their exact sample.

[29] HHZ provide an example where reported results are strongly affected by sample loss due to imperfect matching. See Internet Appendix Table IA-20.

6. Several papers include sample restrictions that impose survivorship bias. For example, by using a balanced panel, FHK require a firm to survive throughout 2001-2010. Given that the SEC experiment was a randomized trial with no evidence of differential attrition, there was no good reason for this restriction. [30] HHZ, GMW CFW impose survivorship bias by requiring R3000 membership in 2005.

7. As Ohlson (2022) discusses, one contributor to false positive results is the failure of authors to conduct a broad range of robustness, placebo, and consistency checks (e.g., result A implies result B), that might falsify their hypothesis. This includes studying multiple, related outcomes. A false positive is easier to find for a single outcome than for a set of related outcomes. Two examples from our re-examination: First, any results found during the experiment period should reverse when the experiment ends. Yet only four papers study reversal; of these CFW find reversal; FHK find reversal for PMDA but not HF-score; HHZ and CZC find no reversal. Second, FHK report results only for PMDA. This is a surprising choice as a measure of earnings management since there was no difference in performance between pilot and control firms, due to the randomized nature of the trial. Showing results with other measures of accruals would be a reasonable thing to do. We find insignificant for three other reasonable measures of accruals, using their exact sample (Internet Appendix Table IA-13).[31]

Some robustness checks will be project-specific. For the SEC experiment, one would expect any results to be stronger for NYSE than for Nasdaq firms because the Nasdaq bid test for short-selling was not as restrictive as the NYSE test (Diether et al., 2009).[32]

---

[30] FHK assert in a footnote that their results are "similar" with an unbalanced panel. As we show in detail in the Internet Appendix Table IA-12, their PMDA result is in significant with an unbalanced panel. As we show in a near-exact replication for HF-score (Internet Appendix Table IA-18), coefficients for HF-score with an unbalanced panel are very different from those with a balanced panel.

[31] A further example. GMW study two measures of investment. We study two others both have small, statistically insignificant coefficients (Tables 1 and 2). They also classify small firms based on total assets. Their results vanish when small firms are defined based on market capitalization.

[32] Of the re-examined papers, only Gong assesses whether effects differ between NYSE and Nasdaq firms. She finds that only pilot Nasdaq firms reduce leverage. An effect only for Nasdaq firms makes no sense because the Nasdaq bid test was less restrictive to begin with.

8. Defining sample periods deserves careful attention. Several studies dropped year 2004 from the analysis. The logic for doing so is not clear. For example, LLS rely on a price efficiency channel. However, this channel can operate only after the experiment started in May 2005. For this channel, there is no reason to exclude 2004 from the pre-period. This choice is consequential. We show above for LLS that their significant result for WPS is due to a combination of non-parallel pre-treatment trends and dropping 2004 from their analysis. HHZ offer another example. They conjecture that auditors would increase audit fees in response to fear of higher litigation risk for pilot firms. But audit fees for 2004 would have been set in 2003, when the identity of the pilot firms was not known. Thus, 2004 should be included in the Pre period, rather than excluded 2004 from their analysis. If 2004 were included in the pre-period, the coefficient on Pilot*During would lose significance.[33]

9. Our re-examination illustrates the power and importance of graphing DiD results to assess whether pre-treatment trends are parallel, and whether the time pattern of results during the experiment period makes sense. We view univariate graphs, leads-and-lags graphs, or both, as essential, for any DiD analysis with panel data. Yet none of the re-examined papers use these graphs. For several re-examined papers, pre-treatment trends were not parallel. For others, results were driven by 2007, which is the wrong time for a treatment effect to appear or strengthen.

10. Finally, specification *choice* is unavoidable, but specification *search* is not. By specification search we mean instances where authors consider different outcomes, specifications, and/or samples; report statistically strong results; and either do not report or downplay, weaker results with other outcomes, specifications, or samples. For example, FHK study only PMDA, and report results where matches are selected based on lag ROA which is an unusual choice. They do not report results using simpler accruals measures, and report only balanced panel results. They report results for HF-score, defined with an extreme threshold, but not for the underlying F-score, which would have been much more intuitive. CZC study dividends and share repurchases, but not total cash payout, and do not scale the payouts by firm size. None of the papers using two-way

[33] See Internet Appendix Table IA-20.

clustering report results with one-way clustering on firm.

Overall, an important takeaway from our reanalysis is that even when researchers begin with a randomized trial, they must make many specification choices. These decisions offer opportunities for specification search or "lucky" specification choices to produce false positive results. If results exploiting the SEC short-sale experiment are as sensitive to specification choice as we found, the same could be true for other DiD studies of true and natural experiments. Many natural experiments result from regulatory change and/or involve policy implications; hence it is particularly important that the results be robust and credible. Replication studies, ours included, can be part of the response to specification search.

B. Implications for the True Effects of the SEC Short-Sale Experiment

None of the 10 reexamined studies provides convincing evidence of indirect effects from the SEC experiment. Many of the others are likely also false positives. We view the following as minimum steps toward credibility for these studies: (i) specify a causal channel and provide supporting evidence; (ii) assess whether the results survive using our pre-specified sample, and to the extent feasible our specification; (iii) confirm pre-treatment balance on outcomes and covariates; (iv) assess whether results are strong enough to meet the Heath et al. (2023) critique of multiple hypothesis testing; (v) assess a range of alternate, related outcomes, when these are available; and (vi) post detailed code, including code to create the sample,; and, to the extent possible, the full dataset or identifiers for standard datasets that others can use to replicate the sample. For your own sample and specification, if different, (vii) follow closely the SEC's rules to determine which firms are pilot and which are control.[34]

Conclusion

In this paper we re-examine key findings in ten papers reporting indirect effects from the SEC short-sale experiment, and find no meaningful support for the core results in any of them.

[34] We publicly posted our sample on Prof. Black's website at Northwestern Law School, together with full replication code for BDLYY (2024). We plan to post similar code and data for this project. Many of our specifications should apply broadly, including our sample periods; use of firm and year FE; preference for results without covariates; providing annual graphs for treated and control firms; and assessing reversal after the experiment ends.

Results with both our pre-specified design and with the best-match designs are substantially different from the reported results. Statistical significance disappears in almost all cases, and when it survives, is weak and likely spurious. There is not a single road to the apparent false positive results we find, but instead multiple roads, often complementary.

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.

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 Simulated Power Analysis for Observational Studies: Application to the Affordable Care Act Medicaid Expansion, 213 *Journal of Public Economics* 104713 (2022).

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.

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*, 2024:1-19.

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.

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. (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.

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 (2020), 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.

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, xx *Journal of Finance* yyy-zzz (forthcoming) (http://ssrn.com/abstract=3961574).

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, Feix Holzmeister, Jurgen Habe, Magnus Johanneson, and Michael Kirchler (2004), Computational Reproducibility in Finance: Evidence from 1,000 Tests, *Review of Financial Studies* xx, yyy-zzz.

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

Richardson, Scott A., Richard G. Sloan, Mark T. Soliman, and Irem Tuna (2005), Accrual Reliability, Earnings Persistence and Stock Prices, *Journal of Accounting and Economics* 39, 437-485.

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.

Securities and Exchange Commission, 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. Results with Our Specification

All panels. Table summarizes results with our specification. See Internet Appendix for regression details. 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.

	(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)
B. 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)	
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								
Covariates	None	Limited	Full					
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*	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	R&D	(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					
Pilot*During	-0.0019	-0.0024	-0.0165					
s.e.	(0.005)	(0.005)	(0.016)					
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		
Pilot*Crisis	-0.0017	-0.0023	-0.3246	-2.0329	-0.0023	-0.0398		
s.e.	(0.0028)	(0.0058)	(0.4109)	(2.6401)	(0.0027)	(0.0362)		

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.

	(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]
B. 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]	
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	R&D	(Capex+R&D)				
Pilot*During	**2.8331****	0.2471	**0.5878****	1.0851*				
s.e. (firm clusters)	(**1.1796**)	(0.3770)	(**0.2638**)	(0.6017)				
s.e. (2-way clusters)	(1.5670)	(0.4218)	(0.2264)	(0.6231)				
Overinvest. F-stat (p value)	**5.44 (p=0.02)****	2.32 (p = 0.128)	0.19 (p = 0.660)	1.61 (p = 0.205)				
(2-way clustered s.e.'s)	3.33 (p = 0.106)	1.73 (p = 0.225)	0.26 (p = 0.625)	1.72 (p = 0.226)				
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					
Pilot*During	-0.0119	-0.0106	-0.0236					
s.e. (firm clusters)	(0.006)	(0.006)*	(0.020)					
s.e. (2-way clusters)	**[0.005]****	**[0.004]****	[0.015]					

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)						
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
Pilot*Crisis	-0.0004	-0.0031	n.m.	-1.7797	-0.0021	-0.0237
s.e.	(0.0013)	(0.0041)		(1.4352)	(0.0018)	(0.0196)

Table 3. Reported Results

All panels. Table summarizes reported results. 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.

	(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)*	(1.1842)
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					
Pilot*During	**-0.011****	-	**-0.039****					
s.e. (2-way clusters)	**[0.004]**	-	**[0.012]**					
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. I	Mkt Lev I	Book Lev. II	Mkt Lev II	NPL/Loans	NPL/Equity		
Pilot*Crisis			**0.6904*****	**3.3889****	**0.0094*****	**0.0857*****		
s.e. (firm clusters)			**(0.3124)**	**(1.4995)**	**(0.0021)**	**(0.0237)**		

Table 4 (FHK) Accruals: Moving to FHK Best-Match and Exact Specifications

Regressions of indicated accruals measures on Pilot*During, Pilot*Post, and constant term. Regressions are similar to Table IA-9, Panel A, except we progressively switch to the FHK best-match specification (Table IA-9, Panel B). **Panel A** reproduces results for our specification from Table IA-9, Panel A, including firm and fiscal year FE. **Panel B**, we switch to calendar year periods, over 2001-2010 (and thus to calendar year FE), define Pre as calendar 2001-2004, During as calendar 2005-2007, and Post as calendar 2008-2010, but include 2004 in the Pre period. **Panel C**. Remove calendar 2004 from Pre-period. **Panel D.** Switch to FHK best-match sample. Principal changes: use their definition of utilities, require data on FHK covariates. **Panel E.** Replace firm FE with pilot dummy. In **Panel F**, we replace calendar year FE with During and Post dummies (Pre is the omitted period), and report s.e.'s both clustered on firm and clustered on firm and calendar year. **Panel G**. Exclude operating accrual outliers based on $|accruals_t/assets_{t-1}| > 1.00$. before computing AA and PMDA, instead of winsorizing operating accrual outliers. **Panel H**. Find matching firm for PMDA using lagged ROA instead of current-year ROA. **Panel I**. Switch to FHK exact sample, obtain PMDA and AA from Compustat. **Panel J**. Use FHK exact specification: Impose FHK matching errors (i.e., use FHK posted 2012 PMDA data); winsorize by fiscal year instead of across years; also report s.e.'s with two-way clustering on firm and fiscal year. For Operating Accruals, Total Accruals, and AA, there is no FHK 2012 PMDA Specification to follow, so we used the FHK 2019 PMDA Specification. **All panels**. Balanced panel requires firms to have data to calculate accruals for each year in sample period. Coefficients on Pilot*Post and constant term are suppressed. Sample size shown only when different from previous panel. Standard errors clustered on firm are in brackets. Two-way clustering is on firm and *calendar year*, except as indicated in Panel I. *, **, *** indicates significance at the 10%, 5%, and 1% level, respectively.

	Unbalanced Panel				**Balanced Panel**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Accruals type	**Operating**	**Total**	**AA**	**PMDA**	**Operating**	**Total**	**AA**	**PMDA**
Panel A. Our Specification (Table IA-9 Panel A)								
Pilot*During	-0.0024	-0.0000	-0.0020	0.0029	-0.0037	0.0035	-0.0030	0.0075
s.e. cluster on firm	[0.0035]	[0.0076]	[0.0050]	[0.0080]	[0.0036]	[0.0079]	[0.0054]	[0.0087]
Pilot (Control) Firms	702 (1,413)	702 (1,413)	698 (1,401)	698 (1,401)	517 (947)	517 (947)	492 (906)	492 (906)
Panel B: Switch to Calendar Year Periods								
Pilot*During	-0.0025	-0.0039	-0.0050	-0.0043	-0.0055	-0.0007	-0.0070	-0.0004
s.e. cluster on firm	[0.0035]	[0.0074]	[0.0050]	[0.0080]	[0.0036]	[0.0077]	[0.0055]	[0.0089]
Pilot (Control) Firms	702 (1,413)	702 (1,413)	698 (1,401)	698 (1,401)	513 (934)	513 (934)	487 (896)	487 (896)
Panel C: Remove 2004 from Pre Period								
Pilot*During	-0.0044	-0.0016	-0.0040	-0.0022	-0.0065	0.0022	-0.0064	-0.0010
s.e. cluster on firm	[0.0040]	[0.0079]	[0.0055]	[0.0086]	[0.0043]	[0.0085]	[0.0061]	[0.0095]
Panel D: Switch to FHK Best Match Sample								
Pilot*During	-0.0046	-0.0001	-0.0040	0.0014	-0.0069	0.0036	-0.0074	-0.0024
s.e. cluster on firm	[0.0039]	[0.0077]	[0.0055]	[0.0086]	[0.0042]*	[0.0082]	[0.0061]	[0.0095]
Pilot (Control) Firms	701 (1,412)	701 (1,412)	697 (1,400)	697 (1,400)	489 (880)	489 (880)	466 (846)	466 (846)
Panel E: Remove Firm FE								
Pilot*During	-0.0048	-0.0021	-0.0021	0.0034	-0.0067	0.0041	-0.0065	-0.0024
s.e. cluster on firm	[0.0039]	[0.0077]	[0.0054]	[0.0084]	[0.0042]	[0.0082]	[0.0062]	[0.0095]
Panel F: FHK Best Match Specification (Remove Calendar Year FE)								
Pilot*During	-0.0047	-0.0033	-0.0030	0.0038	-0.0068	0.0027	-0.0073	-0.0022
s.e. cluster on firm	[0.0039]	[0.0077]	[0.0054]	[0.0085]	[0.0042]	[0.0082]	[0.0062]	[0.0096]

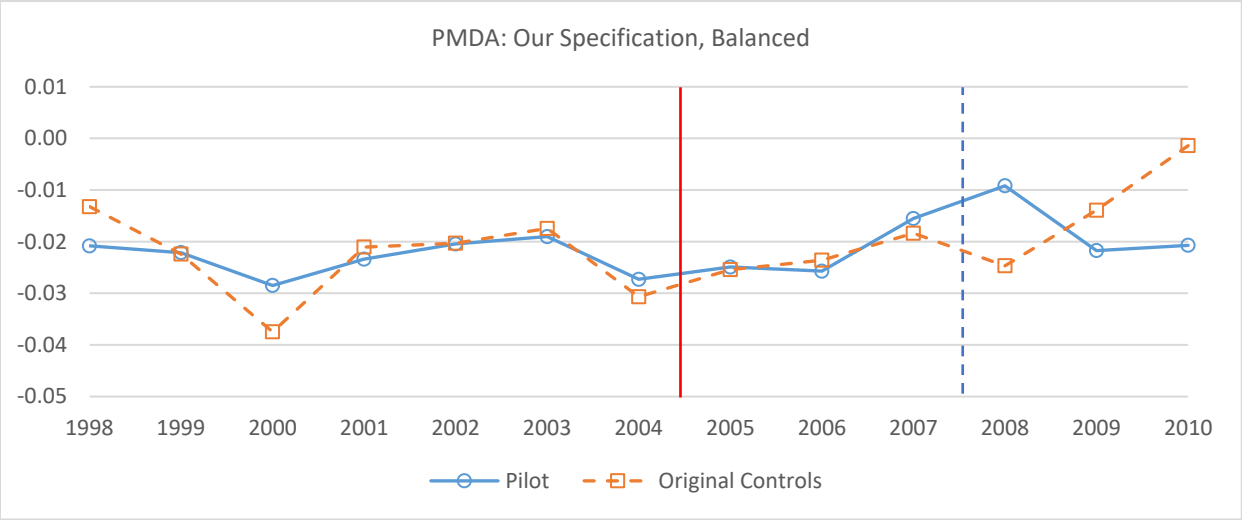
	Unbalanced Panel				Balanced Panel			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Accruals type	**Operating**	**Total**	**AA**	**PMDA**	**Operating**	**Total**	**AA**	**PMDA**
s.e. [2-way clustering]	[0.0034]	[0.0064]	[0.0045]	[.]	[0.0048]	[0.0094]	[0.0064]	[0.0030]
Panel G. Exclude Operating Accrual Outliers instead of Winsorizing								
Pilot*During	-0.0051	-0.0048	-0.0031	-0.0002	-0.0073	0.0005	-0.0053	-0.0040
s.e. cluster on firm	[0.0037]	[0.0074]	[0.0036]	[0.0059]	[0.0039]*	[0.0077]	[0.0039]	[0.0064]
s.e. [2-way clustering]	[0.0034]	[0.0068]	[0.0038]	[0.0044]	[0.0043]*	[0.0093]	[0.0039]	[0.0062]
Pilot (Control) Firms	701 (1,412)	701 (1,412)	697 (1,400)	697 (1,400)	478 (861)	478 (861)	451 (824)	451 (824)
Panel H: Find PMDA Matching Firm Using Lagged ROA								
Pilot*During	-0.0051	-0.0048	-0.0031	-0.0036	-0.0073	0.0005	-0.0053	-0.0092
s.e. cluster on firm	[0.0037]	[0.0074]	[0.0036]	[0.0059]	[0.0039]*	[0.0077]	[0.0039]	[0.0065]
s.e. [2-way clustering]	[0.0034]	[0.0068]	[0.0038]	[0.0049]	[0.0043]*	[0.0093]	[0.0039]	[0.0065]
Pilot (Control) Firms	701 (1,412)	701 (1,412)	697 (1,400)	697 (1,400)	478 (861)	478 (861)	451 (824)	451 (824)
Panel I: Use FHK exact sample, obtain PMDA and AA from Compustat								
Pilot*During	-0.0031	-0.0010	-0.0018	-0.0077	-0.0033	0.0049	-0.0020	-0.0114
s.e. cluster on firm	[0.0037]	[0.0077]	[0.0036]	[0.0058]	[0.0043]	[0.0089]	[0.0042]	[0.0070]
s.e. [2-way clustering]	[0.0035]	[0.0072]	[0.0041]	[0.0042]*	[0.0021]	[0.0083]	[0.0030]	[0.0062]*
Pilot (Control) Firms	730 (1,494)	730 (1,494)	730 (1,494)	730 (1,494)	388 (709)	388 (709)	388 (709)	388 (709)
Panel J. FHK Exact Specification; Impose FHK Matching Error, Winsorize PMDA by Fiscal Year, cluster on firm and *fiscal* year								
Pilot*During	-0.0028	-0.0020	-0.0017	-0.0069	-0.0027	0.0045	-0.0015	-0.0095
s.e. cluster on firm	[0.0038]	[0.0077]	[0.0036]	[0.0058]	[0.0042]	[0.0088]	[0.0041]	[0.0070]
s.e. [2-way clustering]	[0.0032]	[0.0067]	[0.0039]	[0.0040]*	[0.0019]	[0.0079]	[0.0031]	[0.0059]
s.e. cluster firm, *fiscal year*	[0.0028]	[0.0069]	[0.0030]	[0.0045]	[0.0022]	[0.0085]	[0.0028]	**[0.0040]****
Firm-Year Obs.	16,515	16,515	16,515	16,515	9,873	9,873	9,873	9,873
Pilot (Control) Firms	730 (1,494)	730 (1,494)	730 (1,494)	730 (1,494)	388 (709)	388 (709)	388 (709)	388 (709)

Figure 1. Annual Means for Selected Outcomes: Our Specifications

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; uses CFW Adjusted Best-Match Specification. **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.

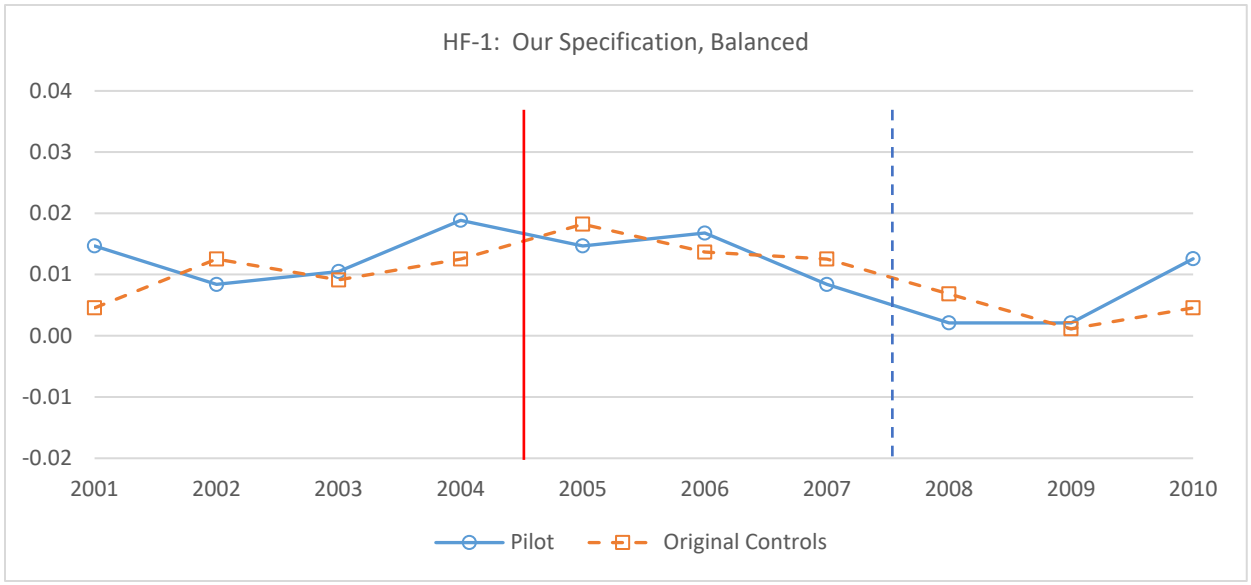
Panel A. FHK Accruals (PMDA, Balanced Panel)

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



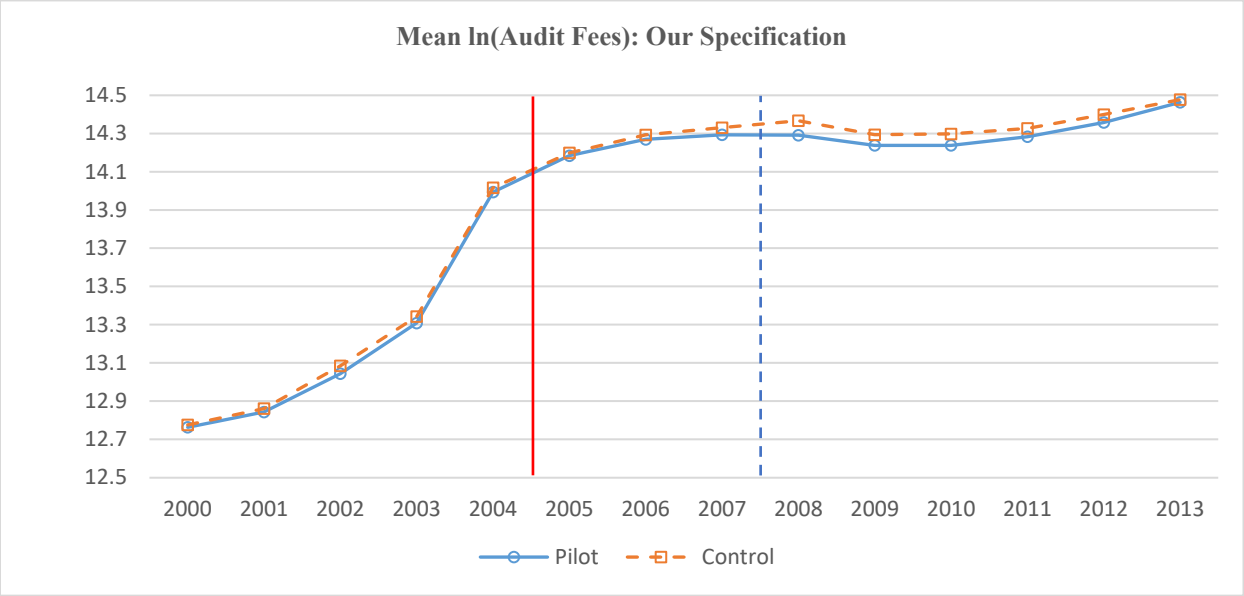
Panel B. FHK HF-Score (HF-1, Balanced Panel).

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


HF-1: Our Specification, Balanced
0.04
0.03
0.02
0.01
0.00
-0.01
-0.02
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Pilot
Original Controls

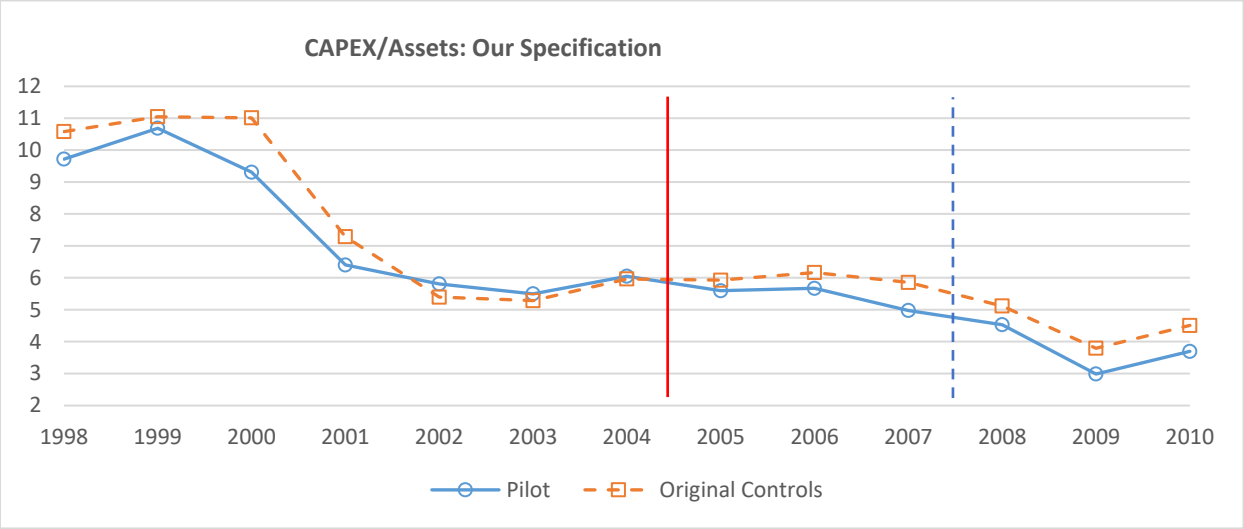
Panel C. HHZ Audit Fees

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



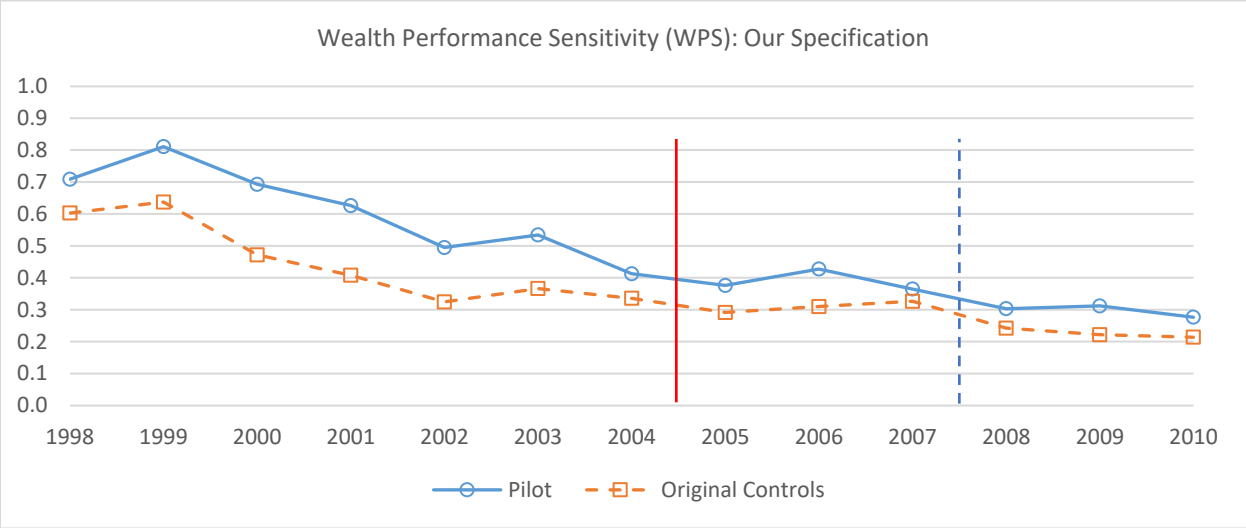
Panel D. GMW Capex/Assets

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



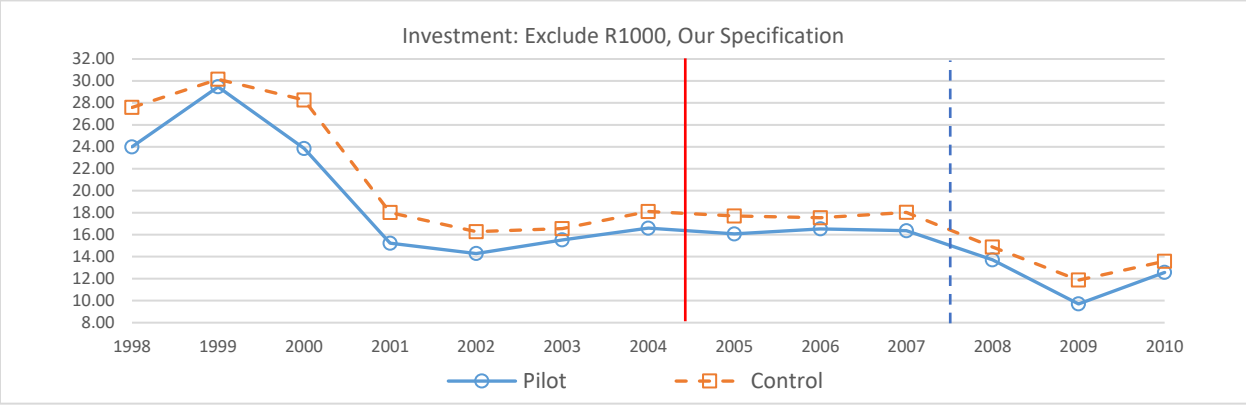
Panel E. LLS CEO Wealth Performance Sensitivity

Interpretation: No evidence for a treatment effect. Non-parallel trends in Pre period (2001-2004).



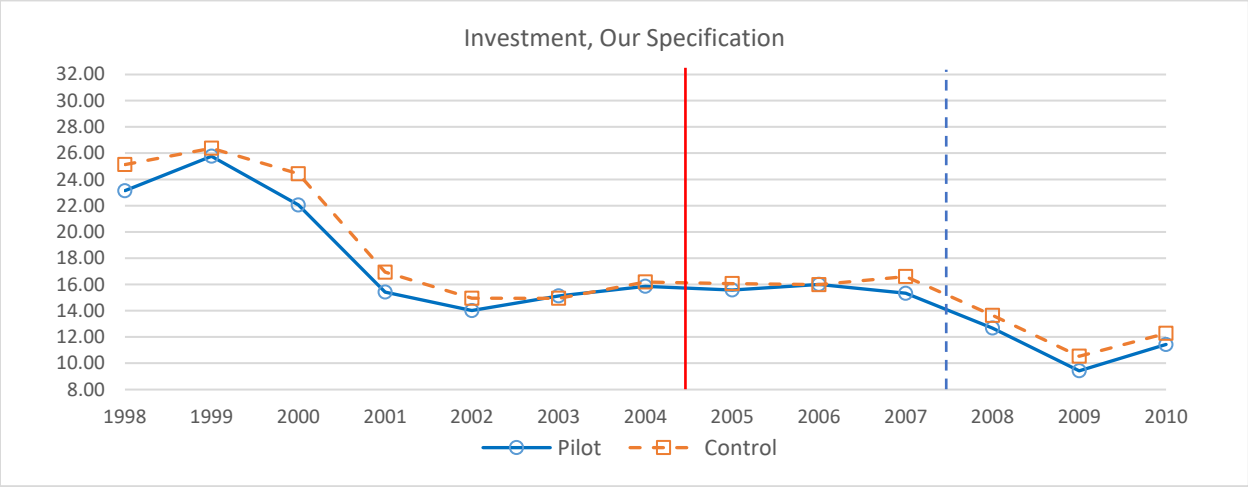
Panel F. CFW Under investment

Interpretation: No evidence for a treatment effect. Non-parallel trends in Pre period (2001-2004). No reversal in Post period.



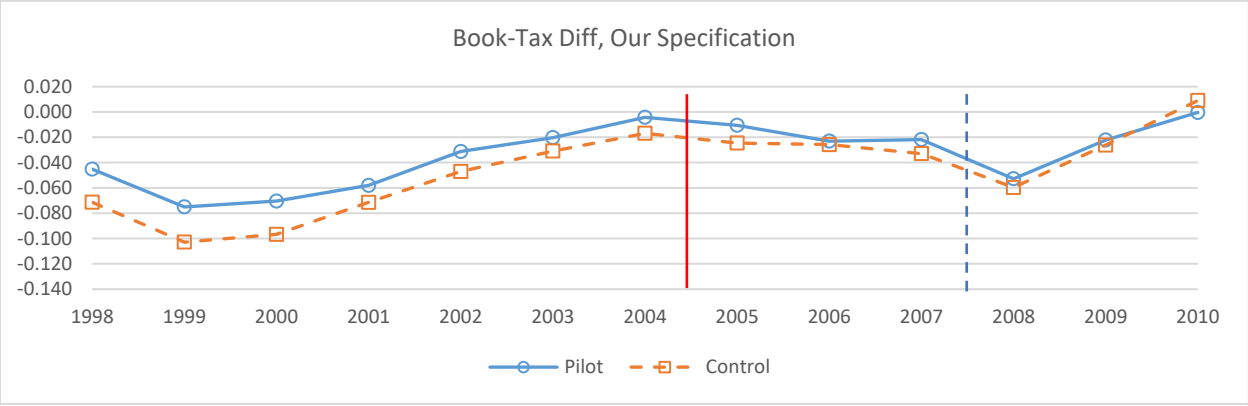
Panel G. TWX Underinvestment

Interpretation: Similar to CFW. No evidence for a treatment effect. Non-parallel trends in Pre period (2001-2004). No reversal in Post period..



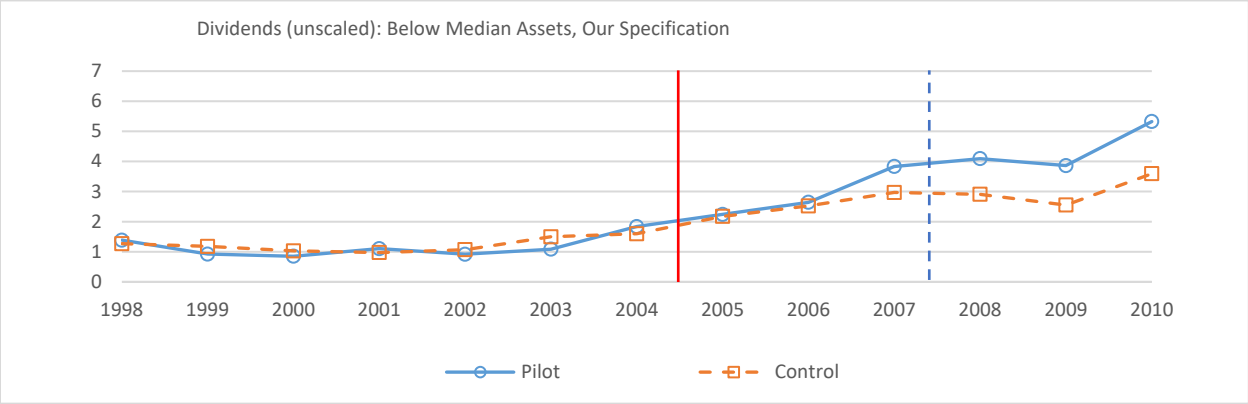
Panel H. KLP Book-Tax Difference

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



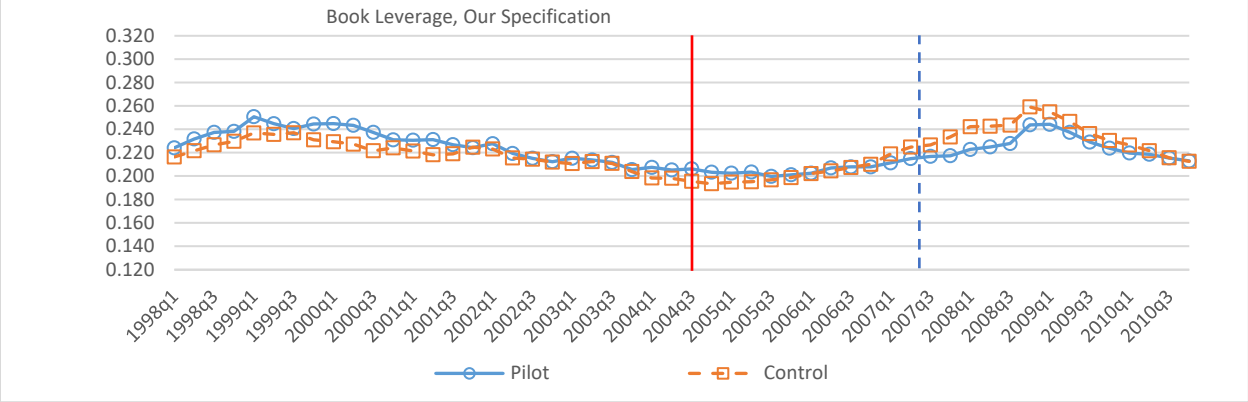
Panel I. CZC Dividends (unscaled)

Interpretation: Evidence for a treatment effect only in 2007. Larger treatment-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.

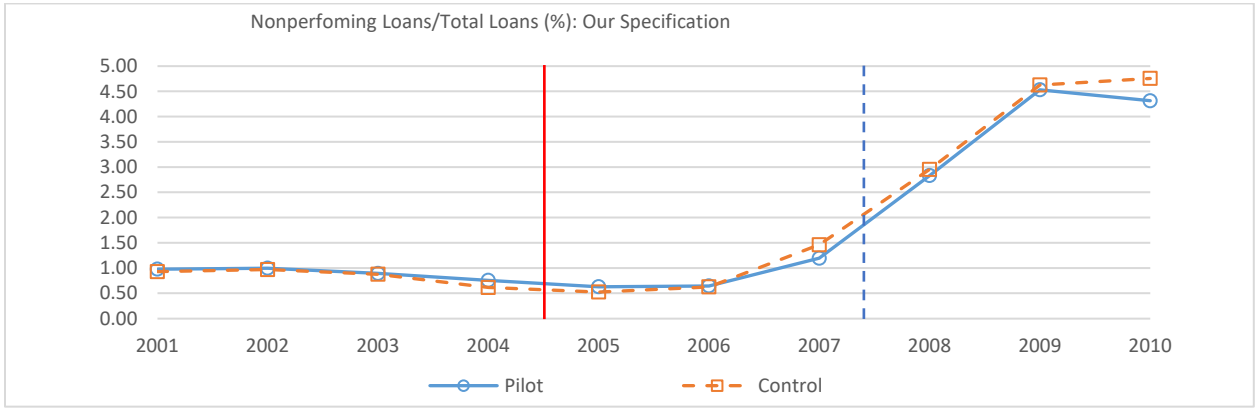


Figure 2. Statistical Significance Across Specifications

Figure reports *t*-statistics for our specification, best-match specifications (using one-way clustering) and reported results, for all 29 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 effect *sign* is opposite from the reported sign. For CFW overinvestment, we report a *t*-statistic with the same p-value as the reported value. For TWX overinvestment, we assume $t = F^{0.5}$. For BHLN book leverage II, we report the coefficient without covariates. Vertical lines at ± 1.96 indicate 5% significance**.**

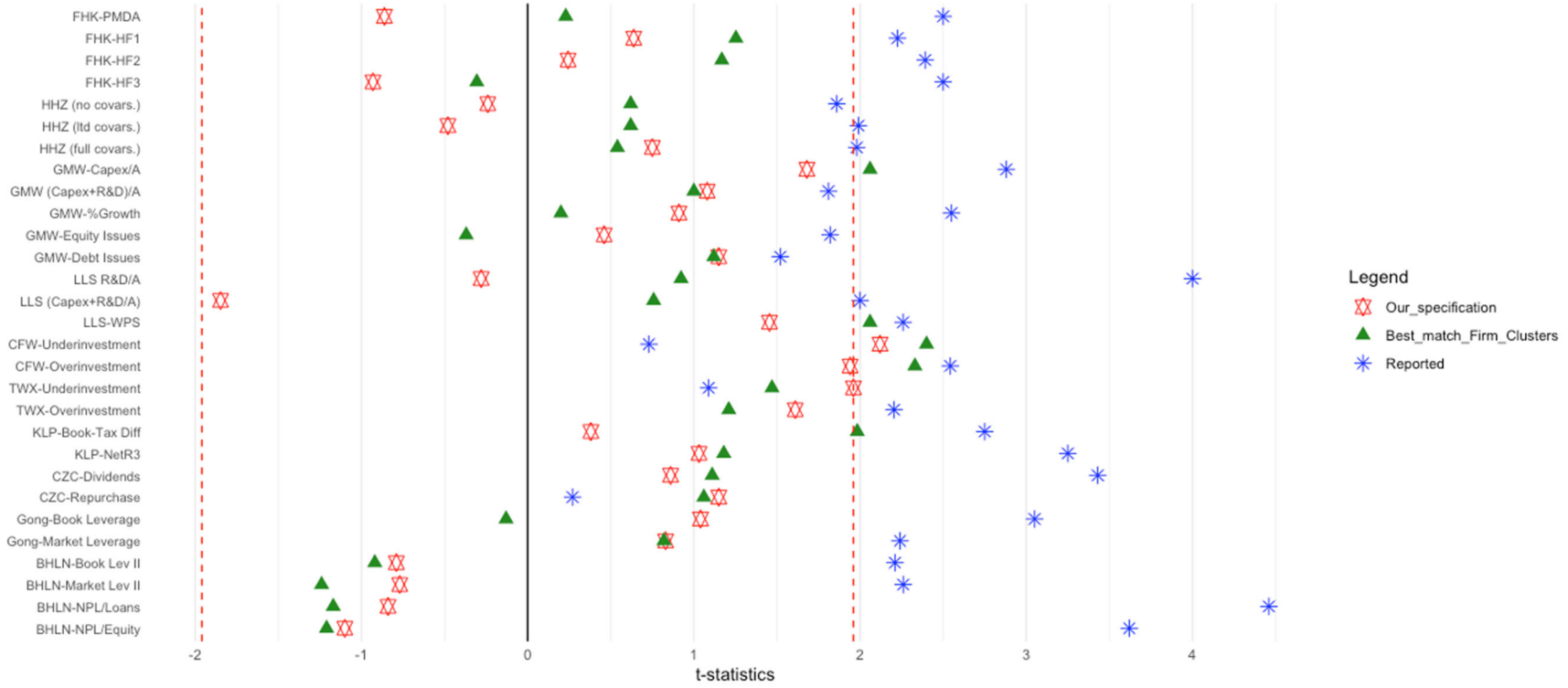


Figure 3. Comparing Coefficients Across Specifications

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, each scaled to coefficients 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.

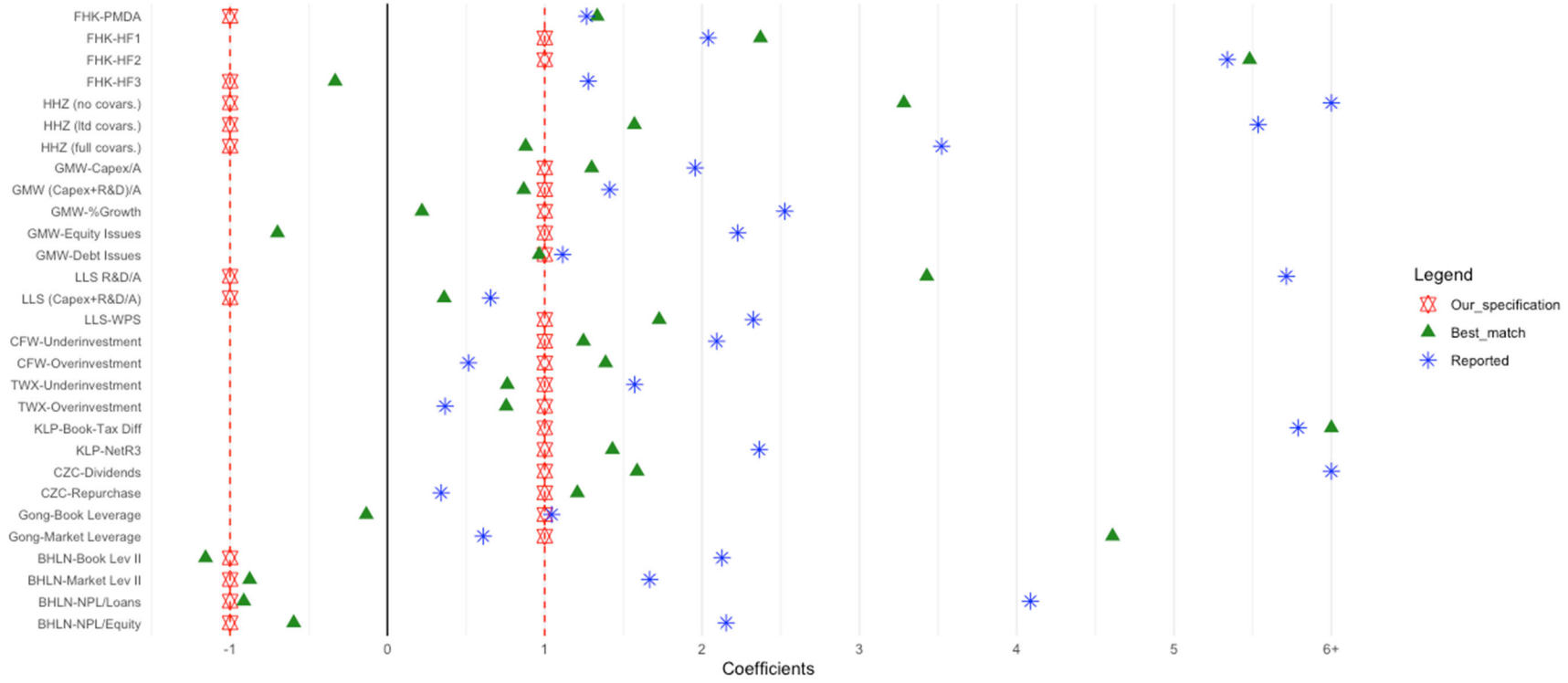


Figure 4. Comparing Standard Errors: Our Specification v. Best Match

Figure shows, for indicated 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). 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. For CFW underinvestment and TWX underinvestment, see Figure 2 for *t*-statistics; we report here the corresponding s.e.'s.

.

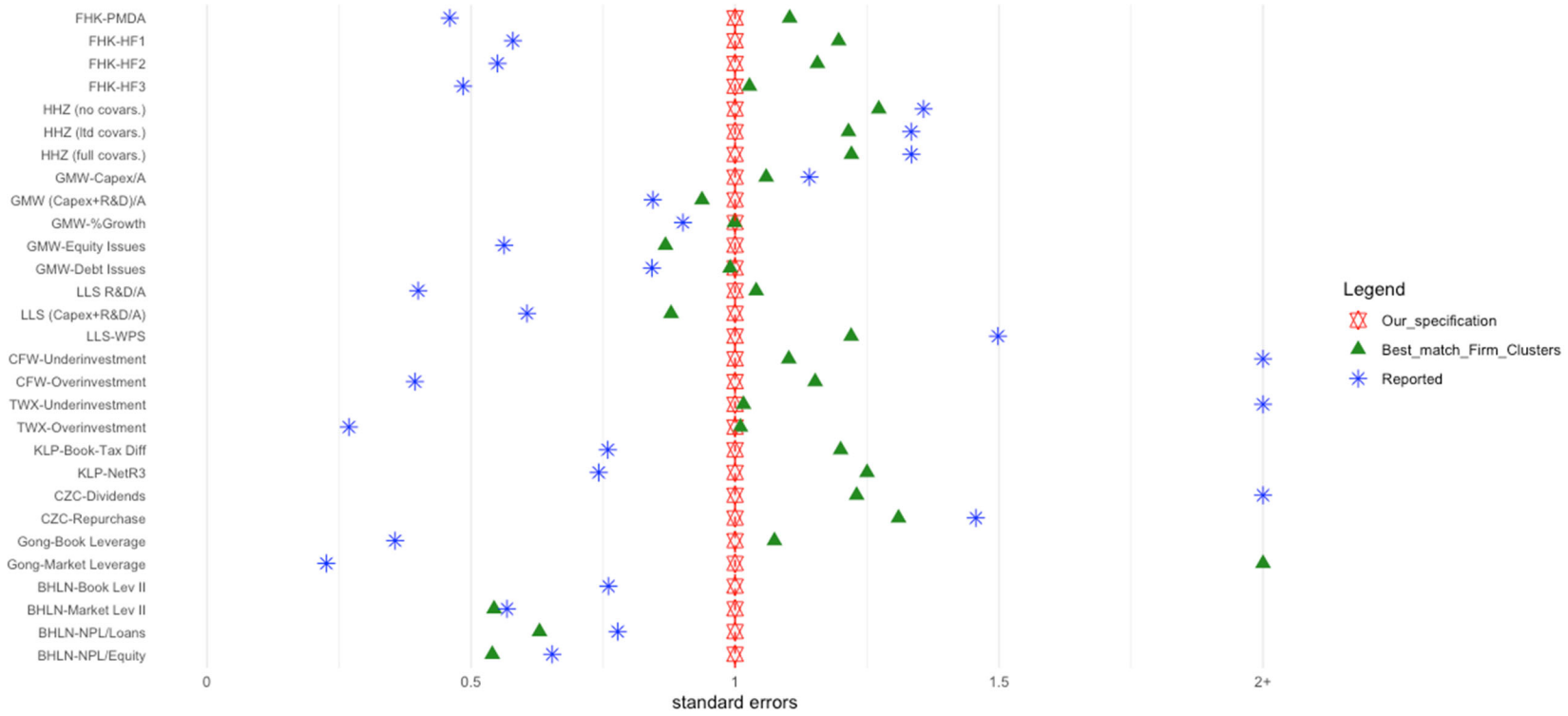


Figure 5. Comparing Standard Errors: One-Way versus Two-Way Clustering

Figure shows, for indicated papers using two-way clustering, and indicated outcomes, best match s.e.'s (with s.e.'s clustered on firm), scaled to 1.00; and best match s.e.'s with two way clustering on firm and year, scaled relative to s.e.'s clustered on firm. Two-way clustering is on firm and year for all studies except Gong (who uses quarterly data). FHK organized their data by calendar year but clustered on firm and *fiscal* year (determined from their exact code), without saying so in their paper. The FHK Best-Match Specifications assume clustering on firm and calendar year.

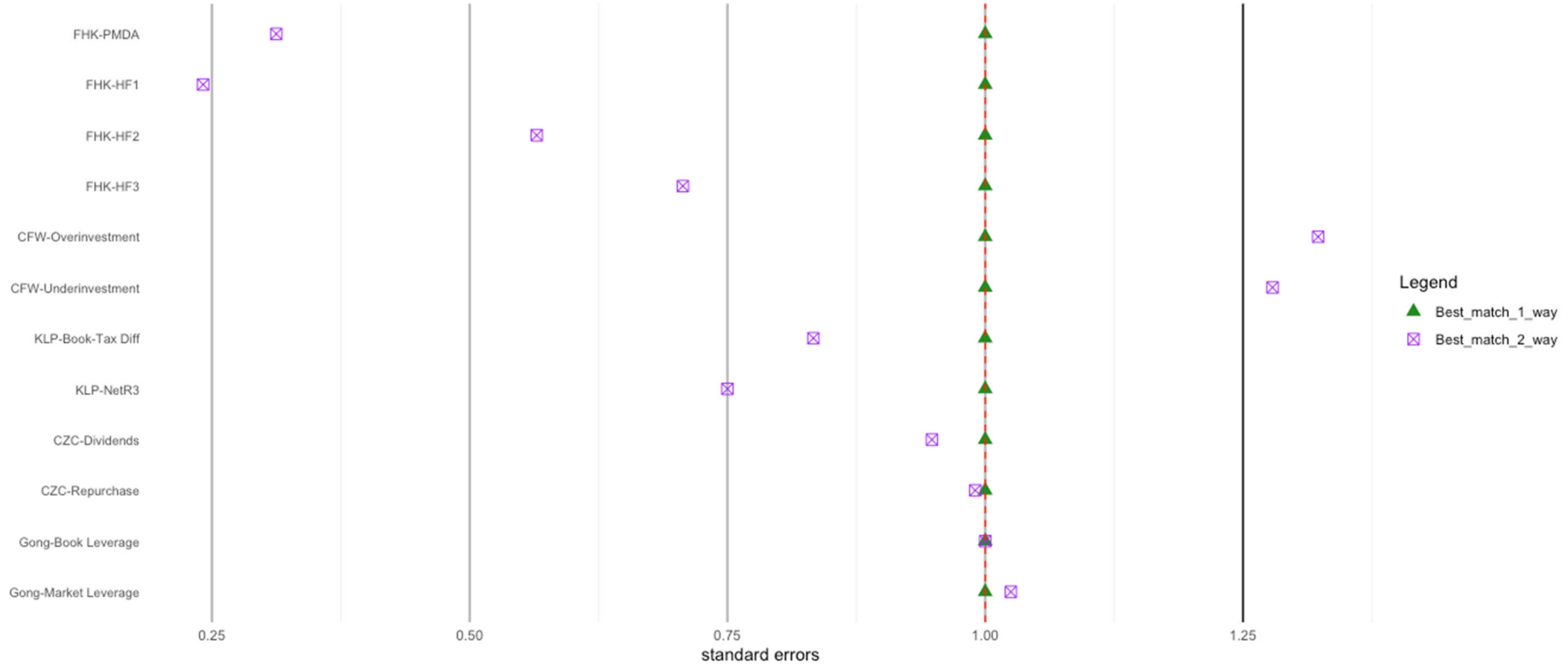
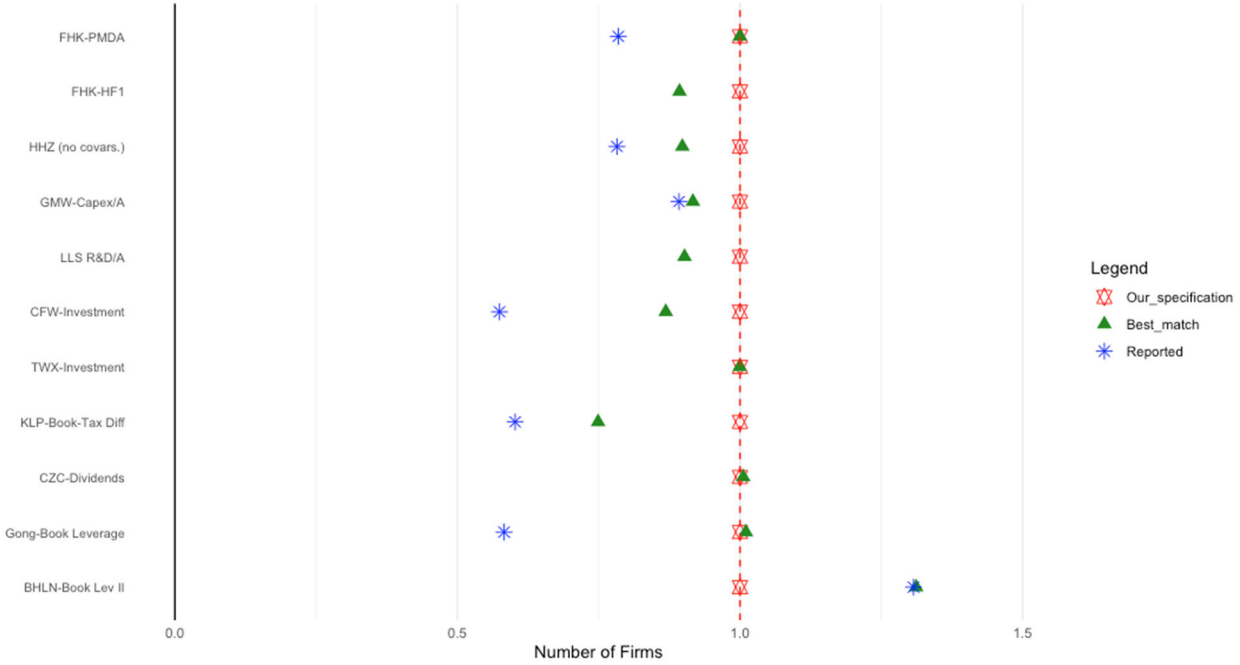


Figure 6. Comparing Sample Sizes

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.

Panel A. Number of Firms



Panel B. Number of Firm-Year Observations