# The SEC's Short-Sale Experiment: Evidence on Causal Channels and on the Importance of 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[1]
University of Arizona, Dhaliwal-Reidy School of Accountancy

(draft January 2022)

Northwestern University, Pritzker School of Law, Law and Economics Working Paper Series # 20-06

European Corporate Governance Institute
Finance Working Paper No. 813/2022



[1] We thank Parth Lalkiya and Lauren Fiotakis for research assistance. We thank Gauri Bhat, Doug Hanna, Ankit Jain (discussant), Pab Jotikasthira, Sanjay Kallapur, Umang Khetan (discussant), Srini Krishnamurthy, Charles Lee, Dan Millimet, Shiva Rajgopal, Mehrdad Samadi, Shyam Sundar, Jake Thomas, Sorabh Tomar, Ram Venkataraman and workshop participants at Bar Ilan University, Carnegie Mellon, Columbia Business School, Haifa University, Hebrew University of Jerusalem, Indian School of Business, Southern Methodist University, Stanford Law School, Tel Aviv University and the University of Texas at Arlington and participants at AELA, Midwest Finance Association, FMA Annual Meetings, for comments and discussions on an earlier draft.

http://ssrn.com/abstract=3657200

# The SEC's Short-Sale Experiment: Evidence on Causal Channels and on the Importance of Specification Choice in Randomized and Natural Experiments

**Abstract:** During 2005-2007, the Securities and Exchange Commission (SEC) conducted a randomized trial in which it removed short-sale restrictions from one-third of the Russell 3000 firms (pilot firms). Early studies found modest market microstructure effects of removing the restrictions but no effect on short interest, pilot firm returns, or price efficiency. More recently, many studies have attributed a wide range of indirect outcomes to this experiment, mostly without assessing the causal channels for those outcomes. We examine the three most often cited causal channels for these indirect effects: short interest, share returns and managerial fear. We find no evidence to support any of these channels. We then reexamine the principal findings in four recent studies using a pre-specified research design (similar across the four reexaminations) and a larger sample that closely matches the actual experiment, and find no support for the reported outcomes in any of these papers. We then switch to best-match specifications that closely match the samples and specifications reported in each paper, and still find only minimal support for the reported results. For two papers, we have the authors' original data and code; the reported results technically replicate but are highly fragile. Our findings highlight the importance of confirming a causal channel in randomized trials or natural experiments as well as the importance of sample selection and other aspects of specification choice for the statistical significance of reported results.

**Keywords:** natural experiments; causal channels; specification choice; Regulation SHO; SEC experiment

**The SEC's Short-Sale Experiment: Evidence on Causal Channels and on the Importance of Specification Choice in Randomized and Natural Experiments**

"A fragile inference is not worth taking seriously" (Leamer, AER, 1985)

## I. Introduction

In July 2004, the SEC announced a randomized trial (contained in Reg SHO) in which it suspended short-sale restrictions (price tests) for one-third of the firms ("pilot" firms) in the Russell 3000 Index (R3000), that traded on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), or the Nasdaq national market (Nasdaq). Specifically, for the pilot firms, the SEC suspended the uptick rule for the NYSE and AMEX firms and the similar but less restrictive bid test for Nasdaq firms during a roughly two-year period (May 2, 2005 through July 5, 2007). It left some but not all of the prior restrictions in place for the remaining, roughly 2,000 firms ("controls"). The price tests essentially ensured that short sale orders could not be executed at a price below the last trade The SEC's objective in conducting the experiment was to study the effects of relatively unrestricted short selling on market volatility, price efficiency, and liquidity and to monitor trading behavior (SEC, 2004a and SEC 2004b).[1]

It is unclear that these rules constrained fundamentals or valuation based short selling (Barclay, 1989). While, the uptick rule slightly delayed execution of short sales on NYSE, this should matter little to value-based short sellers, who typically build positions over time and hold them for long periods. Moreover, several developments over the years significantly diminished the effectiveness of the price tests in constraining valuation based short selling. First, the progressive reductions in tick size, from $0.125 to a penny by the time of the experiment, further reduced any effect of the uptick rule (Alexander and Peterson, 2002). Moreover, the Nasdaq bid test was weaker than the NYSE uptick rule, so the impact of the SEC experiment on Nasdaq firms was expected to

[1] Under NYSE Rule 440B, in effect prior to the experiment, a short sale was permitted only following a plus tick or a zero-plus tick ("uptick rule"). Under Nasdaq Rule 3350, short sales in National Market Securities had to be above the current bid if that bid was below the previous bid ("bid test"). Below, we generally refer to both rules as the "uptick rule." Where we want to distinguish between them, we call them the "NYSE uptick rule" and the "Nasdaq bid test." The American Stock Exchange (AMEX) rule was similar to the NYSE. We generally refer to both NYSE and AMEX firms as NYSE firms.

be less than for NYSE firms. For example, Christophe, Ferri and Angel (2004) find that the Nasdaq bid test had little impact on short trade execution; and based on studying how the experiment affected Nasdaq firms, Alexander and Peterson (2008, p. 86) conclude that "the bid test is relatively inconsequential."

Second, regional exchanges, which traded the shares of larger NYSE firms, did not impose the uptick rule, and electronic exchanges, which accounted for around 40% of trading volume in Nasdaq firms, did not impose the bid test. This further weakened any effect for the firms traded on these venues. Many larger firms also had traded options, which are an alternate way to express bearish sentiment, with low incremental cost compared to direct short sales (Battalio and Schultz, 2006).

Third, in addition to fully suspending price tests for pilot firms, the SEC also suspended the NYSE rule for control firms in the Russell 1000 (below, R1000) for after-hours trading (from 4:15 p.m. until the opening on the next trading day). The SEC also suspended the NYSE uptick rule for all firms at times when the consolidated transaction reporting system was off (generally from 8:00 p.m. until 4:00 a.m. the next day), and the Nasdaq bid test always only applied during limited hours (generally 9:30 a.m. to 4:00 p.m.). These factors reduced the differences in the rules that applied to pilot versus control firms.

Consistent with the arguments above, initial studies of the experiment found little to no direct impact of removing short-sale restrictions on short interest, share returns and volatility (OEA, 2007; Alexander and Peterson, 2008; Diether, Lee and Werner, 2009). However, as expected, these studies did find some improvement in execution of short trades, especially for NYSE firms. For example, Diether et al. (2009) and Alexander and Peterson (2008) find improved execution of short sales for NYSE firms resulting in smaller trade size and higher short volume. For Nasdaq firms, consistent with the bid test being less restrictive than the NYSE uptick rule, the suspension of the bid test had limited impact on pilot firms. Based on these findings, the SEC in 2007 removed these restrictions for all firms.

Despite little evidence of *direct* impact of the Reg SHO experiment on pilot firms, over 60 papers in accounting, finance, and economics report that suspension of the price tests had wide

ranging *indirect* effects on pilot firms, including on earnings management, investments, leverage, acquisitions, management compensation, workplace safety, and more (see Internet Appendix, Table IA-1 for a summary). Some of these papers find that the Reg SHO experiment affected behavior of third parties such as auditors and analysts.

The broad range of indirect effects attributed to the Reg SHO experiment is surprising. First, since the uptick rule and bid tests did not meaningfully constrain short selling, one would not expect lifting these restrictions to generate wide ranging indirect effects. This increases the risk that indirect effects may be false positives (Harvey, 2017). Second, to credibly attribute indirect effects to the Reg SHO experiment, there should be evidence supporting a causal channel through which removing short-sale constraints could generate those indirect effects. However, prior work found no reliable evidence that removing the price tests affected short interest, share returns, price efficiency, volatility, etc. Without a clear causal channel, indirect effects are more likely to be false positives.

We study the three causal channels most commonly cited in the indirect effects literature, through which the Reg SHO experiment could have affected the behavior of firms or third parties: (i) short interest, (ii) returns and (iii) manager's fear of bear raids. For a fourth channel, price efficiency, we rely on prior work finding no evidence for this channel (Alexander and Petersen, 2008; Diether et al., 2009) or the speed with which share prices respond to negative information (Bai, 2008).

For short interest and share price channels, prior work found no support for these channels. However, Grullon, Michenaud and Weston (2015, below, GMW) report evidence for an effect on short interest between experiment announcement and launch (a period not previously studied), and a drop in share prices for small pilot firms (defined as below median in total assets) before experiment announcement (again, a period not previously studied). We reexamine that evidence using a more comprehensive sample which closely matches the actual pilot and control firms, and a longer period. For short interest, we find no support for the GMW finding of rising short interest between experiment announcement and launch and explain why their measure (cumulative abnormal short interest) is misspecified. Also, prior work studies only mean short interest for pilot

versus control firms. We extend that work by examining the distribution of short interest across firms. We again find no evidence of difference between pilot and control firms. For share returns, the GMW finding is not robust to sample choice, including use of the full experiment sample or defining "small" pilot firms based on market capitalization or trading volume instead of assets. There is also no effect, even for their sample, for NYSE firms – for whom any effect should be stronger, as discussed above. We also discuss why a pre-announcement effect, when the list of pilot firms was unknown, is implausible.

The third, commonly cited channel is manager fear: Even if the Reg SHO experiment did not actually affect short interest or returns, pilot firm managers could have feared being targeted by short sellers and taken pre-emptive actions (Fang, Huang, and Karpoff, 2016, below, FHK). We cannot directly study this channel, but we can assess its plausibility. If firm managers were fearful that relaxing the price tests would affect them, one might expect them to voice concerns in various ways: speaking with business news reporters; writing to the SEC when it sought public comments; seeking meetings with SEC officials to express opposition. For example, when the Financial Accounting Standards Board proposed expensing of employee stock options in 1993, there was major opposition to this proposed rule from firm managers, including appeals to Congress, and significant media coverage of manager opposition. In contrast, the short-sale experiment generated very little publicity. We found no evidence of manager opposition when the rule was proposed in 2003, when it was announced in 2004, or when the SEC proposed to abolish the short-sale rule in 2006. One might also expect any fear to shrink in 2006 and 2007, once it became apparent that the bears were not charging at pilot firms, and thus any indirect effects should shrink in magnitude.

The weak evidence supporting the principal causal channels asserted in the indirect effects studies reinforces the doubts noted above about whether the short-sale experiment meaningfully affected substantive short sellers. We therefore sought to assess whether skepticism is warranted about the robustness of the wide range of indirect effects reported in the literature. To keep the task reasonable, we reexamine four studies – as many, we believe, as any one project can tackle with sufficient care. We chose these studies for several reasons. First, all are published in top

journals, and thus are likely to represent the best of this literature. Second, they rely on data from standard sources, which eased the task of reexamination. Third, they rely on different causal channels. Two of the studies (FHK and Hope, Hu, Zhao, 2017, below HHZ) were familiar to us as three of us served as discussants of these papers. We chose GMW (2015) because they reported evidence supporting causal channels through short interest and share price, in contrast to prior studies. We later added Lin, Liu, and Sun (2019; below, LLS) because of its publication in the *American Economic Review* and because they study outcomes similar to those in GMW.

FHK rely on the manager fear channel. They conjecture that in response to a greater threat of short selling, pilot firms' managers reduced earnings management to preemptively deter short sellers. Their measure of earnings management is performance-matched discretionary accruals (PMDA), a measure that compares the accruals of pilot and control firms to those of firms matched based on performance (return on assets, or ROA). They also report that pilot firms have a lower likelihood of having a very high "F-score" (a measure of the likelihood of an accounting misstatement).

GMW conjecture that short-sale constraints result in overvaluation of firms, which can induce overinvestment, and that removing these constraints will reduce share prices and thus investment. GMW report that small pilot firms reduced investment and raised less capital.

HHZ test for an effect of the short-sale experiment on the behavior of auditors. They conjecture that the experiment increased litigation risk for auditors because pilot firms would face higher risk of large share price drops, followed by securities litigation alleging auditing errors. The auditors would pre-emptively respond to increased litigation risk by increasing audit fees.

LLS rely on a price efficiency channel. They posit that share prices of pilot firms became more informative and that, with more informative prices to guide managerial actions, shareholders will perceive lower need for direct CEO performance incentives, so pilot firms will have lower sensitivity of CEO wealth to performance. They also hypothesize that managers will rely on the more informative prices to guide business decisions, which will make investment more sensitive to Tobin's *q*.

To assess the evidence on causal channels and then reexamine these studies, we first developed a pre-specified sample and research design (Black et al, 2019, below BDLYY, 2019). Our pre-specified analysis plan specified how we would approach the FHK, GMW, and HHZ research questions, if we wanted to test their conjectures. We added LLS later. We defined the pilot and control samples closely following the SEC's rules, defined pre-experiment ("Pre"), during-experiment ("During"), and post-experiment ("Post") periods consistently across the reexamined papers, specified regressions, and made other specification choices with the outcomes in these papers hidden, to ensure that our specification choices could not be affected by knowledge of how a particular specification would affect our results. With our prespecified design, we find no support for the principal conjectures in any of these four studies. Across 23 outcomes (13 from these papers plus 10 related outcomes that we study to assess robustness), none are statistically significant with the predicted sign.

We also developed "best-match" specifications for each paper, in which we did our best to match each paper's sample and design, based on the descriptions in each paper. All but two results (one for GMW, one for LLS) remain insignificant, and those two are fragile. The gaps between the best-match and reported results suggest that each paper makes important specification choices that our best-match approach could not capture. In response to this project, FHK posted data and code for their PMDA result (but not their HF-score results) and replied to an earlier draft of this project (Fang, Huang and Karpoff, 2019). HHZ also provided their data and code to us. Using the exact FHK and HHZ samples and specifications, we can technically replicate one of their results, but as we show below, both sets of results are fragile.[2]

Given our failure to find support for the causal channels, and our finding that the results in the four papers we reexamine are not robust, it becomes more likely that many other indirect-

[2] We discuss the technical replication briefly in Part V, and in more detail in the Internet Appendix. We have posted the Stata code and datasets needed to generate the results in this paper, and the SAS code we used to generate starting datasets, both our own and the best-match datasets, on our website. We posted our pre-analysis plan on SSRN, and indicate there and in the Internet Appendix the limited deviations from the original plan. These principally involved adding detail to our specification choices and adding LLS as a paper to reexamine.

effects results would also prove not to be robust, if closely examined. The Heath et al. (2020) critique of multiple hypothesis testing strengthens those concerns.[3]

The sensitivity to specification choice that we find for the four studies that we reexamine has implications for the credibility of other DiD studies of randomized trials and natural experiments. There is substantial work in other disciplines concerning how often results can be technically replicated or are robust to alternative specifications.[4] There has been less work in finance or accounting, as Harvey (2014, 2017, 2019), Welch (2019), and Hail, Lang, and Leuz (2020) have observed. Our results support the need for attention to these concerns.

This paper proceeds as follows. Section II provides evidence on causal channels. Section III summarizes our re-examination methodology. Section IV presents our re-examination results. Section V discusses technical replication of FHK and HHZ. Section VI discusses broader implications from our project. Section VII concludes. Given the scope of this project, which both examines causal channels and reexamines four papers, we relegate many supporting results to the Internet Appendix.

## II. Evidence on Causal Channels

### A. Importance of Causal Channel in Testing Indirect Effects of the SEC Experiment

We examine the three causal channels most often cited in the indirect effects literature: short interest, returns and managerial fear. Aside from the manager-fear channel, any indirect effects from the Reg SHO experiment should follow from a causal channel that begins with the direct effects of the experiment on substantive short selling. The most natural channel would involve an increase in substantive short selling, for which the best evidence would be an increase in short interest for pilot firms relative to control firms. Greater substantive short selling might, in turn, affect the share prices of pilot firms (a share price channel). Effects on short interest or price

[3] Our concerns with robustness involve papers reporting indirect effects of the SEC experiment on the behavior of managers, firms, and third parties such as auditors. They do not apply to market microstructure papers, which study direct effects of the experiment on trading markets.

[4] See, e.g., Ioannidis (2005), Kaplan and Irvin (2015); Olken (2015), Open Science Collaboration (2015).

might in turn affect manager behavior, or the behavior of third parties such as auditors and analysts. However, as noted earlier, the early studies of the experiment found no significant change in short interest or returns (Alexander and Peterson, 2008; Diether et al., 2009).

Many indirect-effects papers invoke greater short-selling as a causal channel, either without discussing the studies that found no increase in short interest or citing only evidence for an increase in short-sale trading volume. However, higher trading volume, without higher short interest, is consistent with more arbitrage trading, but does not imply increased substantive short-selling. Similarly, many indirect-effects papers invoke a share price channel, due to presumed greater substantive short selling, without showing evidence of either increased short interest or lower returns to pilot firms.[5]

We view evidence supporting a causal channel as central to the credibility of the indirect effects studies. A DiD design (the typical approach in most of these studies) with a presumed causal channel is similar to an instrumental variable (IV) design, in which the instrument is the shock (a firm being assigned to pilot status), and the instrumented variable is a measure of the channel (short interest or share returns).[6] A DiD design with a weak or absent causal channel is similar to an IV design with a weak first stage. Inference from IV with a weak first stage is unreliable due to the "weak instruments" problem (e.g., Angrist and Pischke, 2009, ch. 4.6). Stock, Wright and Yago (2002) propose $F > 10$ for multiple instruments ($t > 3$ for a single instrument), as a minimum threshold.

### B. Sample Selection

To assess whether the experiment affected returns around the time of announcement, or on change in short interest between announcement and experiment launch, the appropriate sample to

[5] GMW recognize the importance of providing evidence for a causal channel, and assert that the early studies did not look at the right periods to evaluate short interest or returns. They report that short interest increased before the experiment started and pilot firms' returns fell relative to control firms before the experiment was announced. We discuss their evidence below.

[6] Without covariates, the IV estimate (called a Wald estimate) is [(average treatment effect for all firms)/(proportion of complier firms)], Angrist, Imbens, and Rubin (1996). The numerator is the DiD estimate.

use is the full sample of pilot and control firms at the time of announcement, which we term the "2004 Announcement Sample." The SEC announced the short-sale experiment on July 28, 2004 and launched the experiment on May 2, 2005. It created separate lists of NYSE, AMEX, and Nasdaq national market firms included in the R3000 index as of June 25, 2004, and assigned one-third of the firms in each list to be treated, effectively at random.[7] The SEC's original list of pilot firms includes only 986 pilot firms (which the SEC called "Category A" firms) instead of 1,000 because the SEC excluded, apparently prior to randomization, firms that were either not listed on these exchanges as well as firms that became public after April 30, 2004.

To construct the 2004 Announcement Sample, we start with the R3000 list as of June 30, 2004, from Bloomberg, which maintains monthly historical lists of R3000 index firms. We merge this list with the CRSP monthly stock file for June 2004 and can match all 3,000 firms, including the 986 pilot firms. The SEC randomized the R3000 Index into pilot firms (one-third of the R3000), for which it suspended the uptick rule completely, and control firms (the remaining two-thirds of the R3000). Then, for control firms in the R1000, which the SEC termed "Category B", the SEC suspended the uptick rule *after trading hours*. The SEC never published a list of all control firms, only lists of the Category A (pilot) and Category B firms. To create a full list of control firms, we follow the SEC's exclusion rules and exclude 32 firms (all traded on the Nasdaq small cap market) that were not listed on NYSE/AMEX or Nasdaq national market, and 12 firms that began trading after April 30, 2004. We also exclude two firms that were delisted prior to the announcement date. This leaves 2,954 firms, consisting of 985 pilot firms (one initial pilot firm was delisted on June 28, 2004) and 1,969 control firms, as of July 28, 2004. We assess in the remainder of this part the evidence for the most commonly cited causal channels for indirect effects of the SEC experiment: short interest, share prices, and manager fear.

[7] The R3000 list is updated every year on the last Friday in June. The SEC adopting release, Securities Exchange Act Release 34-50104 (July 28, 2004), states that the SEC used the R3000 list as of June 25, 2004. The SEC conducted, in effect, a block randomized experiment, in which within each trading market (NYSE, AMEX, and Nasdaq national market), it ranked these firms by trading volume over June 2003 through May 2004, and chose every third firm (the 2nd, 5th, 8th, etc. in the within-market lists) to be treated.

## C. Causal Channel 1: Impact on Short Interest

To examine the impact of Reg SHO experiment on short interest, we focus on short interest over July 2003-December 2007. Data on short interest is available for the middle and end of each month. We use mid-month short interest as the outcome, treat July 2003 through July 2004 as the pre-announcement period; August 2004 through April 2005 as the pre-launch period; May 2005 through June 2007 as the experiment period; and July to December 2007 as the post experiment period.[8]

### *C.1. Graphical Evidence*

We start with graphical evidence. In Figure 1, we plot monthly short interest of pilot firms and control firms from July 2003 through December 2007. Panel A shows results for the 2004 Announcement Sample (2,954 firms). Panel B shows results using a narrower "2005 Analysis Sample" (the 2004 Announcement Sample updated to the start of the experiment period, and excluding financial and utility firms; 2,115 firms, see details below). Panel C shows results using the GMW Best-Match Sample, Vertical lines indicate the experiment start and end; a dotted line shows experiment announcement. For all three samples, there is no evidence of higher short-interest for pilot firms, during either the pre-launch period or the experiment period. The lack of an effect of the experiment on short interest is consistent with the price tests not meaningfully affecting substantive short selling.

GMW report evidence of an increase in short interest during the pre-launch period, which is stronger for small firms (below-median in assets). We therefore also report, in Panel D, a similar comparison for small firms within the GMW Best-Match Sample. Here too, there is no meaningful separation between pilot and control firms. It is also unclear to us why one would expect, on theoretical grounds, a rise in short interest before the experiment began.[9]

---

[8] The experiment was announced on July 28, 2004, so July 2004 mid-month short interest precedes the announcement. All trading days in May 2005 are in the experiment period. The experiment ended on July 5, 2007. Regression results for the experiment period are not sensitive to whether we treat July 2007 as part of the experiment period, coming after the experiment period, or drop this month.

[9] GMW provide a graph (their Figure 1) of "Cumulative Abnormal Short Interest" which shows a steady

*C.2. DiD Regression Results for Short Interest*

Next, we undertake a more formal analysis of whether the SEC experiment affected short interest. We use a DiD approach to estimate the causal effect of the experiment on pilot firms' short interest, during both the pre-launch and the experiment period.

$$y_{its} = \alpha_s + f_{is} + \lambda_{ts} + \gamma_s^a * Ann_t + \gamma_s^e * During_t + (\beta_s^a * P_i * Ann_t) + (\beta_s^e * P_i * During_t) + \epsilon_{its} \quad (1)$$

Here *y* is short interest as percentage of shares outstanding; $P_i$ is a pilot firm dummy (=1 for pilot firms); $Ann_t$ and $During_t$ are dummy variables for the pre-launch period (August 2004 through April 2005) and the experiment period (May 2005 through June 2007); *i* indexes firms, *t* indexes time in calendar months; *s* indicates the sample; the $\lambda_{ts}$ are month fixed effects, the $f_{is}$ are firm fixed effects. The non-interacted pilot dummy is absorbed by the firm fixed effects. Short interest is likely to be persistent within a given firm across time, so we cluster standard errors (s.e.'s) on firm. The coefficients of principal interest are on the interaction terms: $\beta_s^a$ for the pre-launch period, and $\beta_s^e$ for the experiment period.

In Table 3, we present DiD regressions, following eqn. (1), for the four samples shown in Figure 1: 2004 Announcement Sample, 2005 Analysis Sample, GMW Best-Match Sample, and small firms (based on assets) within the GMW Best-Match Sample. We also report results for small firms within the 2004 Announcement sample (based on market capitalization), small firms within the GMW Best-Match Sample (based on market capitalization); and small firms within the GMW Best-Match Sample (based on trading volume). We use market capitalization and trading volume as alternative measures of "small" because membership in the R3000 is based on market capitalization and the SEC ranked firms based on trading volume, when selecting pilot firms. We

increase during the pre-launch period, reaching a total of 4% of outstanding shares. However, this measure is misspecified. Studying *cumulative* abnormal short interest is akin to an event study showing a graph of cumulative *price*, rather than cumulative *returns*. It is also unclear why, starting with a randomized experiment, one should study abnormal short interest rather than the simpler measure of mean short interest. In the Internet Appendix (Figure IA-1), we present results for the GMW measure of abnormal short interest. This measure is noisy, with no evidence that it rises for pilot firms in either the pre-launch or the experiment period.

report results without covariates but in the Internet Appendix (Table IA-5), find similar results if we include the GMW covariates. For both the pre-launch and experiment periods, all coefficients on the interaction terms are small and statistically insignificant. We thus find no support for an increase in short interest for small pilot firms during the pre-launch period or the experiment period.[10]

*C.3. Distribution of Short Interest*

Although there is no evidence for higher mean short interest for pilot firms, the SEC experiment could still have allowed short-sellers to target some firms more heavily. We therefore also test for a difference in the distribution of short interest between pilot and control firms, and report results in the Internet Appendix (Figure IA-3). There is no visual evidence of a difference in the distribution of short interest between pilot and control firms during the pre-announcement period, the pre-launch period, or the experiment period. A Kolmogorov-Smirnov fails to reject the null of equal distributions of short interest for pilot and control firms.

In sum, there is no evidence that the Reg SHO experiment led to an increase in short interest for pilot firms, and thus no evidence for a causal channel that runs through short interest.

## D. Causal Channel II: Impact of Short-Selling Restrictions on Share Prices

We next examine evidence for a second possible channel, lower returns for pilot firms. Prior studies found no effect of the short-sale experiment on returns (Diether et al., 2009) or extreme price changes (OEA, 2007). However, GMW report small but statistically significant negative abnormal returns to small pilot firms (based in assets), relative to control firms during the two weeks *before* the experiment was *announced*. We therefore reconsider the evidence for an effect on returns, focusing on this period. We report "raw" rather than abnormal returns (as do GMW), because there should be only chance imbalances in market model β's between pilot and

[10] In the Internet Appendix (Figure IA-2), we report leads and lags graphs showing monthly estimated treatment effects during the sample period. Across the four samples, pre-announcement trends are reasonably parallel, as one would expect for a randomized experiment, there are no apparent trends in either the pre-launch or experiment periods, and all monthly coefficients are statistically insignificant.

control firms; we also confirm pre-experiment balance on β's (see Table 2, Panel B). Thus, a comparison of raw returns to pilot versus control firms should be unbiased.

*D.1. Graphical Evidence*

GMW report both graphical and regression evidence of modestly negative relative returns of small pilot firms, over a (-10, +1) window around the announcement date. We reconsider their evidence. In Figure 2, we report buy and hold relative returns (BHRRs; buy and hold raw returns for pilot firms relative to control firms) over June 1-Sept. 30, 2004 (the same time period as in GMW Figure 2).[11] As discussed earlier, since we are evaluating the impact of the announcement of the SEC experiment, the 2004 Announcement Sample is the preferred sample to use. We therefore present results for: (i) the 2004 Announcement Sample; and (ii) small firms within this sample (based on market capitalization). For comparison to GMW, we also report results for several samples based on the GMW Best-Match Sample: (iii) small firms (based on assets); (iv) small firms (based on market capitalization; (v) small firms (based on trading volume); (vi) small NYSE firms (based on assets); and (vii) small Nasdaq firms (based on assets). We exclude two firms with share price < $1 at June 30, 2004 but obtain similar results (not reported) with a $5 share price minimum (used by GMW). Vertical lines indicate the experiment announcement, start, and end. Shading indicates the (-10, +1) window.

In Figure 2, there is evidence of a relative drop in BHRRs over the (-10, +1) event window for only two of the seven samples: small pilot firms (based on assets) within the GMW Best-Match Sample, and small Nasdaq firms within this sample. For both samples, pilot firm returns are initially higher than control firm returns following SEC approval of the experiment on June 23, but then decline during the two weeks before the formal announcement. Considering all seven samples, the negative BHRR reported by GMW is not robust to sample choice. In particular, it is not found for the preferred, larger sample (the 2004 Announcement Sample), nor for NYSE firms,

[11] To address firms leaving the sample during this period, we compute daily average relative returns for each group, and then compound them to obtain BHRR. This effectively assumes that shares in the departing firm were sold at the closing price on the last listing day, and the funds reinvested in the remaining pilot (or control) firms. See The Internet Appendix for details. Note that GMW refer to what we call BHRR as a "BHAR equivalent."

even though the experiment was expected to have a larger impact on NYSE firms than Nasdaq firms.[12]

*D.2. Evidence on BHRRs*

In Table 4 we report BHRRs for the samples shown in Figure 2 over several event windows around the announcement of the experiment: (i) day -10 to +1; day -1 to +1 and day 0. We also report *t*-statistics for a two-sample difference in means between pilot and control firms. The BHRRs for the (-1, +1) window and the announcement date are small and statistically insignificant across samples. The day 0 returns are near zero for all samples and are positive (opposite from predicted) for the 2004 Announcement Sample (both all firms and small firms). For the longer (-10,+1) window, the BHRR are negative and statistically significant only for small firms (based on assets) within the GMW Best-Match Sample and for small Nasdaq firms within this sample. For all GMW small firms (based on assets), the BHRR is -1.55% and barely statistically significant (*t* = 2.02). By comparison, GMW report BHRR of -2.35% for small firms (based on assets).[13]

Overall, across samples, we find only very limited evidence for a share price drop for pilot firms prior to experiment announcement. This evidence is not robust to sample choice and is driven by negative returns to small Nasdaq firms (based on total assets) even though the Reg SHO experiment was expected to have a larger impact on NYSE firms. Thus, we do not find support for a causal channel that runs through share returns.

*D.3. Access to the SEC's List of Pilot Firms Prior to Public Announcement*

There is also a serious causal channel question in attributing negative relative pre-announcement returns to pilot firms prior to the SEC announcement to the experiment. To trade in advance of the announcement, someone would need access to the SEC's list of pilot firms prior

[12] In the Internet Appendix, Figure IA-5, we report weekly BHRRs (buy-and-hold relative returns over each week) over 2003-2007 along with 95% confidence intervals, for the 2004 Announcement Sample and small firms within this sample (based on market capitalization). A few weekly relative returns are significant at the 5% level, but no more than would be expected from chance alone. Overall, these weekly BHRR graphs provide no evidence of an effect of the SEC experiment on returns to small pilot firms.

[13] GMW report BHRR but not its statistical significance. They report statistical significance for a regression-based measure of relative returns. In the Internet Appendix, Table IA-6, we consider this measure; results are consistent with those reported in Table 4.

to public release of this list. This seems highly remote.[14] Further, we find no evidence of a significant price reaction either over a narrower (-1,+1) window or on day 0 (July 28, 2004), when the SEC released the actual list. This further suggests that one cannot treat the modest negative pre-announcement relative returns, for a particular sub-sample of small firms based on total assets, shown in GMW, as a true response to the experiment.

### E. Causal Channel III: Managerial Fear

FHK and a number of other studies rely, as a causal channel for indirect effects, on the assertion that managers changed their behavior due to fear of increased short selling.[15] This channel cannot be directly tested empirically, but we can assess its plausibility. Managers would have to believe that the price tests represented an economically meaningful impediment to substantive short selling. This would be contrary to widespread beliefs at the time, discussed above, that by 2004 the price tests had little effect on substantive short-selling – beliefs that are consistent with the lack of evidence for an increase in short interest or lower returns to pilot firms.

But suppose many managers (enough to drive the results that FHK and others report) worried nonetheless. How would they be likely to react? We believe that if managers felt that the experiment would have an important negative effect on their firms, they would have written to the SEC expressing their opposition and there would have been significant media coverage of the rule and manager opposition. A FASB proposal to change accounting for employee stock options

[14] There were enough nuances and details in the SEC's process for choosing pilot firms, not known prior to the announcement on July 28, 2004, so that no one could have known which specific firms would be pilot and which would be control. The SEC often holds material market-sensitive information; for example information about investigations of particular firms. We are unaware of instances of leakage or theft of this information. Billett, Liu and Tian (2020) note that SEC staff held meetings with market participants during June and July 2004 to discuss the logistics of conducting the experiment. We can think of no reason why, at those meetings, the SEC staff would have provided nonpublic information about which firms would be pilot firms.

[15] FHK (at 1255) assert that "The decision to eliminate all short-sale price tests prompted a huge backlash from managers and politicians." However, as support, FHK cite stories from the financial crisis period in 2008, long after the experiment. GMW (at 1739) assert that "In public comments, NYSE officials, specialists, and member firms all expressed support for short-sale restrictions." However, as support, they cite only comment letters from the NYSE and its specialist association. As we discuss below, member firms supported the experiment. Most other studies that rely on a manager fear channel either simply assert the potential for manager fear or cite FHK or GMW.

provides a relevant example. When FASB proposed requiring expensing at the fair value of options, this provoked major, well-publicized opposition from many firm managers, including significant press coverage, and an appeal to Congress to override the proposed rule.[16] Thus, we undertake a detailed examination of comment letters to the SEC on the proposed experiment and press coverage of the experiment.

We searched the business press (including the Dow Jones News Service (DJNS), Bloomberg, the Wall Street Journal (WSJ), and the New York Times) for news stories and other information about the experiment, during the pre-launch period (September 2003-April 2005) and the experiment period (May 2005-July 2007), and present results of the search in Internet Appendix Table IA-7, Panel A.

We find that the short-sale experiment attracted very little business press attention. The SEC's June 23, 2004 approval of the experiment was covered in a short, favorable WSJ story, which does not mention manager opposition. The story quotes the head of the SEC's Division of Market Regulation as saying that the SEC had expanded the scope of the pilot program, relative to the proposing release, because of "overwhelming comment" from the industry to "expand the pilot program." The July 28, 2004 announcement of the pilot was not covered in any of the standard business news sources. Coverage of the experiment between the announcement and the actual launch was sparse, and contained only technical explanations of how the experiment would work.[17] The experiment launch, in May 2005, was noted in a DJNS story a few days earlier, with a WSJ summary the next day. In 2006, the SEC extended the experiment, originally scheduled for one year, for a second year, with minimal press attention and no apparent controversy.[18]

---

[16] The first effort by the Financial Accounting Standards Board to require expensing at the fair value of employee stock options, in June 1993, produced more than 700 comment letters opposing the requirement and significant media coverage of the proposed rule. The companies lobbied the Congress and the SEC Chair Arthur Levitt urged FASB to ease off on the proposed rule. See, e.g., Bodie, Kaplan, and Merton (2003), Gleason and Glendening (2019).

[17] The next business press story we found was on 30 November 2004 (four months after the SEC announced the experiment). This story explained that the SEC was delaying the experiment launch to give the exchanges time to make programming changes needed to implement the experiment. This was not even a separate story about the short sale experiment; instead, it was appended to a main story about another SEC rule. Judith Burns, SEC Delays Short-Sale Pilot, Seeks NMS Comment, *Dow Jones News Service* (Nov. 30, 2004).

[18] We found one DNJS story about the extension, with a WSJ summary the next day. Neither story was long

The SEC's December 2006 proposal to repeal the short-sale rule also attracted minimal attention and no apparent opposition. A New York Times story about the repeal explains:[19]

> You may not have read of this proposal. It was virtually ignored by the news media, and if any companies are upset about it, they have not made themselves known. A pilot program that exempted some companies from the so-called uptick rule starting in 2005 drew little attention.

This is the *only* New York Times story about the experiment we found.

The SEC formally approved repeal in June 2007. A *Wall Street Journal* story on the repeal explained that the rule had become "more of an annoyance than a hindrance" to short-sellers, discussed researchers' view that "the uptick rule's usefulness has disappeared," and did not mention any opposition to repeal.[20] The official repeal announcement, issued on July 3, 2007, received no press coverage.[21]

As a further check on whether there was substantial managerial concern, we reviewed all comment letters the SEC received for its November 2003 proposal for the short-sale experiment.[22] Internet Appendix Table IA-7, Panel C lists all comment letters from organizations (as opposed to individuals), and indicates whether they supported the experiment without changes, supported the experiment but proposed change, or opposed the experiment.[23]

As the SEC noted in the adopting release, most of the comments it received supported the experiment.[24] Of the 23 comments that the SEC received on the experiment from organizations,

---

enough to warrant a byline. SEC Pilot Program To Halt `Uptick' Rule, *Wall Street Journal* (April 29, 2005); SEC to Extend Test On Short-Sale Rules, *Wall Street Journal* (April 22, 2006).

[19] Floyd Norris, 70 Years Later, A Scapegoat Gets a Break, *New York Times* (Dec. 8, 2006). The NYT also published an op-ed article in October 2006, supporting repeal of the short-sale rule. Richard Sauer, Bring on the Bears, *New York Times* (Oct, 6, 2006).

[20] Spencer Jakab, Short-Sellers May Owe ETFs Some Thanks – Dropping of 'Uptick' Rule By SEC Comes as Growth Of Stock Baskets Is Soaring, *Wall Street Journal* (June 15, 2007).

[21] SEC Release 34-55,970 (July 3, 2007).

[22] SEC Release No. 34-48709, 68 Federal Register 62972 (Nov. 6, 2003). The comment letters are available at https://www.sec.gov/rules/proposed/s72303.shtml.

[23] Of 43 comments from individuals, 36 supported the experiment.

[24] See SEC Release 34-50103 (July 28, 2004), 69 Federal Register (Aug. 6, 2004), at 48008 and 48012.

20 were favorable.[25] The only negative letters came from the NYSE, the NYSE's specialist association, and a small quantitative trading firm, Susquehanna International Group. Nasdaq supported the experiment, as did several other exchanges and trading platforms (Chicago Board Options Exchange, Chicago Stock Exchange, Archipelago Holdings), major banking and investment banking firms (Charles Schwab, Citigroup, Goldman Sachs, JPMorgan Securities, Merrill Lynch, Morgan Stanley, UBS Securities) and mutual fund groups (Investment Company Institute, Managed Funds Association). Many comments supported expanding the number of firms to be covered by the experiment – a change that the SEC adopted. There were no comments from firm managers, other than the favorable comments from financial firms.

We also reviewed all comments received by the SEC on its 2006 proposal to repeal the rule.[26] Table IA-7, Panel C summarizes the nine comments from organizations; all are favorable. All supported repeal, including the NYSE, which had originally opposed the pilot.[27] Two comments recommended retaining the uptick rule for small-cap firms, which had not been part of the experiment. Once again, the SEC received no comments from firm managers.

We view this examination of comment letters as confirming financial industry support for the experiment, no evidence of opposition from firm managers, and more broadly, lack of evidence that firm managers viewed the short-sale rule as important, one way or the other. This analysis cannot disprove the existence of managerial concern, but we believe that it does show lack of widespread managerial fear.

How else might initially fearful managers react? By 2006, and certainly by 2007, they would likely realize that the bears were not charging. Any initial effects, found in 2005, would likely diminish. In contrast, random differences drift at random and thus might even increase in

[25] We include in organizational letters a comment from Prof. James Angel of Georgetown University, a well-known expert on securities markets. We count a joint submission by Citigroup, Goldman Sachs, Merrill Lynch, and Morgan Stanley as four comments and a joint submission by JPMorgan Securities and UBS Securities as two comments.

[26] SEC Release No. 34-54891 71 Federal Register 75068 (Dec. 13, 2006). The comment letters are available at https://www.sec.gov/comments/s7-21-06/s72106.shtml.

[27] The SEC also received 8 comments from individuals; of these four supported and four opposed repeal.

2006 and 2007. None of the indirect effects papers tests for a decline in treatment effects in 2006 and 2007. Each of the four re-examined papers finds the opposite (larger effects in 2007). The statistical significance of the reported results depends heavily on the larger magnitudes in 2007.[28]

In sum, there are several substantive reasons to discount the likely magnitude of a potential fear channel. These include minimal news attention to the experiment; no news stories suggesting manager opposition; no comment letters from firm managers to the SEC opposing the experiment; and observed treatment effects are often stronger in 2007 than in 2005 or 2006.

**F. Other Possible Causal Channels**

The three channels discussed above are the principal ones relied on by the indirect-effects studies, but other causal channels are possible. For example, LLS and a few other papers posit a price efficiency channel, where the greater ease of short selling makes market prices more responsive to new negative information and thus more informative, managers realize this, and therefore rely more heavily on market prices when making decisions. While we do not directly reassess this channel, we note that: (i) if substantive short selling does not change, it is not obvious why one would expect a change in either price efficiency or the closely related concept of price informativeness (Bond, Edmans, and Goldstein, 2012); (ii) if price efficiency for negative information did increase, this would suggest, even if not strictly require, negative relative returns for pilot firms, which are not observed; (iii) prior studies have not found evidence of improved price efficiency (Alexander and Petersen, 2008; Diether et al., 2009; Bai, 2008**)**; and (iv) the lack of evidence for manager attention to the experiment, discussed above for the manager fear channel, makes it unlikely that managers would have changed their behavior based on assumed greater price efficiency or informativeness.

[28] For FHK (see Internet Appendix, Figure IA-6, Panel D3), for GMW (see Figure 4; for HHZ (who rely on auditor fear), see Figure 5; for LLS (who rely on a price efficiency channel), see Figure 6.

## III. Re-examination Methodology

We now turn to re-examining core results from the four studies mentioned earlier. We began our re-examination by asking how *we* would design a study to answer the principal research questions in the re-examined papers. In a number of instances, we prefer simpler specifications that rely on the initial randomization as a basis for balance between pilot and control firms. For example, we prefer to avoid covariates; doing so increases sample size and avoids potential bias from using covariates that might be affected by the experiment.

### A. Sample Selection for Reexamination of FHK, GMW, HHZ, and LLS

To study the effects of the actual experiment, one must update the 2004 Announcement Sample to account for changes between announcement and experiment launch; we also exclude financial and utility firms. These changes lead to a "2005 Analysis Sample," comprising 2,115 firms, which we use in our re-examinations. This sample closely follows the SEC's experiment design and is substantially larger than those in the studies we re-examine. This may be one reason why our results diverge from the papers we re-examine. We summarize our sample selection choices and other specification choices here; the Internet Appendix provides additional details.

We construct the 2005 Analysis Sample as follows. Although the SEC did not publish in 2004 a list of either Category B firms or other control firms, it did publish on April 13, 2005 (shortly before the start of the experiment on May 2, 2005) updated lists of Category A and Category B firms. We use the updated lists to exclude 38 pilot firms from the 2004 Announcement Sample and move one firm from control to pilot status. For Category B firms, we exclude the 15 firms that the SEC excluded due to mergers or acquisitions. We also exclude 67 additional firms (5 pilot, 62 control) that ceased trading as of May 2, 2005. This leaves 943 pilot and 1,891 control firms. We then exclude financial firms (SIC 6000-6999), utilities (4900-4999) and three firms that did not file 10-Ks for fiscal 2004, resulting in a 2005 Analysis Sample of 2,115 firms (702 pilot and 1,413 control).[29] HHZ, GMW, and FHK all impose additional sample restrictions. HHZ and

[29] FHK, GMW, and HHZ also exclude financial and utility firms; LLS do not.

GMW require firms to be in the R3000 in both 2004 and 2005 and FHK use a balanced sample of firms with data over 2001-2010.

## B. Specification Choices

### *1. Firm and Year Fixed Effects*

In our pre-specified research design, we organize firm-year observations based on fiscal year, and use firm and fiscal year fixed effects (FE), which is a standard DiD specification. GMW and LLS also use firm and fiscal year FE; HHZ use firm FE but not year FE; FHK's main models do not use either firm or year FE, but they report similar results with year FE.

### *2. Sample without Covariates, and are Covariates Appropriate?*

Our pre-specified design does not include time-varying covariates. Given the initial randomization, one can obtain unbiased estimates without covariates, especially since firm FE can absorb chance imbalances between pilot and control firms. The re-examined papers use a variety of covariates, but also report either univariate results (FHK, HHZ, GMW), or results with firm and year FE but without covariates (LLS). In practice, using their covariates makes very little difference in coefficient estimates or precision.[30]

### *3. Balanced versus Unbalanced Panel*

In our pre-specified design, we rely on an unbalanced panel. This permits a larger sample and is appropriate since there is no reason to expect differential attrition between pilot and control firms. We confirm the absence of differential attrition in our pre-analysis plan. HHZ, GMW, and LLS all use unbalanced samples. FHK use a balanced panel of firms with data throughout 2001-2010, but state that they found similar results with an unbalanced panel. Our reexamination of FHK considers both unbalanced and balanced panels.

[30] We present results for FHK, GMW, and LLS with covariates in the Internet Appendix due to space constraints. We present results for HHZ (who study a single outcome) in the text both with and without covariates.

### *4. Sample Periods and Transition Period*

Each reexamined paper makes a different choice regarding the Pre, During, and Post periods, and whether to exclude a transition period (roughly, 2004) between experiment announcement and launch (FHK, HHZ, and LLS do so), and include a Post period in the study (FHK and HHZ do so).

In our pre-specified design, we use fiscal years 2001-2010 as our sample period. We define which firm fiscal years are in this period using the Compustat convention, under which, if the fiscal year-end month is January-May, the fiscal year is (calendar year – 1); and if the fiscal year-end month falls in June through December, fiscal year is the calendar year in which the fiscal year ends. Thus, our sample period includes fiscal year ends from June 2001 through May 2011.

We believe the logic for including a Post period is compelling. Any treated-minus-control difference found during the experiment should reverse once the experiment ends. Observing whether reversal occurs is an important robustness check. Since the experiment started in May 2005, it is unclear if the year 2004 should be excluded but this may depend on the posited causal channel. [31]

Judgment is needed on which firm fiscal years fall within the Pre, During, and Post periods. The experiment was announced in late July 2004 and ran from early May 2005 to early July 2007.[32] In our pre-specified design, we treat firm fiscal years for which half of the year or more falls within the experiment period as within the During period (firm 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. We include the period from experiment announcement in July 2004 through launch in May 2005 in the Pre period; in contrast FHK, HHZ, and LLS exclude this period.

[31] For some research questions exclusion of 2004 is not appropriate. For example, LLS conjecture that prices become more informative *after* the price tests are eliminated. For this causal channel, there does not appear to be a good reason to exclude a transition period, rather than include 2004 in the Pre period.

[32] See SEC Release 34-55970 (July 3, 2007) (ending the short-sale restrictions effective July 6, 2007).

### *5. Winsorization*

We winsorize outcomes at the 1% and 99% levels across all sample years. When we use covariates, we winsorize across all sample years, usually at the 1% and 99% levels; see Table 2, Panel A for details. Each re-examined paper makes different winsorization choices; see the Internet Appendix for details.

## C. Methodology: FHK and Accruals Measures

We use both FHK's PMDA measure and three simpler accruals measures – operating accruals, total accruals, and abnormal accruals (AA), which we measure using the modified Jones model. Given the randomization, the simpler measures can provide unbiased estimates of mean pilot-versus-control changes in earnings management. We discuss in the Internet Appendix (Section IV) why, given the randomization and a large sample, the simpler measures may be preferable and provide evidence that they are less sensitive to specification choice. At a minimum, they offer useful robustness checks.

Following Healy (1985) and Sloan (1996), much prior research has studied operating accruals or AA. However, Richardson, Sloan, Soliman and Tuna (2005) show that investing accruals are also mispriced, and Desai, Krishnamurthy and Venkataraman (2006) provide evidence that short selling prior to restatements is related to both operating and total accruals (operating plus investing accruals). Dechow, Ge, Larson and Sloan (2011) and Larson, Sloan, and Giedt (2018) recommend a comprehensive accruals measure. Therefore, we consider both operating and total accruals, defined as:

$$\textit{Operating accruals (OPACC)} = (\textit{Earnings} - \textit{CFO}) / \textit{LagAssets}$$

$$\textit{Total accruals (TOTACC)} = (\textit{Earnings} - \textit{CFO} - \textit{CFI}) / \textit{LagAssets}$$

where Earnings is earnings before extraordinary items from the statement of cash flows (Compustat data item IBC), and CFO (OANCF minus XIDOC) and CFI (IVNCF) are cash flow from operations and cash flow from investing activities, respectively. LagAssets is the book value of total assets (AT) at the end of the prior fiscal year. Following Hribar and Nichols (2007), we

winsorize accruals at 1% and 99% across all fiscal years to reduce the influence of outliers and address data entry errors in Compustat.

We also study AA and its derivative, PMDA. We compute AA and PMDA using all firms on Compustat. We follow FHK and estimate accruals cross-sectionally within each fiscal year and Fama-French 48 industry, using the modified Jones model:

$$\frac{TA_{i,t}}{AT_{i,t-1}} = \beta_0 + \beta_1 \frac{1}{AT_{i,t-1}} + \beta_2 \frac{\Delta REV_{i,t}}{AT_{i,t-1}} + \beta_3 \frac{PPE_{i,t}}{AT_{i,t-1}} + \varepsilon_{i,t} \quad (1)$$

where *i* indexes firms and *t* indexes fiscal years. $TA_t$ is earnings before extraordinary items (IBC) minus operating cash flows (OANCF minus XIDOC) for year *t*. $AT_{t-1}$ is total assets at the end of fiscal year *t*-1. $\Delta REV_t$ is the change in sales revenue (SALE) from year *t*-1 to *t*. $PPE_t$ is gross property, plant and equipment (PPEGT) at the end of year *t*. We then estimate equation (1), requiring a minimum of 10 industry-year observations, and use the estimated coefficients to calculate normal accruals $NA_{i,t}$:

$$NA_{i,t} = \widehat{\beta_0} + \widehat{\beta_1} \frac{1}{AT_{i,t-1}} + \widehat{\beta_2} \frac{(\Delta REV_{i,t} - \Delta AR_{i,t})}{AT_{i,t-1}} + \widehat{\beta_3} \frac{PPE_{i,t}}{AT_{i,t-1}} \quad (2)$$

where $\Delta AR_{i,t}$ is the change in accounts receivables (RECT). We then calculate firm-year-specific abnormal accruals and PMDA as:[33]

$$AA_{i,t} = (TA_{it} / AT_{i,t-1}) - NA_{i,t}. \quad (3)$$

$$PMDA_{i,t} = AA_{i,t} - AA_{j,t} \text{ for matched firm } j$$

To find the matching firm *j*, we follow Kothari, Leone and Wasley (2005) and match each firm-year observation with another observation from the same year and industry with the closest same-year return on assets ($ROA_{Kothari,t}$, defined as net income divided by total assets ($AT_t$)).[34]

[33] Following Hribar and Nichols (2007), we winsorize the variables used to estimate AA and PMDA at 1%/99% within each fiscal year. We winsorize again, across all sample years, in estimating equations (4)-(5).

[34] We note, however, that using a time-varying match is not advisable in a "causal" project. Since accruals affect ROA, one should match, if at all, on pre-treatment values of ROA, determined in 2004.

## D. Difference-in-Differences (DiD) Specification

### *1. Simple DiD Specification*

To test whether pilot firms had lower accruals than control firms during the experiment period, we estimate the following DiD model for each accruals measure over 2001-2010.

$$y_{i,t} = \beta_0 + \gamma_t + f_i + \beta_1 Pilot_i * During_t + \beta_2 Pilot_i * Post_t + \varepsilon_{i,t} \quad (4)$$

Here $y_{i,t}$ is the accruals measure; $Pilot_i$ = 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 $\gamma_t$ and $f_i$ are year and firm FE. Non-interacted terms are omitted because they are absorbed by the firm and year FE. A negative coefficient $\beta_1$ on Pilot*During provides evidence that pilot firms reduce earnings management during the experiment period. The expected coefficient $\beta_2$ on Pilot*Post should be close to zero because short-sale restrictions were removed for all firms following the experiment. We can also test for a sign reversal in the Post period, relative to the experiment period, by replacing Pilot*During with (Pilot *(During or Post)) in equation (4). With this specification, in equation (5), FHK predict a positive coefficient on Pilot*Post, similar in magnitude to the negative coefficient on Pilot*During in equation (4).

$$y_{i,t} = \beta_0 + \gamma_t + f_i + \beta_3 Pilot_i * (During\ or\ Post)_t + \beta_4 Pilot_i * Post_t + \varepsilon_{i,t} \quad (5)$$

We cluster standard errors on firm and do not include covariates, but in the Internet Appendix, we also use a specification with covariates, in which we add $\boldsymbol{\lambda}*\mathbf{x}_{i,t}$ to eqns. (4)-(5), where $\mathbf{x}_{i,t}$ is a vector of the covariates used in each paper (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 treatment period, we use two graphical approaches. The first is a plot of univariate differences in means for pilot versus control firms. When data are available, we extend these plots back to 1998. Showing a longer pre-treatment period is useful in identifying non-parallel pre-treatment trends (which could arise by chance), and in assessing random variation in means between the two groups, during the pre-experiment period. The second is a "leads-and-

lags" specification, in which we estimate a separate "treatment effect" for each year, before, during, and after the experiment period, and plot the annual coefficients and 95% confidence intervals (CIs) in leads-and-lags graphs. We present and discuss the lead-and-lags graphs in the Internet Appendix; they provide no surprises relative to the annual means graphs.

### *3. F-score and the FHK HF-score Variant*

FHK also study the effect of the SEC experiment on the likelihood of future misstatement of earnings by pilot firms. Dechow et al. (2011) develop an F-score measure which predicts the likelihood of a material misstatement. They develop three summary measures of the likelihood of a future restatement (we refer to them as F-1, F-2 and F-3). The first measure (F-1) is based on financial statement variables such as operating performance and accruals. The second measure (F-2) adds off-balance sheet items, such as operating leases, and non-financial measures, such as change in the number of employees. The third measure (F-3) adds market-related variables such as market-adjusted returns.

FHK create a binary variable ($HF_{i,t}$) that equals 1 if firm *i*'s F-score for year *t* is in the top 1% of F-scores in their sample across all sample years, and 0 otherwise. They study the resulting HF-1, HF-2, and HF-3 measures. FHK report that the coefficient on Pilot*During is significantly negative (HF of pilot firms declines relative to controls during the experiment period) for all three measures, but do not find reversal after the experiment ends. We report results for both F-score and HF-score.

Our DiD specification for F-score uses equations (4)-(5), with different dependent variables. For HF-score, which is a binary variable, we follow FHK in using probit estimation, drop firm FE to avoid the incidental parameters problem, and add Pilot dummy.

## E. GMW Specifications

GMW study five outcomes: capital expenditures/assets, (capital expenditures + R&D)/assets, percent growth in assets, equity issuance, and debt issuance, for all firms and for small firms. We also study R&D/sales, because we believe that when studying capital expenditures and (capital expenditures plus R&D), one should also study R&D by itself. In our

pre-specified design, we chose to study R&D/sales rather than R&D/assets. Both denominators are used by researchers. As a consistency check for GMW's choice to study percent asset growth, we also study *ln*(assets). In a panel specification with firm and year FE, *ln*(assets) provides an alternate measure of asset growth.

Our DiD specifications use equations (4)-(5), with these dependent variables. GMW do not examine whether results from the experiment period reverse after the experiment ends. We study both the During and the Post periods.

**F. HHZ Specifications**

HHZ study only one outcome, *ln*(audit fees). Our DiD specifications use equations (4)-(5), with *ln*(audit fees) as the dependent variable. Our first specification uses firm and fiscal year FE, but no covariates. With *ln*(audit fees) as the outcome variable, a case can be made for controlling for firm size, since size is a known, powerful predictor of audit fees, even though an effect on growth is one possible outcome of the experiment. Therefore, in a second regression model, we control for *ln*(assets). We also use regression models which include, respectively, the short list of covariates in HHZ Table 5, model 1 (which controls for *ln*(sales) as a measure of size); and the longer list in HHZ Table 5, model 2.

**G. LLS and Triple Difference Specification**

LLS study wealth-performance sensitivity (WPS) using a DiD specification, and also R&D /assets and (capital expenditures + R&D)/assets using a triple difference specification, with Pilot*During interacted with Tobin's *q*. We study these outcomes as well as capital expenditures/assets, because we believe that when studying R&D and (capital expenditures plus R&D), one should also study capital expenditures by itself. We study both R&D/sales and R&D/assets. For WPS, we again rely on eqns. (4)-(5). For the other outcomes, we replace equation (4) with a triple difference specification:

$$y_{i,t} = \beta_0 + \gamma_t + f_i + \beta_1 * q_{i,t} + \boldsymbol{\gamma} * \boldsymbol{DBL_{i.t}} + \beta_2 * q_{i,t} * Pilot_i * During_t + \beta_3 * q_{i,t} * Pilot_i * Post_t + \varepsilon_{i,t} \quad (6)$$

Here $q$ is Tobin's q; **DBL** is a matrix 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 vector of coefficients. The core coefficients are those on the triple interactions, $\beta_2$ and $\beta_3$. We make analogous changes to equation (5) to measure sign reversal. LLS do not study whether results during the experiment period reverse after the experiment ends. We study both the During and the Post periods.

## IV. Reexamination Results

In this part, we reexamine core results from each paper. Our focus is on the robustness of the reported results to alternative specifications. We present results using: (i) our pre-specified research design and sample; and (ii) best match samples and specifications—samples and regression specifications that match those from each paper as closely as we can, based on what each paper states it does. We constructed the FHK and HHZ best-match specifications without knowing the authors' actual choices and samples, which we obtained later. In the Internet Appendix, we provide results with covariates. We treat results as statistically significant if they are significant at the 5% level or better in a two-sided test. In regression tables, we denote "marginally significant" results (significant at the 10% level) with a single *.

### A. Summary Statistics and Pre-Treatment Balance

Table 2, Panel A reports variable definitions. Panel B provides evidence for balance on pre-treatment outcomes and covariates for fiscal 2004, the year prior to experiment launch. Pilot and control firms are similar on both outcomes and covariates, as expected given the initial randomization.

### B. FHK: Results for Accruals

For FHK, we present results for four measures: operating accruals, total accruals, abnormal accruals, and PMDA, and for both unbalanced and balanced panels. FHK report results only for PMDA and a balanced panel, but assert that unbalanced panel results are similar. For the balanced panel results, we require the FHK covariates (*ln*(Total Assets); Market-to-book ratio; Return on assets; and Leverage) to be non-missing for all sample years, but do not include them in

regressions.

*B.1. Graphical Evidence*

Given the initial randomization and no evidence of differential attrition, a simple comparison of means can provide valuable information. In Figure 3, we provide annual means for all four accruals measures, separately for pilot and control firms. There is no evidence of a treatment effect for any of the measures. Balanced panel graphs using our specification, and graphs using the FHK Best-March Specification with either unbalanced or balanced panel, are similar (see Internet Appendix, Figure IA-6).

*B.2. Regression Evidence*

Table 5, Panel A presents regression results for both unbalanced and balanced panels. We report coefficients for Pilot*During, Pilot*Post, and Sign Reversal (the *change* in the coefficient from During to Post). Panel B presents results using the FHK Best-Match specification; we discuss the best-match specifications below. Panel C presents the FHK reported result, for comparison. In Panel A, all 16 relevant coefficients (4 accruals measures; unbalanced and balanced panels; coefficients on Pilot*During and Sign Reversal) are insignificant, with mixed signs. The PMDA coefficients have the opposite sign (positive) relative to FHK's prediction and result.

For the best-match specifications in Panel B, the coefficients for Pilot*During and Sign Reversal are more often negative, but all remain insignificant. PMDA continues to take a positive sign for the unbalanced panel. With a balanced panel (the FHK specification) the PMDA coefficient changes sign from positive with our specification to negative, but at -0.0022 is far below the FHK reported coefficient of -0.010 and not close to statistical significance.

Thus, neither graphical nor the regression evidence supports the FHK conjecture, with either our specification or the FHK Best-Match Specification.

*B.3. One-Way versus Two-Way Clustering*

In Table 5, we present both standard errors (s.e.'s) clustered on firm and standard deviations (s.d.'s) based on randomization inference (explained below). The other three papers

use standard errors clustered on firm, as does our pre-specified design. We also present results with randomization inference to assess the validity of the two-way clustered s.e.'s that FHK report, which are much lower than ours (0.004 versus our 0.0087).

Clustered standard errors with a small number of clusters can be downward biased (e.g., Cameron, Gelbach and Miller, 2008). We investigated the large discrepancy between one-way and two-way clustered errors. At a qualitative level, s,e,'s annual treatment effects for PMDA are around 0.0175 (see Internet Appendix, Figure IA-12). Averaging over the three-year experiment period can increase precision at the usual $n^{0.5}$ rate. Thus, the s.e. for Pilot*During should be around $0.0175/3^{0.5} \approx 0.01$. This is close to the observed s.e. clustered on firm. Conversely, FHK's reported s.e. of 0.004 for a 3-year experiment period therefore appear too small.

To confirm which s.e.'s are correct (if any), we used randomization inference to obtain the exact s.d.'s of coefficient estimates. For each sample of pilot and control firms in Table 5, Panel A, we drew at random the correct number of pilot firms. For example, for PMDA with a balanced panel, the sample includes 1,398 firms (492 pilot and 906 control). We drew 492 of these firms at random, assigned them to be pseudo-pilot firms, and computed the pseudo-coefficients on Pilot*During and Pilot*Post using the same regression specification as in Table 5 (from equation (4)). We repeated this procedure 1,000 times to obtain an empirical distribution of the pseudo-coefficients, and measured the s.d. As Panel A shows, the randomization inference s.d.'s are very close to s.e.'s with firm clusters, and for PMDA they are twice as large as the s.e. that FHK report.[35] Thus, downward bias in two-way clustered s.e.'s drives the statistical significance for their reported coefficient, which would be insignificant with one-way clustered s.e.'s or randomization-based s.d.'s.

[35] For the FHK Best-Match Specification, standard errors clustered on firm (not reported) are very close to the randomization inference s.d.'s reported in Table 5. In work in progress, we are assessing the performance of one-way clustering on firm, versus two-way clustering on firm and year, compared to randomization inference, for a number of other finance and accounting datasets.

## C. FHK: Results for F-Score and HF-Score

We next re-examine a second FHK result. FHK report that pilot firms had a lower likelihood of a material misstatement, using their binary HF-score measure. We study both F-score and the FHK measure.

In Table 6 columns (1)-(6), we report the three F-score measures, using an OLS specification similar to Table 5: unbalanced and balanced panels; no covariates; with firm and fiscal year FE. In columns (7)-(12), we report the corresponding HF measures, using probit estimation. We report results for our specification in Panel A, for the FHK Best-Match Specification in Panel B, and the FHK reported results in Panel B. Note, however, that panels A and C are not directly comparable because we report marginal effects (which have interpretable meaning), while FHK report probit coefficients (which do not). With our specification, across all measures, there is no evidence of a treatment effect. For F-score, the coefficient signs are consistently *positive*, opposite from the FHK prediction. For HF-score, all coefficients are small and insignificant, with mixed signs.

With the FHK Best-Match Specification (Panel B), all F-score coefficients are insignificant; the coefficients are positive with a balanced panel. For HF-score, we report probit coefficients, for better comparability to FHK. For HF-1 and HF-2 with a balanced panel, the coefficients on Pilot*During for HF-1 and HF-2 are similar in magnitude to those that FHK report, but are insignificant using randomization inference s.d.'s (or s.e.'s clustered on firm, not reported). For HF-3, the coefficient is positive (opposite from predicted). Graphical results for annual means (Internet Appendix Figure IA-7) also show no evidence of a treatment effect, with either our specification or the FHK Best-Match Specification. There is no evidence for sign reversal.

In the Internet Appendix (Figure IA-8), we consider HF-score thresholds less extreme than the 1% threshold used by FHK. We consider thresholds from 2.5%- 20%. The marginal effects are positive with a 2.5% or 5% threshold (opposite from predicted), with either our specification or the best-match specification. Overall, considering both F-score and HF-score, there is no evidence that pilot firms act in ways that reduce the probability of a misstatement.

## D. GMW: Firm Investment and Growth

GMW report evidence of lower investment, lower asset growth, and lower equity issuance for small firms (based on assets). We present results for their outcomes (capital expenditures/assets, (capital expenditures + R&D)/assets, percent asset growth, equity issuance, and debt issuance), for R&D/sales as an alternate measure of investment, and for *ln*(assets) as an alternate outcome for assessing asset growth.

We begin with graphical evidence. In Figure 4, we present the annual means for small firms, using our specification. We find minimal evidence for reduced investment, and no evidence for slower growth or less equity issuance:

> *Investment (Capex/Assets, (Capex+R&D)/Assets, and R&D/sales).* There is a relative decline in Capex/Assets, and (Capex + R&D)/Assets, during the experiment period, consistent with a treatment effect. However, the effect is largest in 2007, which is the wrong time for a true treatment effect to appear. There is also: (i) a similar or larger gap during 2000-2001, which then closes, and (ii) no post-experiment reversal. And R&D/sales for pilot firms *rises* in 2006 relative to control firms (opposite from predicted). Thus, there is at most mild, mixed evidence of lower investment by pilot firms.
>
> *Growth (Percent Asset Growth and ln(Assets)):* For percent asset growth, pilot firms show a relative dip in 2007. This is the wrong time for a treatment effect to emerge: the bears were not charging and most of the 2007 data points come from firms with fiscal years ending in December 2007, well after the experiment ended. For *ln*(Assets), annual means move in parallel. Thus, we find no evidence of a treatment effect.
>
> *Equity and Debt Issuance*. In Figure 4, there is no evidence of a treatment effect for equity issuance.

Graphs using the GMW Best-Match Specification are visually very similar (Internet Appendix, Figure IA-10).

We provide DiD regression results for small firms in Table 7. We present results for our specification in Panel A, results with the GMW Best-Match Specification in Panel B, and the GMW reported results in Panel C. Note that GMW report univariate results, without firm or year FE and regression results with firm and fiscal year FE *and covariates*. They do not report regression results with firm and fiscal year FE but *without covariates*, which would be more directly comparable to our results.[36]

With our specification, all seven outcomes in Table 7 are insignificant. Most coefficients are negative, but R&D/Sales takes a positive coefficient (opposite from predicted) and the coefficient for *ln*(Assets) is close to zero. With the Best-Match Specification, the coefficient for CAPEX/Assets is negative and mildly significant ($t = 2.06$). However, significance is driven by a gap that increases in 2006 and again in 2007, with no post-experiment reversal (Figures 4 and IA-10). All other coefficients are insignificant and equity issuance takes a positive coefficient.

In sum, considering both graphical and regression evidence, and both our specification and the GMW Best-Match Specification, there is minimal evidence for the GMW hypothesis of lower investment and slower growth by small pilot firms.

**E. HHZ: Auditing Fees**

HHZ report that pilot firms have about 5% higher audit fees during the experiment period, with bare statistical significance (insignificant for univariate difference in means, $t = 1.99$ with limited covariates and $t = 1.98$ with more extensive covariates).

In Figure 5, we provide annual means for *ln*(Audit Fees), separately for pilot and control firms. The two sets of means move closely in parallel with each other, and provide no evidence of a treatment effect. The pilot firm mean is somewhat below the control firm mean for 2007, but this is the wrong time for a treatment effect to appear, since the experiment ended in mid-2007. Moreover, the gap does not close in the Post period.

[36] GMW also present triple-difference results, in which the third difference is between large and small firms. However, the DiD and triple-difference specifications will produce identical results in a fully interacted model, which includes interactions between small firm dummy and the year dummies, the constant term, and the covariates.

We present regression results in Table 8, using an unbalanced panel, as do HHZ. We present results for our specification in Panel A, results with the HHZ Best-Match Specification in Panel B, and the HHZ reported results in Panel C. Column (1), with firm and fiscal year FE but no covariates, is our preferred specification; column (2), which controls for *ln*(Assets) (which correlate strongly with audit fees) is a sensible alternative. In column (3), we use HHZ's short list of covariates, and in column (4) their full list. With our specification, all coefficients are small and slightly *negative* (opposite from predicted).[37] The same is true for a balanced panel (Internet Appendix, Table IA-18). With the HHZ Best-Match Specification, all coefficients become positive, but are far lower than those reported by HHZ and not close to statistical significance.

In sum, neither graphical nor regression evidence supports the HHZ hypothesis of higher audit fees for pilot firms, for either our specification or the HHZ Best-Match Specification.

**F. LLS: Sensitivity of Investment and CEO Wealth to Share Price**

We present results for the LLS outcomes, plus R&D/Sales (they study R&D/Assets) and Capex/Assets. We begin with graphical evidence and show annual means for pilot and control firms in Figure 6. In brief:

*WPS*. WPS is higher for pilot firms in the pre-experiment period, but the gap shrinks over 1999-2004 and remains roughly constant over 2004-2007. Thus, there is no evidence of a treatment effect. However, the non-parallel pre-treatment trends will drive a spurious negative coefficient on Pilot*During in a DiD regression, especially one that drops 2004, as LLS do.[38]

*Investment (Capex/Assets, R&D/Assets, and R&D/Sales)*: The coefficient on Capex/Assets is significant and *negative* (opposite from predicted). R&D/Assets moves in parallel for

[37] In the Internet Appendix Table IA-18, we report results for HHZ with a balanced panel as a robustness check. All regression coefficients are negative (opposite from predicted).

[38] For the LLS sample period, the DiD coefficient compares the pilot-minus control average for 2005-2007 to the pilot-minus-control average for 2002-2003. It is apparent from Figure 6 that this coefficient will be negative, even though there is no meaningful change in the pilot-minus-control difference over 2004-2008.

both groups. R&D/Sales rises for pilot forms in 2005 and 2006, but reverses in 2007. Thus, there is no overall evidence for a treatment effect.

*(Capex + R&D)/Assets*. There is a small relative decline for pilot firms (opposite from predicted) in 2006-2007, and thus no evidence for a treatment effect.

Graphs using the LLS Best-Match Specification are somewhat different, but also provide no evidence for the treatment effect. With this specification, WPS is very similar for pilot and control firms over 2004-2007 (Internet Appendix, Figure IA-11).

We turn to regression results in Table 9. We present results for our specification in Panel A, results with the LLS Best-Match Specification in Panel B, and the LLS reported results in Panel C. With our specification, consider first WPS, for which LLS predict a drop for pilot firms during the experiment period. We find an insignificant drop, but the negative coefficient is driven by non-parallel pre-treatment trends. For their other outcomes, LLS predict a positive sign on the triple interaction Pilot*During*Tobin's *q*. With our specification, most coefficients are negative (opposite from predicted), with a significant negative coefficient for Capex/Assets.

Turning to the LLS Best-Match Specification, only one of the five LLS outcomes is significant. WPS takes a barely significant coefficient of -0.1292 ($t$ = 2.05). However, this negative coefficient is driven by non-parallel pre-treatment trends; the coefficient becomes insignificant if we include 2004 in the Pre period.[39]

In sum, considering both the graphical and the regression evidence, and both our specification and the LLS Best-Match Specification, there is no support for the LLS conjectures.

**G. Summary**

We find no meaningful support for the hypotheses about earnings management in FHK, firm investment and growth in GMW, auditing fees in HHZ, or greater managerial sensitivity to share price in LLS. With our specification, across 23 outcomes (for FHK, using both balanced

[39] LLS rely on a price efficiency channel, which depends on actual short selling. The short-sale restrictions were in place in 2004. We therefore believe that 2004 should be included in the Pre period for their study.

and unbalanced panels)[40] and 36 total regressions, we find one significant coefficient, with the wrong sign (LLS, capex/assets). Our failure to find evidence for their conjectures is consistent with our prior view that the short-sale experiment, which had only minor direct effects on trading markets, likely had only minor indirect effects on firms.[41]

With the best-match specifications, the coefficients on Pilot*During generally move toward the reported results in each paper, but most remain insignificant and far from the reported coefficients. Only two of 23 coefficients are statistically significant with the predicted sign (on for GMW, one for LLS), both have t-statistics barely above 2, and neither is convincing when viewed together with the graphical results. Overall, there is no evidence for a treatment effect for FHK, HHZ or LLS, and minimal evidence for GMW.

In the Internet Appendix, we move step by step from our specification to each best-match specification, and report results at each step (Internet Appendix, Tables IA-10 (FHK), IA-17 (HHZ), IA-22 (GMW), and IA-23 (LLS)). Across these intermediate specifications, there are no significant results for FHK or HHZ; no significant results for GMW other than for Capex/assets (also seen for the GMW Best-Match Specification); while for LLS, WPS is weakly significant for two of the four intermediate specifications (as it is for the LLS Best-Match Specification, and R&D/assets take a positive and weakly significant coefficient (t = 2.11 and 2.18), in two intermediate specifications.[42]

When we began this project, we expected to find differences between the reported results and those with our sample and specification. However, we also expected that we would come close to the reported results with the best-match samples and specifications. The large differences

[40] Ten outcomes for FHK (4 accruals measures, 3 F-score measures; 3 HF-measures); 7 for GMW, 1 for HHZ; and 5 for LLS, including the additional outcomes we study for GMW and LLS.

[41] For FHK and HHZ, two coauthors wrote statistical code independently and confirmed that they obtained the same results. For FHK, we also compared our AA and PMDA code to the code publicly posted by Dan Taylor and Joost Impink.

[42] One can use FHK (accruals), to illustrate this process. The intermediate steps are: (i) switch from periods based on fiscal year to periods based on calendar year; (ii) remove 2004 from Pre period; (iii) switch to FHK sample; (iv) remove firm FE; the final step is to remove calendar year FE .

between our best-match results and the reported results confirm the importance of specification choice, including choices that we could not reproduce from the descriptions in each paper.

## V. FHK and HHZ Technical Replication

Our focus in this paper is on the robustness of the results in the four re-examined papers, not on technical replication. However, in response to an earlier version of this paper, FHK posted their sample and code for their PMDA result, but not their HF-score results. HHZ provided their data and code to us. In the Internet Appendix, we used their exact datasets and codes to conduct technical replication of their results and assess robustness using variations on their samples and specifications. We summarize selected findings here.

### A. FHK PMDA: Technical Replication but Sensitivity to Specification

In the Internet Appendix, we confirm technical replication for their PMDA result using their exact sample and code. However, their result is fragile as discussed briefly below and in more detail in the Internet Appendix.

First, their reported decrease in PMDA for pilot firms turns on quirks of the matching process. In both our specification and the best-match specification, we followed Kothari et al. (2005) and found matching firms for PMDA based on *current year* ROA.[43] From their code, we learn that FHK instead matched on lagged ROA. This choice drives their results. We compare the effects of matching on current versus lagged ROA in Figure 7 using FHK's exact sample. In Panel A, the left-hand figure shows mean AA for pilot and control firms during the Pre, During, and Post periods. AA for the pilot and control firms move in parallel, with no evidence that pilot firms reduced their accruals relative to control firms. The middle figure shows AA of matching firms for the pilot and the control firms, when matching firms are selected based on lagged ROA. AA for the two groups is non-parallel. The non-parallel changes in AA for the matching firms

[43] FHK state that their measure of "earnings management is the [PMDA] measure of Kothari et al. (2005), and that "We match each sample firm with the firm *from the same fiscal year-industry* that has the closest [ROA] as the given firm." Kothari et al. (2005) consider matching on either current-year or lagged ROA, and recommend matching on current year ROA as producing better-specified tests (p. 167). We inferred that FHK had followed Kothari et al. and matched on current-year ROA.

produces the FHK result for PMDA. The right-hand figure shows PMDA for pilot firms (AA minus the AA for their matching firms) and for control firms, and replicates FHK Figure 2.

In Panel B, we instead follow Kothari et al. (2005) and choose matching firms based on current ROA, as Kothari et al. (2005) recommend. The left-hand graph is identical to Panel A: AA for pilot and matching firms move in parallel. The middle graph shows AA for the pilot and control matches; the matches also move closely in parallel from the Pre to the During period. Parallel changes both for pilot and control firms, and for their matches, leads in the right hand graph to parallel changes in PMDA, and thus to no evidence of a treatment effect. Thus, the FHK result for PMDA is driven by nonparallel trends for AA of the **matching** firms, when matching using lagged ROA. Taking both panels together: (i) there is no evidence that pilot firms reduced AA in response to the experiment; and (ii) the evidence for a reduction in PMDA depends crucially on a particular specification choice. We view matching on both current and lagged ROA as reasonable choices. If one reasonable choice leads to a significant result, while another does not, the result cannot be said to be robust.

The FHK result is fragile in other ways, including. Using their exact sample, (i) the coefficients on Pilot*During are insignificant for the other three accruals measures, with either an unbalanced or a balanced panel; (ii) the coefficient on Pilot*During is insignificant for PMDA with an unbalanced sample; (iii) the reported PMDA coefficient loses significance with s.e.'s clustered on firm, (iv) with randomization-inference-based s.d.'s, or (v) with s.e.'s clustered on firm and calendar year (versus the FHK choice to cluster on firm and fiscal year). See Internet Appendix for details.

Figure 8 is presented in the spirit of recommendations from Simonsohn et al. (2020) and others that authors should report results across a range of reasonable but meaningfully different specifications illustrates the overall sensitivity of the FHK accruals results to specification. It shows *t*-statistics for all four accruals measures, across the specifications reported in the text and the Internet Appendix. Solid horizontal lines show 5% significance ($t = \pm 1.96$)**.** Across 4 accruals measures and 142 regressions, only four coefficients are significant (two for PMDA, two for operating accruals), and only mildly so ($t = 1.96$, 2.03, 2.10, and 2.24). Twenty-three coefficients

are positive (opposite from predicted). Manifestly, their accruals result is not robust across a range of reasonable specifications and accruals measures.[44]

### B. FHK Results for HF-Score: Attempted Near-Exact Replication

FHK posted their sample or code for their PMDA result, but not for their HF score results. However, we have their PMDA sample, and the firm-year observations are almost identical for their PMDA and HF-score results.[45] We use their exact PMDA sample, plus the instructions for constructing F-score in Dechow et al. (2011), and the FHK statement of how they defined HF-score, to carry out what should be a near-exact replication of the FHK HF-score results. This replication effort fails. We find that: (i) five of the six HF-score coefficients are positive (opposite from predicted); and (ii) the one negative coefficient, for HF-1 with balanced panel, is -0.090 and statistically insignificant (s.e. = 0.176) versus the FHK reported coefficient of -0.178. See Internet Appendix Table IA-16.

### C. HHZ: Technical Replication but Sensitivity to Specification

Using the HHZ exact sample and code, we confirm technical replication of their results. However, their result is fragile as we show in Internet Appendix Table IA-19. In particular, if we use their exact list of firms, but fill in observations that they missed due to incomplete matching to CRSP, Compustat, and Audit Analytics, the coefficient on Pilot*During, without covariates, drops from the 0.048 they report to 0.021 (t = 0.90).

In Figure 9, we summarize the *t*-statistics that we find for the HHZ outcome, across the specifications and covariate choices in the text and the Internet Appendix. Only four of the 78

[44] We address the FHK Reply (2019) in the Internet Appendix, but note two points bearing on robustness. The FHK Reply reports a significant decline for operating accruals with a balanced panel, but it does so using a different specification than in their paper. See Internet Appendix for details on the specification differences. With their operating accruals specification, (i) the operating accruals coefficient is insignificant with an unbalanced panel and coefficients for the other three accruals measures are insignificant with either a balanced or unbalanced panel. See Figure 8, specification N. The FHK Reply also notes that Massa et al. (2015) and an early version of Heath et al. (2021) find a significant negative coefficient for AA. However, AA is insignificant across all specifications in Figure 8, including FHK's own PMDA and operating accruals specifications.

[45] The FHK sample for PMDA is 9,873 firm-year observations; their HF-score sample is 9,871 firms.

coefficients are significant, and barely so ($t$ = 1.98, 1.99, 2.00, and 2.01**)**. Robustness across a range of samples and specifications is manifestly not present.

## VI. Discussion and Implications

### A. Which Specifications Are Preferred; Which Others Are Reasonable?

Our goal, in re-examining FHK, GMW, HHZ, and LLS, was to assess whether their core reported results were robust to alternative specifications. A natural question, given the differences in results across specifications, is when one specification is preferred over another, versus when two specifications are both reasonable, and simply different.

In some respects, we see our sample and specification as strictly preferred: (i) We were careful to closely follow SEC's approach, including retaining firms actually in the experiment and excluding firms that the SEC excluded; (ii) to not impose sample criteria that can generate survivorship bias; (iii) to include firm and year FE; (iv) to check for post-experiment reversal; (iv) to study annual treatment effects and confirm both parallel pre-treatment trends and that results are not driven by an implausible year (2007); and (v) to avoid using two-way clustered s.e.'s in our setting, where they turn out to be downward biased. Our pre-specified design, including a common sample and sample periods across re-examinations, also has soft advantages in protecting against inadvertent bias in specification choice.

In other respects, alternate choices are reasonable, with no clear basis for preferring one over another. Examples include whether to exclude a transition period; how to define utility firms when excluding them from the sample; how precisely to define the Pre, During, and Post periods; whether to winsorize or exclude outliers; and for FHK, finding a PMDA match using current versus lagged ROA:

We conduct a broad range of robustness checks, including dropping a transition period or not; for FHK, studying four accruals measures rather than one, studying both F-score and HF-score, and using different HF-score thresholds; and for GMW and LLS, studying additional outcomes where one would expect results similar to those for the studied outcomes. More robustness checks and more specifications could, of course be tried. But in the end, if one

specification produces a significant result, while other reasonable specifications do not, the result must be seen as not robust. Especially so if the evidence for a causal channel is weak.

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

We next discuss two sets of implications of our study that go beyond the four papers we re-examine: (i) implications for other studies of the short-sale experiment; (ii) broader implications for research relying on natural or randomized experiments. There is necessarily some overlap between the two sets.

The indirect-effects studies of the SEC experiment rely on DiD. While the DiD approach, unlike IV, does not technically require evidence for a first-stage (a causal channel), a credible causal channel is important for attributing an observed effect to the SEC experiment. Lack of evidence for a causal channel increases the likelihood that the study is underpowered. And when an underpowered study finds statistically significant differences between pilot and control firms are, those differences are likely to be false positives and/or have reported effect magnitudes far greater than true effects (Gelman and Carlin, 2014; Black et al., 2021).

The early studies of the experiment found no meaningful impact on short interest or share prices, and thus provided evidence against the natural causal channels for the conjectured indirect effects. We revisited and confirmed the lack of support for these channels over a broader time period, for a sample that closely tracks the SEC rules for choosing pilot and control firms. We cannot rule out a manager fear channel, but found no support for this channel. Moreover, for this channel, one would expect results to weaken in 2007, both because the experiment ended in July 2007 and because by then, it was apparent that the bears were not charging. This is not observed.

Some indirect-effects studies posit additional channels, but do not test for their existence. For example, HHZ assume both that pilot firm share prices will drop (the share price channel), *and* that this will lead to more securities litigation, but do not assess whether pilot firms experienced more securities litigation. LLS posit that relaxing short-sale restrictions makes share prices more informative about the firm's prospects, *and* that firms rely on these more-informative prices, but

do not test whether pilot firm prices in fact respond more strongly to firm-specific news than control firm prices.

Some reported results would be implausible even if a causal channel existed. For example, He and Tian (2016) report a drop in patenting activity, supposedly reflecting firms' exposure to "patenting-related litigation initiated by short sellers." It seems doubtful that firms would alter their multi-year research strategies in response to a planned one-year experiment. The authors also provide no evidence that short-sellers engage in patent-related litigation. Gong (2020) reports that pilot Nasdaq firms, but not NYSE firms, reduce leverage during the experiment. But the Nasdaq bid test was weaker than the NYSE uptick rule, so finding an effect only for Nasdaq firms is surprising. [46]

The magnitude of some reported results is quite large. For example, GMW report a 0.97% drop in Capex/Assets, which is 17% of the pilot firm mean of 5.7%. This effect would imply a very high elasticity of investment to share price, even if the GMW finding of a roughly 1.5% relative price drop for small firms were correct (compare Morck et al., 1990).[47]

Beyond the four papers we re-examined, the weak evidence on causal channels has implications for the confidence one should have in the results in other papers reporting indirect effects of the SEC experiment. We suggest that a Bayesian thumb, whose weight depends on study-specific details, should be applied against results without a clear causal channel. Compare the similar suggestion for Bayesian-based skepticism in Harvey (2017).[48]

---

[46] In unreported results using our specification, we find no significant effect of the SEC experiment on leverage for either NYSE or Nasdaq firms.

[47] As a further example of implausible magnitudes, Bai, Lee, and Zhang (2020) report a 17% increase in workplace accidents at pilot firms. The posited mechanism is cutbacks in "investment in workplace safety to meet short-term [earnings] targets." This effect is implausibly large, more so since it appears quickly (the authors use a one-year lag). Prior safety investments should not dissipate right away, even if new investment is cut. The authors also report that pilot firms had far fewer accidents than control firms during the pre-treatment period, which is inconsistent with random assignment. Pilot firms do not report higher earnings, as we confirm in unreported results for ROA, with a negative (opposite from predicted) but insignificant coefficient on Pilot*During.

[48] In unreported results, we looked for additional papers for which we could readily measure their outcomes, identified four (Wang, 2018; Chen, Zhu and Chang, 2017; Kim, Lu and Peng, 2020; and Gong, 2020), and assessed the robustness of their results using our pre-specified sample and specification. We found only insignificant results for the core results in all four papers. Details are available from the authors on request.

## C. Broader Implications for Researchers

### *1. Implications for Research Using Natural or True Experiments*

If results exploiting the SEC short-sale experiment are as sensitive to specification choice as we found, what are the implications for other DiD studies of true and natural experiments? Our analysis of the short-sale experiment suggests a number of general steps for research on natural experiments, that emerge from our study. They are not intended to be a complete checklist for DiD studies.

1. Consider a pre-specified research design. When feasible, a pre-specified design can greatly enhance credibility. The actual research will often depart from the pre-specified plan to some extent, but the pre-specified base is still important, and ensures transparency for departures from the initial plan.[49]

2. It is important to specify a causal channel and provide supporting evidence in order to reliably attribute indirect effects to a treatment. For an extended causal chain, each step in the chain should be carefully defended. The need to support the posited channel is especially important if prior studies did not find evidence supporting the channel.[50]

3. Researchers should articulate an underlying theory for the results one should expect, and defend each step in the theoretical chain.

4. Sample choice is important, and researchers should be explicit about how their choices affect sample size, in text or an appendix. Even apparently accidental loss of sample (as in HHZ's matching failures) can meaningfully affect result.[51]

---

[49] We found that preparing a careful design is far harder than one might think. Many sample and specification choices must be thought through. But the payoff in credibility can be substantial.

[50] Other areas in which multiple papers have relied on a suspect causal channel include the impact of state antitakeover laws (e.g., Catan and Kahan, 2016; Baker, 2021), and the impact of "universal demand laws" requiring demand to be made on a company's board of directors prior to filing a derivative lawsuit (Donelson et al., 2021).

[51] Here, HHZ provide a model. They specify each step they took and how it affected sample size. This made it easy for us to understand how they lost sample size, relative to our pre-specified and best-match samples.

5. Researchers should conduct consistency checks (are results consistent across alternative specifications and related outcomes). Some examples from our re-examination: (i) for FHK, we assess four measures of accruals, both unbalanced and balanced panels, and both one-way and two-way clustered s.e's; assess both F-score and HF-score, and vary the HF-score threshold; (ii) for GMW and LLS, who study investment, we study R&D alone for GMW and capital expenditures alone for LLS.

6. Placebo checks can be valuable. For the SEC experiment, for example, any difference between pilot and control firms should be insignificant, after the experiment ends. Yet, only a minority of the indirect-effects papers test for reversal.

7. Pre-treatment parallel trends should be confirmed. We saw above that the LLS results for WPS are driven by non-parallel pre-treatment trends. More generally graphs can provide important information about magnitudes and time variation that a panel regression suppresses.

8. Posting of detailed code and, to the extent possible, the full dataset, ideally with extensive comments explaining the code, is important. We believe that posting code should be mandatory. Only through access to the FHK code for accruals could we determine that they found PMDA matching firms using lagged ROA, that they clustered on fiscal rather than calendar year, and that results were not similar for an unbalanced panel. Only with access to the HHZ code could we find the technical errors that lead to statistical insignificance when corrected. Conversely, we lack the FHK code for HF-score, so we cannot determine why our near-exact replication produces very different results than they report.

9. An important aspect of specification choice is the manner of inference. Randomization inference can often be a valuable approach. Conversely, two-way clustering on firm and year can be badly misspecified in short panels. We recommend that authors who are considering two-way clustering should either conduct randomization inference or else report both one-way and two-way clustered s.e.'s, and rely on the larger of the two for inference.

10. Some potential causal channels, such as the manager-fear channel, are not directly provable or refutable. For these soft channels, researchers should report evidence both for and against the channel.

### *2. For Inference, How Reliable is a t-Statistic of Two?*

Most of the indirect-effects papers, including the ones we reexamine, principally report moderate *t*-statistics, often around 2 or only modestly higher. Given the inevitability of specification choice, and the risk of non-neutral choices, the conventional level for statistical significance in finance and accounting ($p < 0.05$, two-sided test, corresponding to $t \geq 1.96$ in a large sample) may be too low to support reliable inference. Astronomy uses minimum *t*-statistic of around 3; physics uses 5. File-drawer bias alone counsels for a *t*-statistic around 3 (McCrary, Christensen and Fanelli, 2016). The Benjamin et al. (2018) consortium of scholars recommends a p-value of .005, roughly equivalent to $t = 2.80$. The Heath et al. (2020) adjustment for multiple hypothesis testing implies a similar threshold. Harvey, Liu and Zhu (2016) recommend a threshold of 3 for studies reporting asset pricing anomalies. In our view, skepticism about the traditional $t = 1.96$ threshold should increase if the evidence for the causal channel is weak, and perhaps also for results with policy implication, as is the case for the SEC experiment.

## VII. Conclusion

During 2005-2007, the SEC conducted a randomized trial in which it suspended price tests contained in Regulation SHO, for approximately 1000 pilot firms traded on the NYSE), AMEX, or Nasdaq. Initial studies of the experiment found little or no impact of removing short-sale restrictions on short interest or share returns to pilot firms, nor any adverse effect on liquidity or volatility. Based on these studies, the SEC removed short-sale restrictions for all firms when the experiment ended in 2007.

Since then, an array of papers have documented a wide range of indirect effects of the experiment on pilot firms. These studies find that pilot firms reduced investment, earnings management, changed their compensation plans etc. Some papers report evidence that third parties such as auditors and analysts changed their behavior in response to the experiment.

In this paper, we first reassess the evidence for the most commonly asserted causal channels and do not find support for them. We then reexamine the evidence for key findings in four papers: FHK, GMW, HHZ, and LLS. Using a pre-specified research design and a sample that closely

follows SEC rules, we find no support for the results in these papers, and only minimal support using best-match samples and specifications. Using the FHK and HHZ exact samples and specifications, their results technically replicate, but are fragile.

An important takeaway from our analysis is that even when researchers begin with a randomized trial, they must make many specification choices. These decisions offer opportunities for one specification choice to produce significant results, when other reasonable choices would not.

## References


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 (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Angrist, Joshua D., Guido W. Imbens, and and Donald B. Rubin (1996), Identification of Causal Effects Using Instrumental Variables, *Journal of the American Statistical Association* 91: 444-455.

Armstrong, Chris, David F. Larcker, Gaizka Ormazabal and Daniel J. Taylor (2013), The Relation Between Equity Incentives and Misreporting: The Role of Risk-Taking Incentives, *Journal of Financial Economics* 109: 327-350.

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.

Barclay, Michael J., Comment on Macey, Miller and Netter, *Cornell Law* Review 74, 836-840.

Battalio, Robert, and Paul Schultz (2006), Options and the Bubbler, *Journal of Finance* 61(5), 2071-2102.

Benjamin, Daniel, James O. Berger, Magnus Johannesson, . . . and Valen E. Johnson (2018), Redefine Statistical Significance, *Nature Human Behavior* 2 (Jan.), 6-10.

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, Alex Hollingsworth, Leticia Nunes and Kosali Simon (2021), The Effect of Health Insurance on Mortality: Statistical Power and What We Can Learn from the Affordable Care Act Coverage Expansions, NBER working paper 25,568, at http://ssrn.com/abstract=3336520.

Bodie, Zvi, Robsert S. Kaplan, and Robert C. Merton (2003), For the Last Time: Stock Options are an Expense, *Harvard Business Review* (March).

Bond, Philip, Alex Edmans, and itay Goldstein (2012), The Real Effects of Financial Markets, *Annual Review of Financial Economics* 4, 339-360.

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 Marcel Kahan (2016). The Law and Finance of Antitakeover Statutes. *Stanford Law Review*, 68, 629-680.

Chen, Hang, Yushu Zhu, and Liang Chang (2019), Short-selling constraints and corporate payout policy, *Accounting & Finance* 59(4), 2273-2305.

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 (2006), 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.

Donelson, D.C., Kettell, L., McInnis, J.M., and Toynbee, S., 2021. The Need to Validate Exogenous Shocks: Shareholder Derivative Litigation, Universal Demand Laws and Firm Behavior. *Journal of Accounting and Economics*, forthcoming.

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. (2019), at http://ssrn.com/abstract=3507033.

Gelman, Andrew, and John Carlin (2014), Beyond Power Calculations: Assessing Type S (Sign) and Type M

(Magnitude) Errors, *Perspectives on Psychological Science* 9(6), 641–651.

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

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(2): 519-543.

Harvey, Campbell R. (2014), Reflections on Editing the Journal of Finance, 2006-2012, *Secrets of Economics Editors* 67-82 (MIT Press).

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

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

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

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

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

Heath, Davidson, Matthew C. Ringgenberg, Mehrdad Samadi, and Ingrid M. Werner (2020), Reusing Natural Experiments, working paper, at http://ssrn.com/abstract=3457525.

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* (Cambridge University Press).

Ioannidis, John P.A. (2005), Why Most Published Research Findings Are False. PLoS Medicine, 2(8): e124.

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

Kaplan, Robert M., and Veronica L. Irvin (2015), Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. PLoS ONE 10(8): e0132382. doi:10.1371/journal.pone.0132382.

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

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.

Leamer, Edward E. (1985), Sensitivity Analyses Would Help, *American Economic Review* 75 (3), 308-313.

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 and Substantive Short Selling: Evidence on Causal Channels and Experiment Design, working paper.

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.

McCrary, Justin, Garret Christensen, and Daniele Fanelli (2016), Conservative Tests under Satisficing Models of Publication Bias, *PLOS One* 11(2): e0149590. doi:10.1371/journal.pone.0149590.

Morck, Randall, Andrei Shleifer, and Robert Vishny (1990), The Stock Market and Investment: Is the Market a Sideshow? *Brookings Papers on Economic Activity* 1990(2): 157-215.

Olken, Benjamin A. (2015), Promise and Perils of Pre-analysis Plans, *Journal of Economic Perspectives* 29 (3), 61-80.

Open Science Collaboration (2015), Estimating the Reproducibility of Psychological Science, *Science* 349, DOI: 10.1126/science.aac4716.

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

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

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

Stock, J., J. Wright, and M. Yogo (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business and Economic Statistics* 20: 518-529.

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. Summary of the Re-examined Papers

In July 2004 the SEC announced a randomized trial in which it temporarily suspended short-sale restrictions, contained in Regulation SHO, for approximately 1,000 ("pilot" or "treated") firms in the Russell 3000 Index. The SEC suspended the NYSE uptick rule and the similar Nasdaq bid test for pilot firms, but left in place some but not all of the prior restrictions for control firms. We reexamine the evidence for causal channels and key results from four recent papers that report indirect effects of the experiment. All references to our results are to results for pilot firms relative to control firms, using pre-specified sample and research design.

| Paper | Conjecture and Main Result | Posited Causal Channel | Plausibility of Causal Channel | Our Evidence |
|---|---|---|---|---|
| FHK | Pilot firms reduced earnings management, measured using performance-matched discretionary accruals (PMDA). | Pilot firms' managers feared being targeted by short sellers and hence reduce earnings management. This would reduce the likelihood of a future restatement. | No evidence of direct effect of the experiment on short interest or share returns, or manager fear channel). Absent support for any of these three causal channels, unclear why managers would take costly action to reduce earnings. Also unclear why their actions would affect only top 1% of F-scores. | No significant change in four accruals measures (operating, total and abnormal accruals, and PMDA) with either balanced or unbalanced sample. Even with their exact sample and specification, no significant results if one selects matching based on current ROA instead of lagged ROA. |
| | Pilot firms had a lower likelihood of a very high "F-score" (measure of likelihood of an accounting misstatement). | | | No significant change in F-score or likelihood of high F-score. |
| GMW | Short-sale constraints result in overvaluation which leads to overinvestment. Thus, removing constraints should reduce share prices and investment. Prices of small pilot firms (below sample median in assets) fell two weeks before SEC experiment *announcement*. Small pilot firms reduced investment and raised less capital. | Lower share prices lead to lower investment, lower growth, and hence less need to raise capital. | Drop in price of small firms before the list of pilot firms was made public is implausible, and is sensitive to sample choice. Also unclear why a small price drop in 2004, even if it occurred, would lead to major change in investment over 2005-2007. | No significant change in any of GMW outcomes or related outcomes. |
| HHZ | Increased short-selling increases auditor securities litigation risk. Auditors respond to litigation risk by increasing audit fees. | Auditor fear that short-sellers will drive down share prices, leading to of increased risk of litigation. which causes auditors to increase audit fees. | No evidence of greater short selling, negative returns for pilot firms, or manager (or auditor) concern about the experiment. Authors do not study whether litigation increased. | Following their exact sample selection steps we end up with a larger sample and find no significant change in audit fees. |
| LLS | Pilot firms share prices become more informative due to reduced short-sale constraints. Shareholders and boards perceive lower need for direct incentives, so CEO wealth sensitivity to performance falls. Managers rely more heavily on prices to guide business decisions, so investment sensitivity to Tobin's *q* rises. | Fewer short-sale constraints lead to more informative prices; boards and managers must believe prices are more informative. | The authors do not assess whether prices become more informative, and prior studies find minor effects of the experiment on price efficiency. | No significant change in any of the LLS outcomes. |

## Table 2. Variables, Summary Statistics, and Balance on Covariates and Outcomes

### Panel A: Variable Definitions

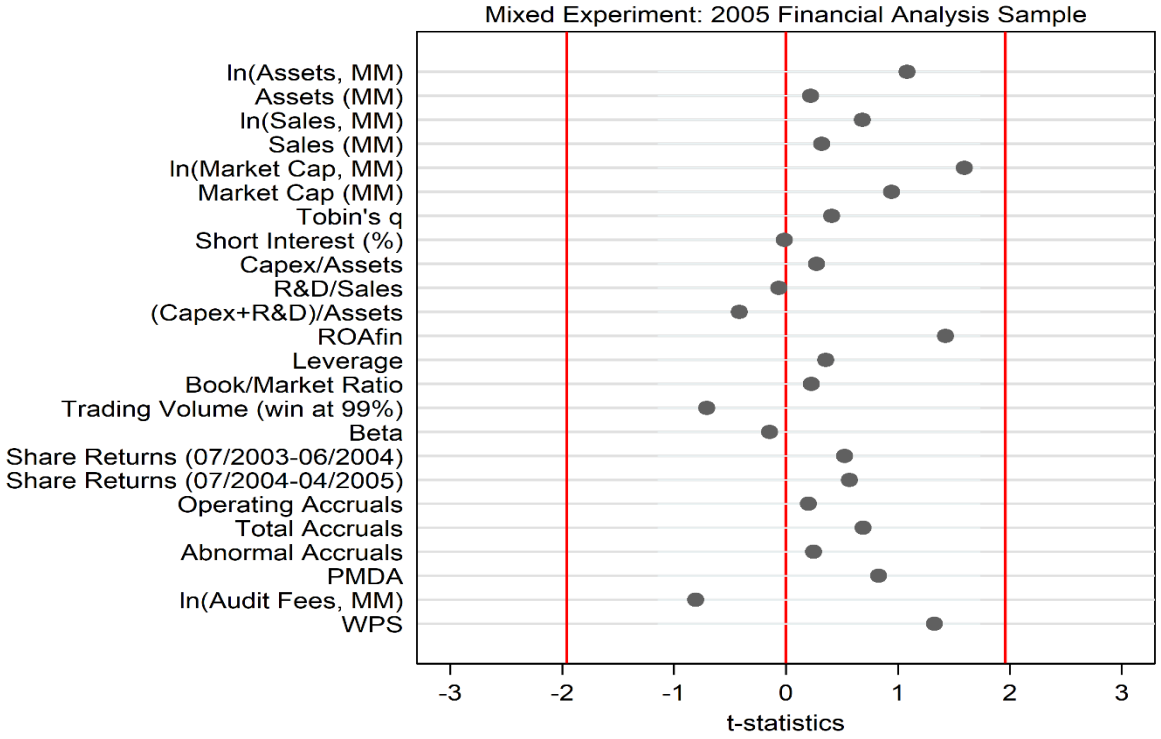
Balance sheet and income statement values are from Compustat Annual. Except as specified below, years are fiscal years (using the Compustat convention for mapping fiscal years to years) and variables are winsorized at 1% and 99% across all fiscal years. Compustat variable names are indicated in the table.

| Variables | Definitions |
|---|---|
| Assets ($M) | Total Assets (AT). |
| Asset Growth | Fractional change in assets $((AT_t - AT_{t-1})/AT_{t-1.}$ |
| Sales ($M) | Net Sales (SALE). |
| Market Cap ($M) | Market Capitalization (PRC*SHROUT/1000). Based on price and shares outstanding from CRSP as of April 29, 2005 for 2005 Analysis Sample. |
| Q | Tobin's *q*, defined as total assets (AT) minus book value of common equity (CEQ) and deferred tax (TXDB), plus market capitalization, scaled by total assets. (AT-CEQ-TXDB+(PRCC_F*CSHO)/AT). |
| Short Interest (% of Shares Outstanding) | monthly open short interest reported on 15th of each month (from Compustat) scaled by shares outstanding at the start of the month (from CRSP)*100 (100*SHORTINT/(SHROUT*1000)). We measure average *% short interest* during 12 months from May 2004 to April 2005 (for covariate balance table). HHZ, in their full set of covariates, measure short interest during last month of year *t*-1. Values above 100% are treated as missing. |
| Capex/Assets | Capital Expenditures scaled by lagged Total Assets ($CAPX_t/AT_{t-1}$). |
| R&D/Sales | R&D scaled by Net Sales ($XRD_t/SALE_t$). Missing R&D is replaced with 0 and negative Net Sales are replaced with 0. |
| R&D/Assets | (Used in LLS). R&D scaled by lagged Total Assets ($XRD_t/AT_{t-1}$). Missing R&D is replaced with 0. Winsorized at 99%. |
| (Capex + R&D)/Assets | Capital Expenditures plus R&D, scaled by lagged total assets ($(CAPX_t + XRD_t)/AT_{t-1}$). Missing R&D is replaced with 0. Winsorized at 99%. |
| $ROA_{fin}$ (finance definition) | Return on Assets, defined as operating income before depreciation and amortization scaled by lagged total assets ($OIBDP_t/AT_{t-1}$). Used by FHK and GMW as covariate. |
| ROA (as outcome) | Income before extraordinary items from cash flow statement/lagged assets ($IBC_t/AT_{t-1}$). We use IBC (rather than IB, which is very similar, and identical for 95% of firms) because we also use IBC to measure operating and total accruals. |
| ROA (Kothari) | Used to find matching firm for PMDA: net income over current total assets ($NI_t/AT_t$) |
| ROA (GMW) | (Used in GMW, who call this "cash flow"): sum of income before extraordinary items and depreciation and amortization expenses scaled by lagged total assets ($IB_t + DP_t)/AT_{t-1}$). |
| ROA (FHK matching) | The measure actually used by FHK to find a matching firm for PMDA, for their 2012 PMDA Specification: Prior year income before extraordinary items /prior year total assets ($IB_{t-1}/AT_{t-1}$). FHK use a different ROA measure, defined above, as a covariate in regressions. LLS use this measure, non-lagged, as a covariate. |
| Leverage | Total Debt/Total Assets ((DLC+DLTT)/(DLC+DLTT+SEQ)). |
| Book/Market | Book-to-Market Ratio (CEQ/(CSHO*PRCC_F)). |
| Trading Volume | Average *fractional trading volume* during 12 months from May 2004 to April 2005, defined as monthly trading volume (from CRSP) scaled by shares outstanding at the end of the month (from CRSP) (100*VOL/(SHROUT*1000)). Winsorized at 99%. |
| Beta | Beta from regression of daily return (RET) on market value weighted return from CRSP (VWRETD) over 250 trading days preceding May 2, 2005. |
| Share Returns | $\prod_i(1 + RET_i) - 1$, where *i* includes 12 months from July 2003 to June 2004 (for 12-month pre-announcement period) or 10 months from July 2004 to April 2005 (for 10-month period between experiment announcement and experiment launch). |
| Operating Accruals | Operating Accruals, defined as Earnings Before Extraordinary items on the cash flow statement (IBC), minus operating cash flows (OANCF) before extraordinary items and discontinued operations (XIDOC), scaled by beginning-of-the-year total assets. ((IBC-(OANCF-XIDOC))/$AT_{t-1}$). We replace missing XIDOC with zero. |

| | |
|---|---|
| Total Accruals | Total Accruals, defined as Earnings Before Extraordinary items on the cash flow statement (IBC), minus operating cash flows (OANCF) before extraordinary items and discontinued operations (XIDOC), minus investing cash flow (IVNCF), scaled by beginning-of-the-year total assets. ((IBC-(OANCF-XIDOC)-IVNCF)/$AT_{t-1}$). We replace missing XIDOC and IVNCF with zero. |
| AA | Abnormal accruals, measured using the modified Jones model, as described in the text. |
| PMDA | Performance-matched discretionary accruals, measured as described in the text. |
| Audit Fees | Audit fees, from Audit Analytics, mapped to Compustat fiscal years. |
| Equity Issues | (Used in GMW) 100 × Sale of Common and Preferred Stock scaled by lagged total assets ($SSTK_t/AT_{t-1}$). |
| Debt Issues | (Used in GMW) 100 × Long-Term Debt Issuance scaled by lagged total assets ($DLTIS_t/AT_{t-1}$). |
| WPS | Wealth-performance sensitivity, as defined by LLS. |
| Additional HHZ Covariates | (used in one HHZ model): ratio of current to total assets; quick ratio, ratio of (inventory + receivables) to total assets, asset growth, ln(1 + no. of business segments), short interest in last month of prior fiscal year, and dummy variables for net income < 0, fiscal year not ending in December, firm has big four auditor, and firm has foreign operations. |
| LLS add'l covariates (WPS as outcome) | dividend dummy (=1 if firm pays any dividend), Firm age (based on the first year the firm appears in Compustat), ownership by all institutional investors, ownership by five largest institutional investors, the ratio of cash to assets, ratio of investment-to-capital (CAPX/PPENT), standard deviation of monthly stock returns (over what period is not stated). Note that LLS made a mistake in their code in measuring cash/assets. Instead of using cash(Compustat data element 162), they use Deferred Tax(data element 126). |

## Panel B: Graphical Evidence on Covariate Balance

Figure provides a graphical overview of covariate balance. Except as stated in Panel A, we use the most recent annual Compustat data date before May 2005, and all variables are winsorized at the 1% and 99% levels. Figure shows *t*-statistics for differences between pilot and original control firms, for the variables listed in Panel A. Vertical lines indicate *t*-statistics of -1.96, 0, and +1.96.



## Table 3. DiD Regressions for Mid-Month Short Interest

Regressions of mid-month short interest (as % of outstanding shares), with firm and month FE, over July 2003-December 2007, for indicated samples, on pre-launch dummy (=1 for August 2004 through April 2005), experiment dummy (=1 for May 2005-June 2007), pre-launch* pilot and experiment*pilot interaction terms, and constant term. *t*-statistics, with standard errors clustered on firm, in parentheses. *, **, *** indicates statistical significance at the 10%, 5%, and 1% levels, respectively. Significant results, at 5% or better, in **boldface**.

| | **2004 Announcement** | | | **GMW Best Match** | | | |
|---|---|---|---|---|---|---|---|
| **Sample** | **All firms** | **Small Firms (Based on Market Cap)** | **2005 Analysis** | **All Firms** | **Small Firms (Based on Assets)** | **Small Firms (Based on Market Cap)** | **Small Firms (Based on Trading Volume)** |
| | (1) | | (2) | (3) | (4) | (5) | (6) |
| Pre-launch × Pilot | -0.022 | 0.098 | 0.089 | 0.088 | 0.248 | 0.331 | -0.041 |
| | (-0.19) | (0.51) | (0.61) | (0.59) | (1.08) | (1.32) | (-0.25) |
| Experiment × Pilot | 0.038 | 0.251 | 0.141 | 0.102 | 0.408 | 0.323 | -0.114 |
| | (0.27) | (1.04) | (0.79) | (0.55) | (1.26) | (0.99) | (-0.63) |
| Pre-launch dummy | **0.814*** ** | **1.474*** ** | **0.843*** ** | **0.816*** ** | **1.270*** ** | **1.665*** ** | **0.876*** ** |
| | **(7.87)** | **(9.11)** | **(6.38)** | **(5.97)** | **(5.65)** | **(7.81)** | **(6.06)** |
| Experiment dummy | **4.014*** ** | **5.633*** ** | **3.960*** ** | **4.237*** ** | **5.879*** ** | **6.502*** ** | **4.709*** ** |
| | **(25.25)** | **(22.63)** | **(20.57)** | **(21.62)** | **(19.35)** | **(21.39)** | **(20.14)** |
| Constant | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| $R^2$ | 0.670 | 0.671 | 0.669 | 0.679 | 0.677 | 0.672 | 0.651 |
| Obs. | 129,284 | 64,746 | 94,280 | 86,956 | 44,736 | 43,519 | 44,384 |
| Pilot (Control) Firms | 875 (1,735) | 428 (897) | 614 (1,237) | 564 (1,133) | 278 (603) | 266 (591) | 288 (570) |
| No. of Firms | 2,610 | 1,325 | 1,851 | 1,697 | 881 | 857 | 858 |

## Table 4. Buy-and-Hold Relative Returns Around SEC Experiment Announcement

Buy-and-hold relative returns (BHRRs) for pilot firms relative to control firms for indicated samples over event windows covering trading days [-10, +1] and [-1, +1] relative to SEC announcement on July 28, 2004, plus the announcement date (day 0), for the indicated samples. *t*-statistics (in parentheses) are for two-sample difference in means. Bottom two rows report GMW's reported BHRR's (they do not report statistical significance). *, **, *** indicates significance at the 10%, 5% and 1% levels, respectively. Significant results at 5% level in **boldface.**

| **Event Window** | **[-10, +1]** | **[-1, +1]** | **Day 0** |
|---|---|---|---|
| **Sample** | | | |
| 2004 Announcement, all firms | -0.0060<br>(1.75) | -0.0011<br>(0.56) | 0.0011<br>(0.98) |
| 2004 Announcement, small firms (based on market cap) | -0.0067<br>(1.16) | -0.0055<br>(1.61) | -0.0001<br>(0.07) |
| GMW Best-Match small firms (based on assets) | **-0.0155**<br>**(2.02)** | -0.0042<br>(1.01) | 0.0000<br>(0.00) |
| GMW Best-Match small firms (based on market cap) | -0.0096<br>(1.27) | -0.0032<br>(0.78) | 0.0001<br>(0.06) |
| GMW Best-Match small firms (based on trading volume) | -0.0006<br>(0.10) | -0.0022<br>(0.69) | 0.0015<br>(0.85) |
| GMW Best-Match small NYSE firms (based on assets) | 0.0044<br>(0.41) | 0.0050<br>(0.80) | 0.0000<br>(0.00) |
| GMW Best-Match small Nasdaq firms (based on assets) | **-0.0224**<br>**(2.31)** | -0.0077<br>(1.50) | 0.0001<br>(0.04) |
| GMW reported (all firms) | -0.0152 | -0.0017 | 0.0004 |
| GMW reported (small firms, based on assets) | -0.0235 | -0.0092 | -0.0021 |

## Table 5. (FHK) Accruals Measures

**Panel A.** Regressions, using our specification, with firm and fiscal year FE, of indicated accruals measures on Pilot*During, Pilot*Post, and constant term, over fiscal years 2001-2010, following equation (4) in the text. During and Post periods are defined in the text. **Panel B.** Regressions are similar to Panel A, but use FHK best-match specification. Two-way clustering is on firm and calendar year, using cluster2.ado. **Panel C.** FHK reported results (their Table III, model (1)). **All Panels.** Sign reversal = coeff. on (Pilot*Post) minus coeff. on (Pilot*During), using equation (5). Coefficient on constant term is suppressed. Unbalanced panel uses all firm-year observations with data to calculate indicated outcome. Balanced panel requires indicated outcome to be non-missing for all sample years. Covariates (must be non-missing for balanced panel but not included in regressions) are *ln(Total Assets);Market-to-book ratio; ROA; and Leverage*. Randomization inference standard deviations (s.d.'s), based on 1,000 repetitions, in parentheses. Standard errors (*s.e.*'s) with indicated clustering in brackets. *, **, *** indicates significance at the 10%, 5%, and 1% level, respectively. Significant results, at 5% level or better, in **boldface.**

| | FHK | Unbalanced Panel | | | | Balanced Panel | | | |
|---|---|---|---|---|---|---|---|---|---|
| | predicted | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| **Accruals type** | sign | **Operating** | **Total** | **AA** | **PMDA** | **Operating** | **Total** | **AA** | **PMDA** |
| **Panel A. Our sample and specification, with firm and fiscal year FE** | | | | | | | | | |
| ***Pilot*During*** | | -0.0024 | -0.0000 | -0.0020 | 0.0029 | -0.0037 | 0.0035 | -0.0030 | 0.0075 |
| *s.d. (rand. inf.)* | negative | (0.0034) | (0.0072) | (0.0050) | (0.0079) | (0.0035) | (0.0078) | (0.0054) | (0.0087) |
| *s.e. (cluster on firm)* | | [0.0035] | [0.0076] | [0.0050] | [0.0080] | [0.0036] | [0.0079] | [0.0054] | [0.0087] |
| ***Pilot*Post*** | | -0.0014 | -0.0052 | -0.0022 | -0.0037 | -0.0032 | -0.0051 | -0.0041 | -0.0007 |
| *s.d. (rand. inf.)* | near zero | (0.0040) | (0.0072) | (0.0055) | (0.0081) | (0.0041) | (0.0072) | (0.0056) | (0.0083) |
| *s.e. (cluster on firm)* | | [0.0039] | [0.0073] | [0.0055] | [0.0081] | [0.0040] | [0.0072] | [0.0057] | [0.0084] |
| **Sign Reversal** | positive | 0.0009 | -0.0052 | -0.0002 | -0.0066 | 0.0005 | -0.0086 | -0.0011 | -0.0081 |
| *s.e. (cluster on firm)* | | (0.0038) | (0.0078) | (0.0050) | (0.0082) | (0.0039) | (0.0079) | (0.0049) | (0.0083) |
| Firm-Year Obs. | | 18,785 | | 18,375 | | 14,640 | | 13,980 | |
| Pilot (Control) Firms | | 702 (1413) | | 698 (1401) | | 517 (947) | | 492 (906) | |
| $R^2$ (within) | | 3.9% | 2.9% | 1.1% | 0.1% | 3.9% | 3.0% | 1.2% | 0.1% |
| **Panel B. FHK Best Match (no firm or year FE)** | | | | | | | | | |
| ***Pilot * During*** | | -0.0047 | -0.0033 | -0.0030 | 0.0038 | -0.0068 | 0.0027 | -0.0073 | -0.0022 |
| *s.d. (rand. inf.)* | negative | (0.0039) | (0.0076) | (0.0055) | (0.0084) | (0.0040) | (0.0081) | (0.0063) | (0.0094) |
| *s.e. (2-way cluster)* | | [0.0034] | [0.0064] | [0.0045] | . | [0.0048] | [0.0094] | [0.0064] | [0.0030] |
| ***Pilot * Post*** | | -0.0011 | -0.0099 | 0.0011 | 0.0018 | -0.0055 | -0.0068 | -0.0061 | -0.0079 |
| *s.d. (rand. inf.)* | near zero | (0.0043) | (0.0076) | (0.0061) | (0.0087) | (0.0044) | (0.0080) | (0.0063) | (0.0090) |
| *s.e. (2-way cluster)* | | [0.0036] | [0.0087] | [0.0052] | [0.0084] | [0.0030]* | [0.0069] | [0.0060] | [0.0086] |
| **Sign reversal** | positive | 0.0036 | -0.0066 | 0.0041 | -0.0020 | 0.0013 | -0.0095 | 0.0012 | -0.0057 |
| *s.e (cluster on firm)* | | [0.0040] | [0.0075] | [0.0052] | [0.0086] | [0.0039] | [0.0077] | [0.0054] | [0.0091] |
| Firm-Year Obs. | | 16,413 | | 16,074 | | 12,321 | | 11,808 | |
| Pilot (Control) Firms | | 701 (1,412) | | 697 (1,400) | | 489 (880) | | 466 (846) | |
| Adj. $R^2$ | | 1.2% | 1.2% | 0.5% | -0.0% | 1.4% | 1.4% | 0.3% | -0.0% |
| **Panel C. FHK Reported** | | | | | | | | | |
| ***Pilot * During*** | | | | | | | | | **-0.010**** |
| *s.e. (2-way cluster)* | | | | | | | | | **[0.004]** |
| ***Pilot * Post*** | | | | | | | | | 0.004 |
| *s.e. (2-way cluster)* | | | | | | | | | [0.004] |
| Firm-Year Obs. | | | | | | | | | 9,873 |
| Pilot (Control) Firms | | | | | | | | | 388 (709) |
| Adj. $R^2$ | | | | | | | | | 0.10% |

## Table 6. (FHK) F-score and HF-score

**Panel A.** Regressions, using our specification, of F-score and HF-score on Pilot*During, Pilot*Post, and constant term. During and Post periods are defined in the text. Columns (1)-(6): Regressions, with firm and fiscal year FE, of indicated F-score measures, calculated using coefficient estimates from Dechow et al. (2011, pp. 60-61), Models (1)-(3). Columns (7)-(12): average marginal effects from probit regressions, with fiscal year FE, of indicated HF-score measure. HF-1 is a dummy variable that equals one if F-1 score $\geq 99^{th}$ percentile of the sample across all sample years, zero otherwise;; HF-2 and HF-3 are similar defined based on F-2 and F-3. Unbalanced panel uses all firm-year observations with data to calculate indicated outcome. Balanced panel requires indicated outcome to be non-missing for all sample years. Randomization inference s.d.'s are in parentheses. $R^2$ is within for F-score, and Pseudo $R^2$ for HF-score. **Panel B.** Regressions, similar to Panel A, but using FHK best-match specification. Two-way clustering is on firm and calendar year, using cluster2.ado. Coefficients on constant term, Pilot, During, and Post dummies are suppressed. For HF-score we report probit coefficients for better comparability to FHK. **Panel C.** FHK results as reported (their Table VI). **All Panels.** Odd (even)-numbered regressions use unbalanced (balanced) panel. s.e.'s with indicated clustering are in brackets. *, **, *** indicates significance at the 10%, 5%, and 1% level, respectively. Significant results, at 5% level or better, in **boldface.**

| | **FHK** | **F-1** | | **F-2** | | **F-3** | | **HF-1** | | **HF-2** | | **HF-3** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | predicted | **(1)** | **(2)** | **(3)** | **(4)** | **(5)** | **(6)** | **(7)** | **(8)** | **(9)** | **(10)** | **(11)** | **(12)** |
| | sign | *Unbal.* | *Balanced* | *Unbal.* | *Balanced* | *Unbal.* | *Balanced* | *Unbal.* | *Balanced* | *Unbal.* | *Balanced* | *Unbal.* | *Balanced* |
| **Panel A: Our Sample and Specification** | | | | | | | | **Marginal Effects** | | | | | |
| ***Pilot*During*** | negative | 0.0096 | 0.0020 | 0.0106 | 0.0071 | 0.0047 | 0.0240 | 0.0004 | -0.0023 | 0.0011 | -0.0009 | -0.0007 | 0.0041 |
| *s.d. (rand. inf.)* | | (0.0179) | (0.0185) | (0.0190) | (0.0204) | (0.0250) | (0.0243) | (0.0034) | (0.0037) | (0.0033) | (0.0038) | (0.0037) | (0.0044) |
| *s.e.[cluster on firm]* | | [0.0167] | [0.0182] | [0.0185] | [0.0205] | [0.0251] | [0.0235] | [0.0034] | [0.0036] | [0.0034] | [0.0037] | [0.0036] | [0.0044] |
| ***Pilot*Post*** | near zero | -0.0060 | -0.0144 | -0.0034 | -0.0056 | -0.0193 | 0.0027 | 0.0061 | 0.0004 | 0.0016 | 0.0025 | -0.0003 | 0.0060 |
| *s.d. (rand. inf.)* | | (0.0194) | (0.0183) | (0.0207) | (0.0208) | (0.0265) | (0.0263) | (0.0048) | (0.0049) | (0.0049) | (0.0048) | (0.0053) | (0.0052) |
| *s.e [cluster on firm]* | | [0.0186] | [0.0197] | [0.0201] | [0.0217] | [0.0275] | [0.0259] | [0.0043] | (0.0044) | [0.0041] | [0.0043] | [0.0047] | [0.0048] |
| Sign Reversal | positive | -0.0156 | -0.0164 | -0.0140 | -0.0128 | -0.0240 | -0.0213 | 0.0058 | 0.0027 | 0.0005 | 0.0034 | 0.0004 | 0.0019 |
| (Post – During) | | (0.0163) | (0.0171) | (0.0176) | (0.0191) | (0.0198) | (0.0214) | (0.0042) | (0.0042) | (0.0041) | (0.0042) | (0.0049) | (0.0045) |
| Firm FE | | Yes | Yes | Yes | Yes | Yes | Yes | No | No | No | No | No | No |
| Fiscal Year FE | | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm-Year Obs. | | 17,996 | 13,550 | 17,739 | 12,880 | 17,462 | 12,360 | 17,996 | 13,550 | 17,739 | 12,880 | 17,462 | 12,360 |
| Pilot (Control) Firms | | 687 (1,380) | 477 (878) | 687 (1,380) | 455 (833) | 686 (1,378) | 436 (800) | 687 (1,380) | 477 (878) | 687 (1,380) | 455 (833) | 686 (1,378) | 436 (800) |
| $R^2$ | | 3.5% | 4.1% | 3.5% | 3.9% | 2.0% | 2.0% | 2.4% | 2.2% | 2.6% | 2.6% | 2.3% | 1.8% |
| **Panel B. FHK Best Match** | | | | | | | | **Probit Coefficients** | | | | | |
| ***Pilot*During*** | | -0.005 | 0.007 | -0.001 | 0.011 | -0.001 | 0.022 | -0.087 | -0.207 | -0.162 | -0.194 | 0.011 | 0.052 |
| *s.d. (rand. inf.)* | | (0.0194) | (0.0208) | (0.0209) | (0.0226) | (0.0250) | (0.0258) | (0.1460) | (0.1706) | (0.1471) | (0.1724) | (0.1484) | (0.1734) |
| *s.e. [2-way cluster]* | | [0.0145] | [0.0165] | [0.0160] | [0.0192] | [0.0177] | [0.0245] | [0.0972] | **[0.0399]*** ** | **[0.0674]** ** | **[0.0938]** ** | [0.1056] | [0.1198] |
| ***Pilot*Post*** | | -0.021 | -0.012 | -0.016 | -0.004 | -0.022 | -0.004 | -0.115 | -0.179 | -0.072 | -0.038 | -0.047 | 0.083 |
| *s.d. [rand. inf.]* | | (0.0216) | (0.0220) | (0.0231) | (0.0238) | (0.0279) | (0.0282) | (0.2027) | (0.2199) | (0.2036) | (0.2238) | (0.2098) | (0.2415) |
| *s.e. [2-way cluster]* | | **[0.0092]** ** | [0.0105] | [0.0084]* | [0.0111] | [0.0141] | [0.0191] | [0.1873] | [0.1721] | [0.0766] | [0.0996] | [0.1217] | [0.1464] |
| Sign reversal | | -0.017 | -0.019 | -0.014 | -0.015 | -0.021 | -0.027 | -0.028 | 0.027 | 0.090 | 0.157 | -0.057 | 0.031 |
| *s.e [cluster on firm]* | | [0.0175] | [0.0179] | [0.0188] | [0.0193] | [0.0208] | [0.0213] | [0.1713] | [0.1710] | [0.1647] | [0.1608] | [0.1779] | [0.1956] |
| Firm-Year Obs. | | 15,499 | 10,890 | 15,499 | 10,890 | 15,499 | 10,890 | 15,499 | 10,890 | 15,499 | 10,890 | 15,499 | 10,890 |
| Pilot (Control) Firms | | 686 (1376) | 431 (779) | 686 (1376) | 431 (779) | 686 (1376) | 431 (779) | 686 (1376) | 431 (779) | 686 (1376) | 431 (779) | 686 (1376) | 431 (779) |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $R^2$ | 0.9% | 1.0% | 0.9% | 1.0% | 0.4% | 0.4% | 1.5% | 1.6% | 1.6% | 1.7% | 1.3% | 1.2% |
| **Panel C. FHK Reported** | | | | | | | | | | | | |
| ***Pilot*During*** | | | | | | | | **-0.178**** | | **-0.189**** | | **-0.200**** |
| *s.e. [2-way cluster]* | | | | | | | | **[0.080]** | | **[0.079]** | | **[0.080]** |
| ***Pilot*Post*** | | | | | | | | **-0.177**** | | **-0.186**** | | **-0.169**** |
| *s.e. [2-way cluster]* | | | | | | | | **[0.087]** | | **[0.087]** | | **[0.086]** |
| Firm-Year Obs. | | | | | | | | 9,871 | | 9,871 | | 9,871 |
| $R^2$ | | | | | | | | 11.5% | | 11.0% | | 10.8% |

## Table 7. (GMW) Investment and Capital Raising

**Panel A.** Regressions, with firm and fiscal year FE, of indicated dependent variables on Pilot × During, Pilot × Post, and constant term, using our specification and small firm sample (using GMW definition, assets, measured in most recent fiscal year ending prior to July 28, 2004, below sample median). During is a dummy variable for the experiment period; Post is a dummy variable for the post-experiment period, Detailed definition of periods is provided the text. **Panel B.** Same, but we switch to the GMW best-match specification. **Panel C.** GMW univariate results for small firms from their Table 5. **Panel D**. Similar to Panel B but includes the GMW covariates (with ln(assets) as the outcome, we do not control for lagged ln(assets)). **Panel E**. GMW reported results for small firms with covariates, from their Table 6. **All panels**. Coefficient on constant term is suppressed. We follow GMW and multiply by 100 to convert amounts to percentages, except for *ln*(assets). *t*-statistics, with standard errors clustered on firm, in parentheses. *, **, *** indicates significance at the 10%, 5%, and 1% level, respectively. Significant results, at 5% level or better, in **boldface.**

| **Dep. Variable (%)** | **Predicted Sign** | **CAPEX/Assets** | **R&D/Sales** | **(CAPEX + R&D)/Assets** | **Percent asset growth** | ***ln*(assets)** | **Equity Issues** | **Debt Issues** |
|---|---|---|---|---|---|---|---|---|
| Reported by GMW | | Yes | No | Yes | Yes | No | Yes | Yes |
| | | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| **Panel A. Small Firms (below median in assets): Our Specification** | | | | | | | | |
| Pilot × During | Negative | -0.496* | 4.443 | -0.743 | -2.423 | -0.0006 | -0.723 | -1.616 |
| | | (1.68) | (0.46) | (1.08) | (0.91) | (-0.02) | (0.46) | (1.15) |
| Pilot × Post | Near 0 (implied) | -0.383 | 13.44 | -0.021 | 1.210 | -0.0571 | 1.249 | 1.753 |
| | | (1.10) | (1.13) | (0.03) | (0.49) | (-0.99) | (0.76) | (1.09) |
| Firm-Year Obs. | | 9,083 | 9,074 | 9,083 | 9,115 | 9,186 | 8,921 | 8,723 |
| Pilot (Control) Firms | | 337 (719) | 338 (714) | 337 (719) | 337 (719) | 338 (719) | 332 (711) | 336 (713) |
| $R^2$ (within) | | 0.050 | 0.006 | 0.023 | 0.044 | 0.267 | 0.034 | 0.008 |
| **Panel B. Small Firms -- GMW Best Match Specification, No Covariates** | | | | | | | | |
| C1. With firm and fiscal year FE | Negative | **-0.644**** | -0.270 | -0.644 | -0.532 | -0.0235 | 0.505 | -1.56 |
| | | **(2.06)** | (0.03) | (1.00) | (0.20) | (-0.76) | (0.37) | (1.12) |
| C2. Without firm or fiscal year FE | Negative | -0.612* | -4.184 | -0.648 | -0.903 | -0.0214 | 0.527 | -1.48 |
| | | (-1.86) | (-0.46) | (-0.94) | (-0.34) | (-0.63) | (0.39) | (-1.02) |
| Pilot (Control) Firms | | 308 (660) | 309 (656) | 308 (660) | 308 (660) | 309 (661) | 302 (652) | 304 (647) |
| **Panel C. Small Firms -- GMW Univariate Results as Reported** | | | | | | | | |
| Pilot × During (univariate, no firm or fiscal year FE) | Negative | **-0.97** | Not studied | -1.05 | **-6.12** | Not studied | -1.61 | -1.80 |
| | | **(2.88)***** | | (1.81)* | **(2.55)**** | | (1.82)* | (1.52) |
| Pilot (Control) Firms | | 313 (629) | | 313 (629) | 306 (622) | | 302 (606) | 296 (603) |

## Table 8. (HHZ) Audit Fees

**Panel A.** Regressions, with firm and fiscal year FE, of *ln*(Audit Fees) on Pilot*During, Pilot*Post, constant term, and indicated covariates. Sample is our 2005 Analysis Sample (unbalanced panel), merged with audit fee data from Audit Analytics. During is a dummy variable for the experiment period; Post is a dummy variable for the post-experiment period, Detailed definition of periods is provided the text. **Panel B.** Regressions similar to Panel A, but using HHZ best-match specification. **Panel C.** HHZ univariate results from their Table 4 (for col. (1)) and multivariate results from their Table 5 (for cols. (3) and (4)). **All panels.** Column (2) uses our preferred firm size control, which is *ln*(Total Assets). Column (3) uses the "limited" covariates specified in HHZ Table 5 model (1) (ln(sales), leverage, book-to-market ratio, and ROA). Column (4) uses full HHZ covariates, from HHZ Table 5, model (2). All continuous variables are winsorized at 1%/99%. Coefficients on covariates and constant term are suppressed. *t*-statistics, using standard errors clustered on firm, in parentheses. *, **, *** indicates significance at the 10%, 5%, and 1% level, respectively. Significant results, at 5% level or better, in **boldface**.

| | **Dependent Variable = *ln*(Audit Fees)** | HHZ predicted sign | (1) | (2) | (3) | (4) |
|---|---|---|---|---|---|---|
| **Panel** | HHZ Covariates<br>*ln*(Assets) as covariate | | No<br>No | No<br>Yes | Limited<br>No | Full<br>No |
| **A. Our specification** | ***Pilot*During*** | Positive | -0.0046<br>(-0.24) | -0.0088<br>(-0.48) | -0.0086<br>(-0.48) | -0.0132<br>(-0.75) |
| | ***Pilot*Post*** | Near zero | -0.0214<br>(-0.85) | -0.0140<br>(-0.65) | -0.0177<br>(-0.85) | -0.0274<br>(-1.34) |
| | Sign reversal | Negative | -0.0167<br>(-1.00) | -0.0052<br>(-0.36) | -0.0090<br>(-0.61) | -0.0141<br>(-0.97) |
| | Firm-Year Obs.<br>Pilot (Control) Firms<br>$R^2$ (within) | | 17,973<br>683 (1,375)<br>0.717 | 17,964<br>683 (1,375)<br>0.756 | 17,651<br>682 (1,368)<br>0.762 | 16,985<br>666 (1,335)<br>0.772 |
| **B. HHZ best-match specification** | ***Pilot*During*** | Positive | 0.0151<br>(0.62) | 0.0107<br>(0.48) | 0.0135<br>(0.62) | 0.0116<br>(0.54) |
| | ***Pilot*Post*** | Near zero | -0.0025<br>(-0.09) | 0.0120<br>(0.50) | 0.0012<br>(0.05) | -0.0051<br>(-0.23) |
| | Firm-Year Obs.<br>Pilot (Control) Firms<br>$R^2$ (within) | | 18,924<br>615 (1,233)<br>0.716 | 18,924<br>615 (1,233)<br>0.774 | 18,924<br>615 (1,233)<br>0.775 | 18,924<br>615 (1,233)<br>0.789 |
| **C. HHZ results as reported** | ***Pilot*During*** | | 0.048<br>(1.09) | Not reported | **0.0476**<br>**(1.99)**** | **0.0465**<br>**(1.98)**** |
| | ***Pilot*Post*** | | -0.014<br>(-0.36) | | 0.0192<br>(0.76) | 0.0146<br>(0.61) |
| | Pilot (Control) Firms<br>Adj. $R^2$ | | 538 (1,072) | | 538 (1,072)<br>0.915 | 538 (1,072)<br>0.920 |

## Table 9. (LLS) Investment to Price and CEO Wealth to Performance Sensitivity

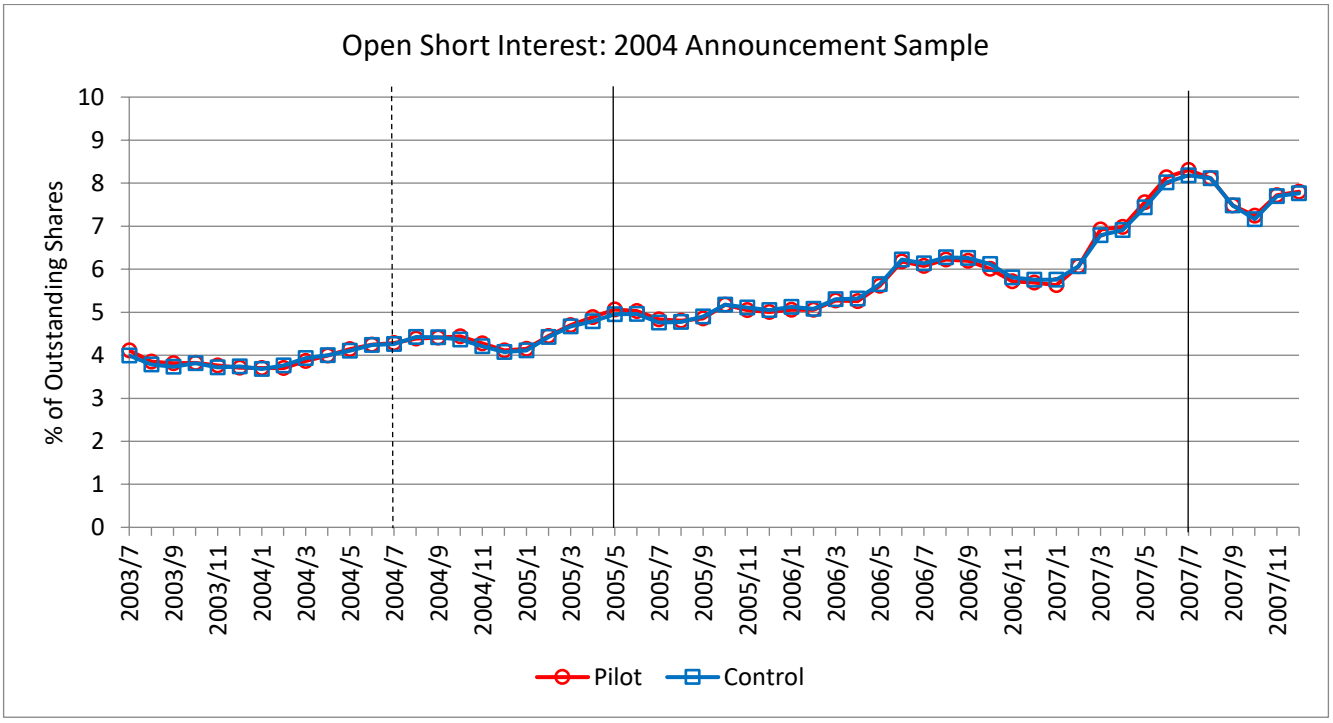
Regressions, with firm and fiscal year FE, of indicated dependent variables (defined in Table 2, Panel A) on indicated variables and constant term. **Panel A.** Uses our specification. All outcomes are winsorized at 1%/99%. During is a dummy variable for the experiment period; Post is a dummy variable for the post-experiment period, Detailed definition of periods is provided the text. **Panel B.** Uses LLS best match specification. Following LLS, we winsorize WPS at 1%/99%, do not winsorize their other outcomes (R&D/assets, and (Capex + R&D)/assets), but do winsorize Capex/assets and R&D/sales. **Panel C.** LLS results as reported. They winsorize WPS at 1%/99%, do not winsorize their other outcomes (R&D/Assets, and (Capex + R&D)/Assets). They report standard errors (s.e.'s), which we show in brackets, and not statistical significance, we infer significance from the ratio of coefficient to standard error. **All panels**. Coefficients of principal interest are the triple interaction terms for Capex, R&D, and Capex + R&D, and in Pilot × During and Pilot × Post for WPS (CEO wealth-performance sensitivity. Coefficients on Q, double interactions (Q × During, Q × Post, Q × Pilot), and constant term are suppressed. *S*tandard errors clustered on firm in parentheses. *, **, *** indicates significance at the 10%, 5%, and 1% level, respectively. Significant results, at 5% level or better, in **boldface.**

| **Dependent variable** | **LLS Predicted Sign** | **Capex/Assets** | **R&D/Sales** | **R&D/Assets** | **(Capex +R&D)/Assets** | **WPS** |
|---|---|---|---|---|---|---|
| | | (1) | (2) | (3) | (4) | (5) |
| **Panel A. Our specification** | | | | | | |
| Q × Pilot × During | Positive | **-0.0046*** ** **(0.0017)** | 0.0338 (0.0300) | -0.0007 (0.0025) | -0.0061* (0.0033) | |
| Pilot × During | Negative for WPS | 0.0066* (0.0037) | -0.0281 (0.0495) | 0.0023 (0.0041) | 0.0097 (0.0061) | -0.0748 (0.0514) |
| Q × Pilot × Post | Negative (implied) | -0.0010 (0.0021) | 0.0463 (0.0775) | 0.0024 (0.0039) | 0.0002 (0.0046) | |
| Pilot × Post | Near 0 (implied) | -0.0009 (0.0043) | -0.0445 (0.1023) | -0.0029 (0.0053) | -0.0016 (0.0074) | -0.0768 (0.0593) |
| Sign Reversal (Post – During) | Neg. (pos for WPS) (implied) | **0.0036** ** **(0.0017)** | 0.0126 (0.0624) | 0.0031 (0.0035) | 0.0063 (0.0043) | -0.0020 (0.0332) |
| Q, double interactions | | Yes | Yes | Yes | Yes | No |
| Firm-Year Obs. | | 17,302 | 17,270 | 17,353 | 17,302 | 12,382 |
| Pilot (Control) Firms | | 694 (1,394) | 694 (1,392) | 694 (1,394) | 694 (1,394) | 517 (999) |
| $R^2$ | | 0.701 | 0.744 | 0.849 | 0.746 | 0.697 |
| **Panel B. LLS best match** | | | | | | |
| Q × Pilot × During | Positive | -0.0006 (0.0016) | 0.0049* (0.0027) | 0.0024 (0.0026) | 0.0022 (0.0029) | |
| Pilot × During | Negative for WPS | 0.0007 (0.0038) | -0.0067 (0.0048) | -0.0018 (0.0044) | -0.0021 (0.0063) | **-0.1292** ** **(0.063)** |
| Firm-Year Obs. | | 8,554 | 8,590 | 8,597 | 8,554 | 9,582 |
| Pilot (Control) Firms | | 637 (1,246) | 637 (1,246) | 637 (1,246) | 637 (1,246) | 672 (1,310) |
| $R^2$ (within) | | 0.811 | 0.934 | 0.836 | 0.772 | 0.725 |
| **Panel C: LLS Results as Reported** | | | | | | |
| Q × Pilot × During | | **Not studied** | **Not studied** | **0.004*** ** **[0.001]** | **0.004** ** **[0.002]** | |
| Pilot × During | | | | -0.006 [0.002] | -0.006 [0.005] | **-0.174** ** **(0.077)** |
| Firm-Year Obs. | | | | 8,307 | 8,267 | 9,400 |
| $R^2$ | | | | 0.921 | 0.845 | 0.709 |

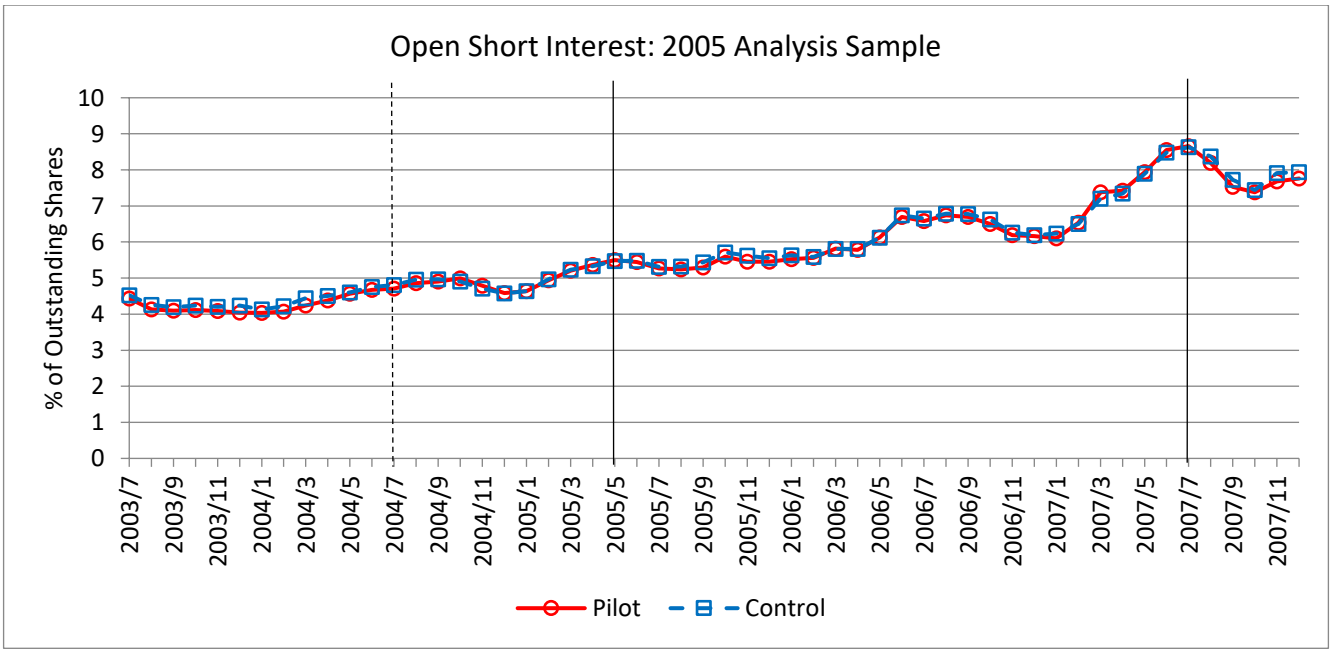
## Figure 1. Monthly Open Short Interest

Mean mid-month short interest, as percentage of outstanding shares, over July 2003-December 2007, for pilot firms and control firms for the 2004 Announcement Sample (**Panel A**); 2005 Analysis Sample **(Panel B)**, GMW Best-Match Sample (**Panel C**), and GMW Best-Match Sample, (small firms; below-median assets) (**Panel D**). SEC experiment was announced on July 28, 2004 and ran from May 2, 2005 through July 5, 2007. Dotted vertical line shows announcement date; solid vertical lines separate pre-announcement, pre-launch, and experiment period. Not all firms have data on short interest in each month.

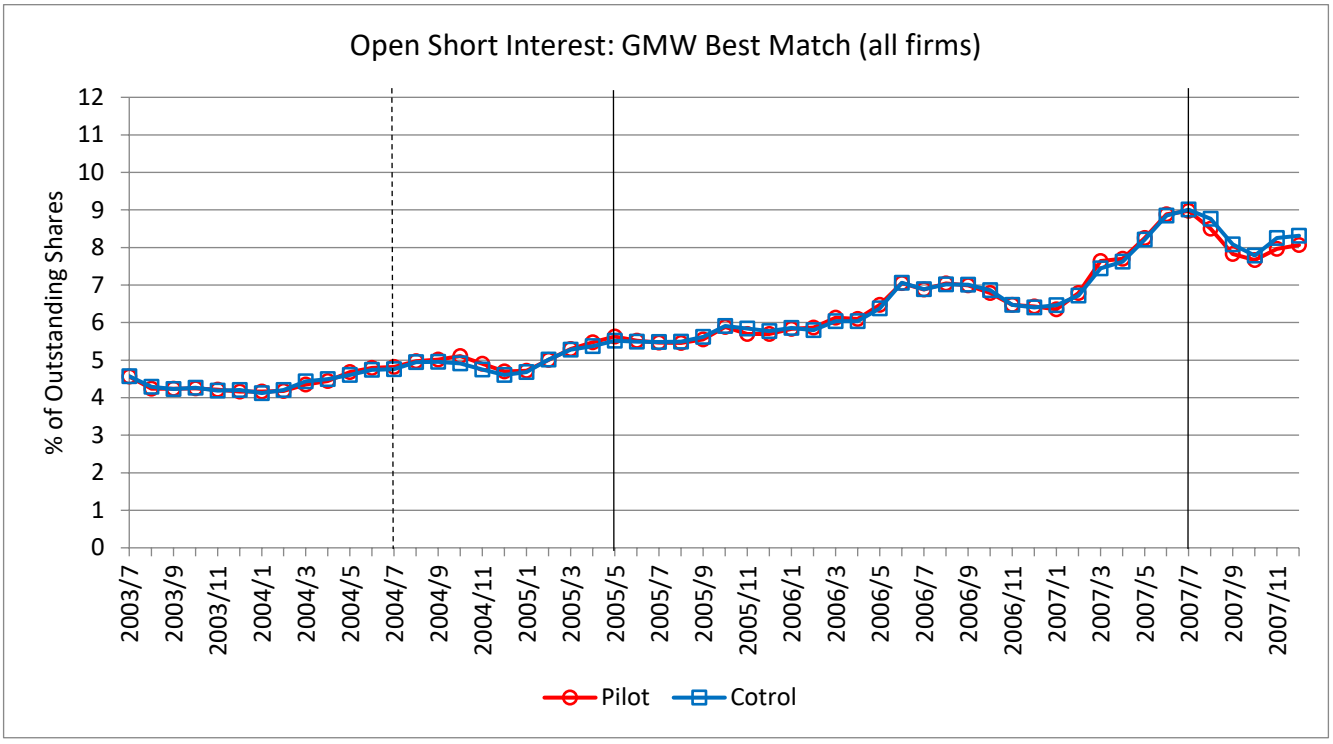
**Panel A. 2004 Announcement Sample**



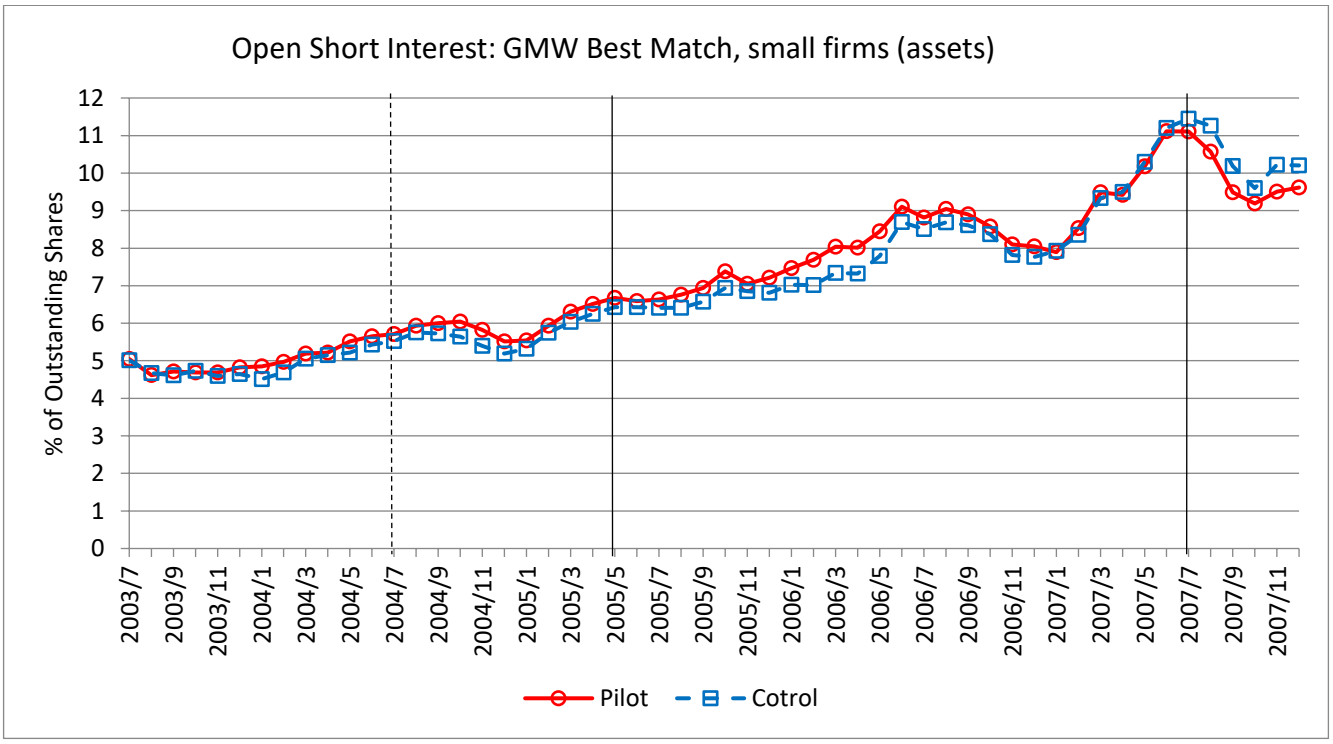

**Panel B. 2005 Analysis Sample**



**Panel C. GMW Best-Match Sample**

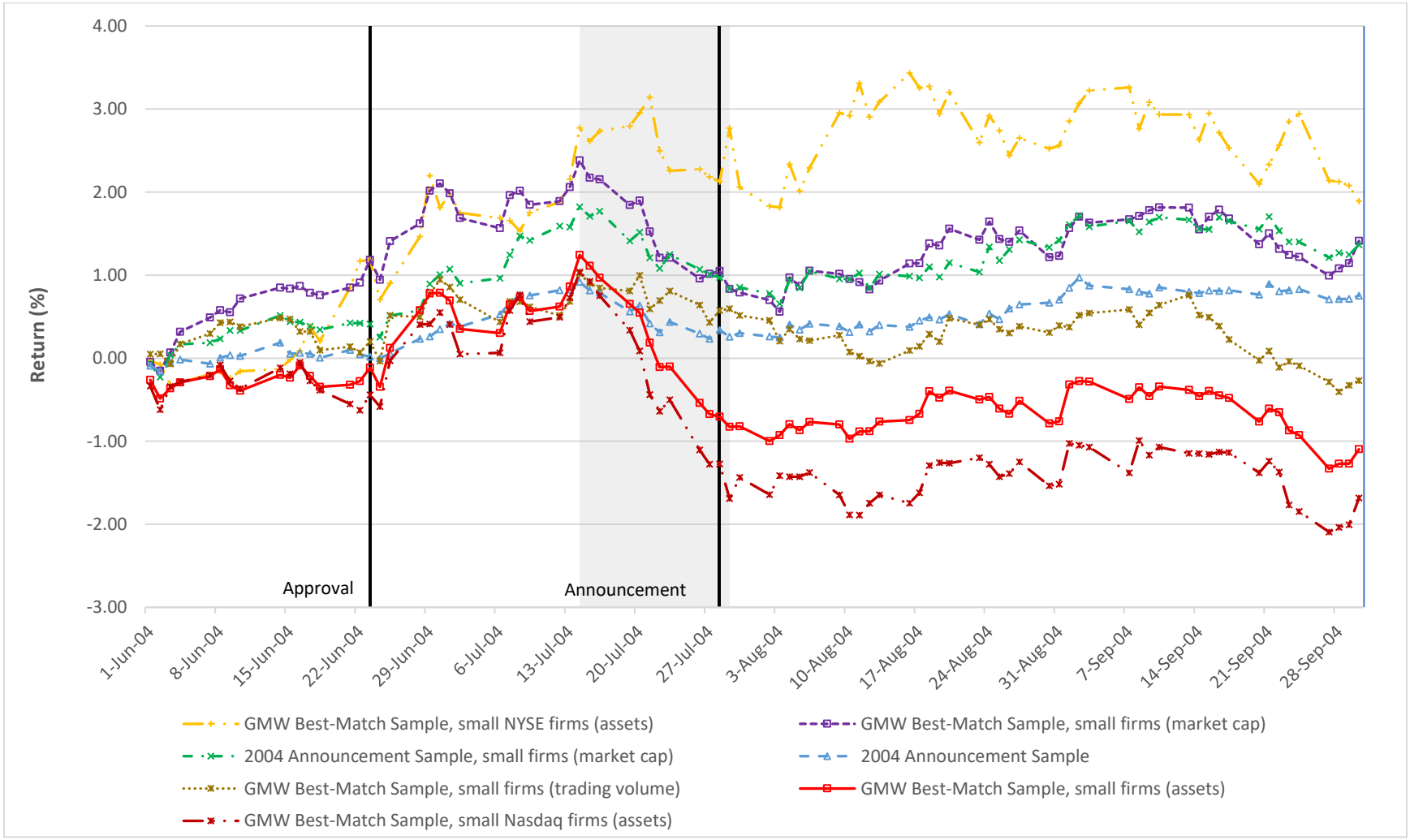



**Panel D. GMW Best-Match Sample: Only Small Firms (Below-Median in Assets)**



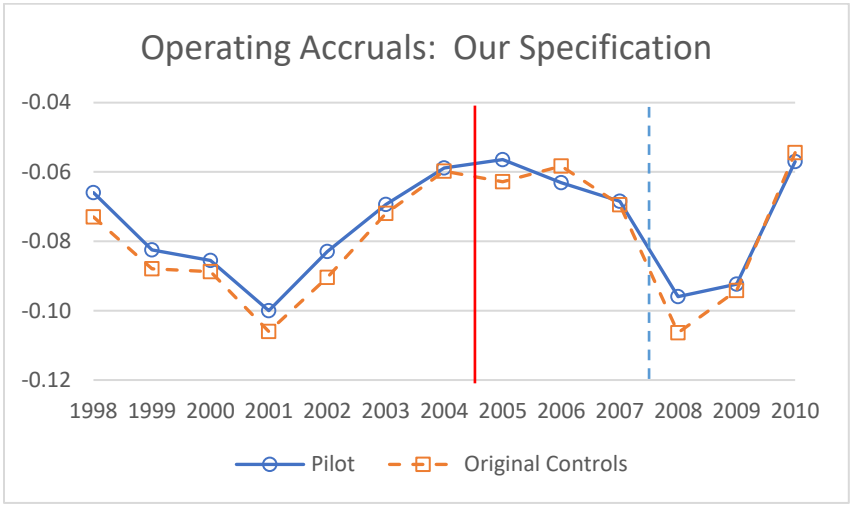
## Figure 2. Buy-and-Hold Relative Returns (BHRRs) for Different Samples

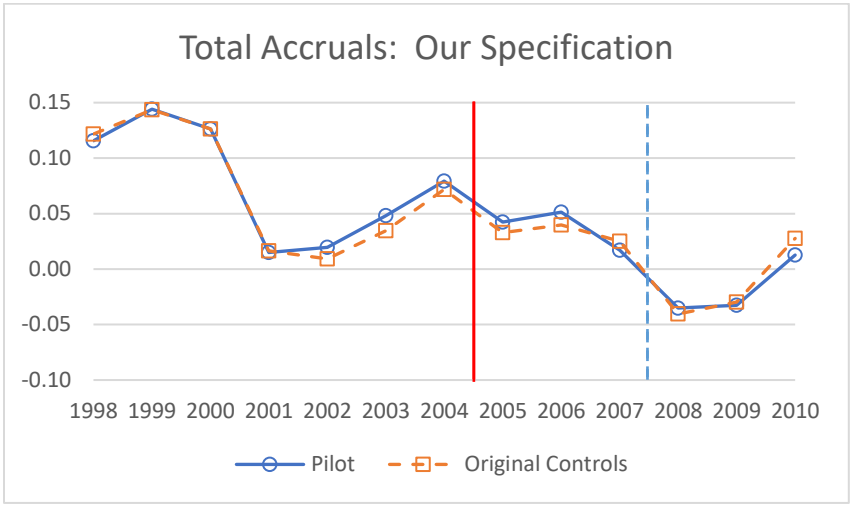
Figure shows cumulative buy-and-hold relative returns (BHRR's) for pilot versus control firms in indicated samples over June 1, 2004-September 30, 2004. Firms are equally weighted. Vertical lines indicate SEC announcement of experiment approval on June 23, 2004, and the SEC formal experiment announcement on July 28, 2004. Shaded area indicates principal event period used by GMW: (-10, +1) relative to the formal announcement date. Samples are: 2004 Announcement Sample; 2004 Announcement Sample, small firms (based on market capitalization); GMW Best-Match Sample small firms (based on assets); GMW Best-Match Sample (based on market capitalization); GMW Best-Match Sample small firms (based on assets), NYSE firms; GMW Best-Match Sample small firms (based on assets), Nasdaq firms.

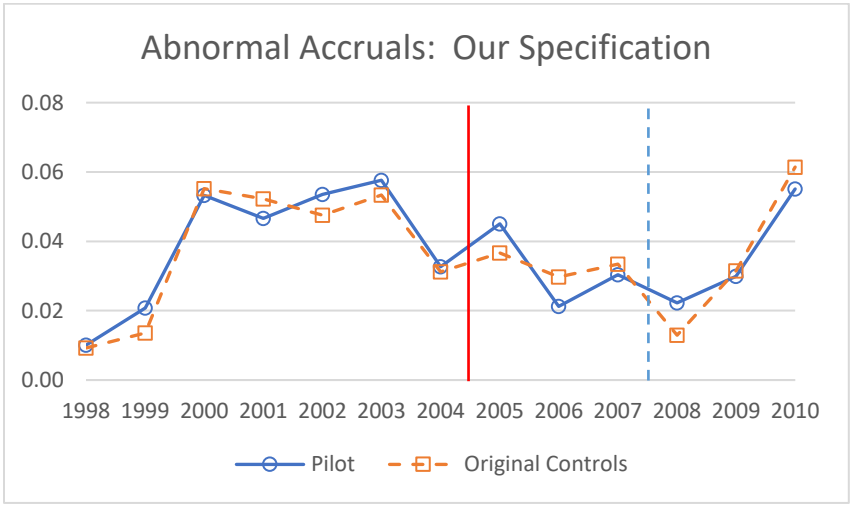


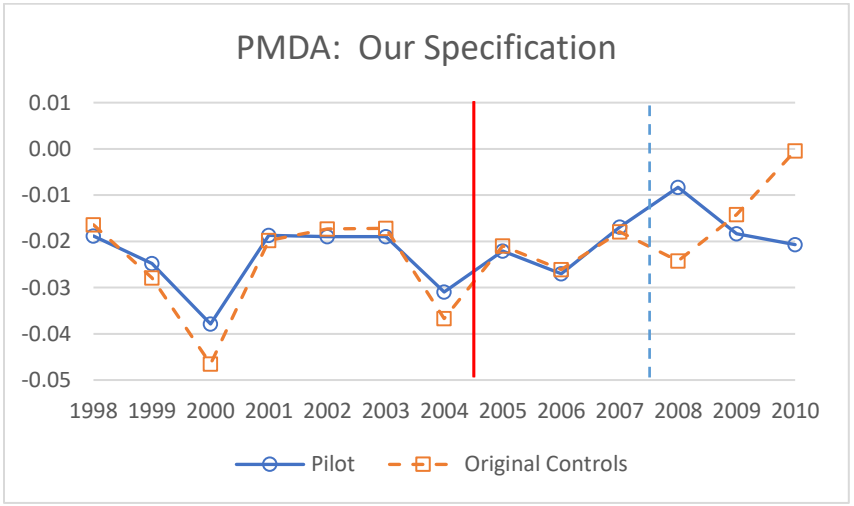
## Figure 3. (FHK) Annual Means for Accruals

Figures show annual means, separately for pilot and control firms, for each accruals measure, over 1998-2010. Figures use our specification and sample (unbalanced panel). Solid and dashed vertical lines separate Pre, During, and Post periods.



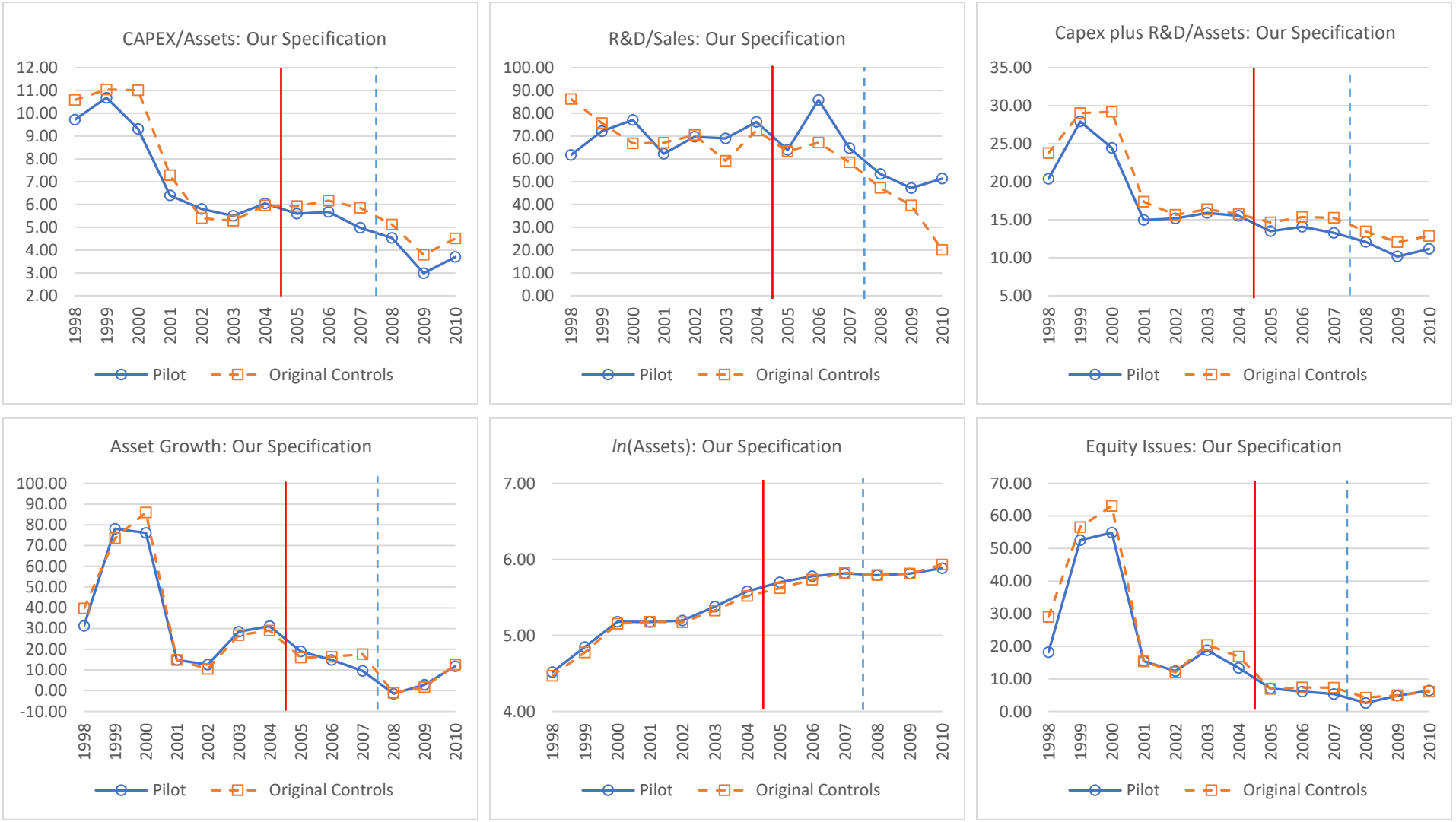



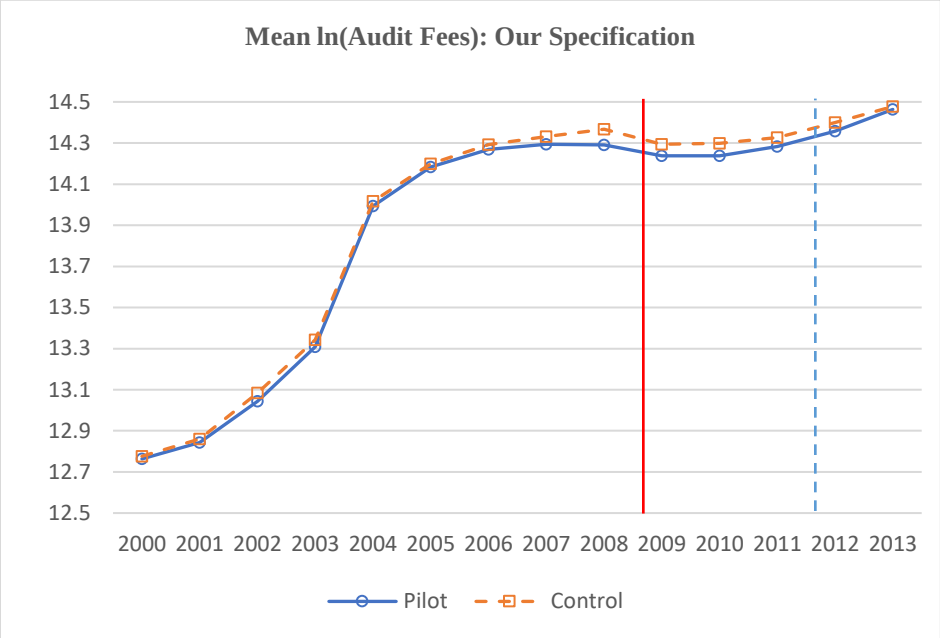
## Figure 4. (GMW) Annual Means for GMW Outcomes for Small Firms (based on assets)

Figures show annual univariate means, separately for small pilot and control firms, using our specification and sample (2005 Analysis Sample), but the GMW definition of small firms (based on assets) Solid and dashed vertical lines separate Pre, During, and Post periods.



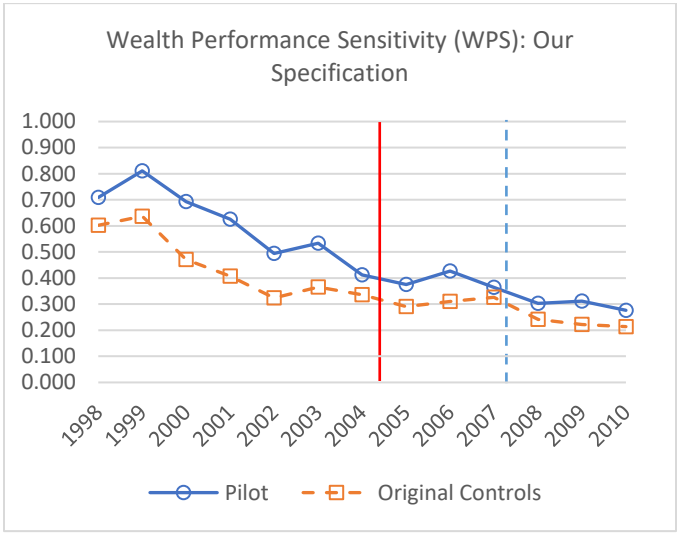
## Figure 5. (HHZ) Annual Means for *ln*(Audit Fees)

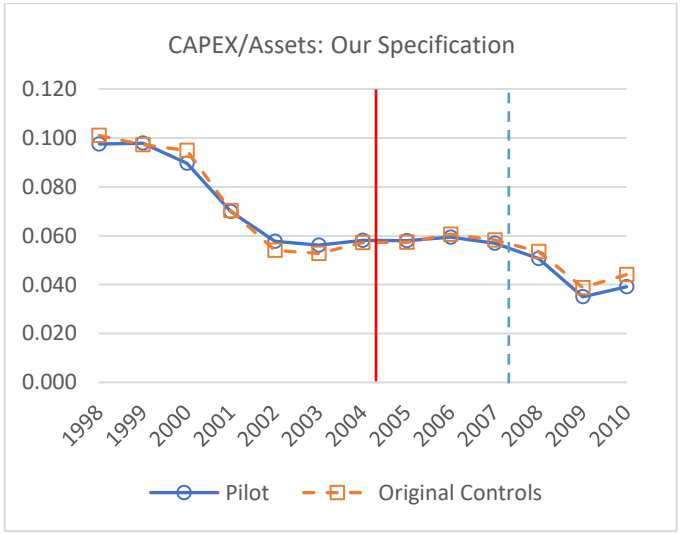
Figure shows univariate means, separately for treated and control firms, of *ln*(Audit Fees), winsorized at 1%/99%, using our specification and sample. Data on auditing fees, needed to extend the time period back to 1998 (as we did for FHK and GMW) is not available. Solid and dashed vertical lines separate Pre, During, and Post periods.

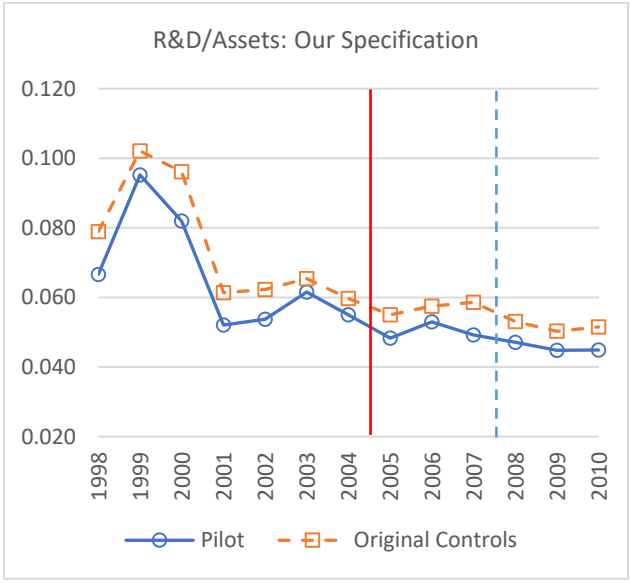


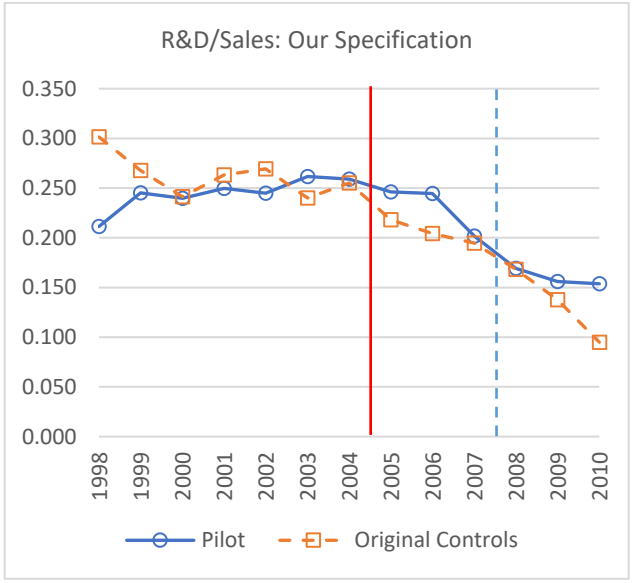
## Figure 6. (LLS) Annual Means for LLS Outcomes

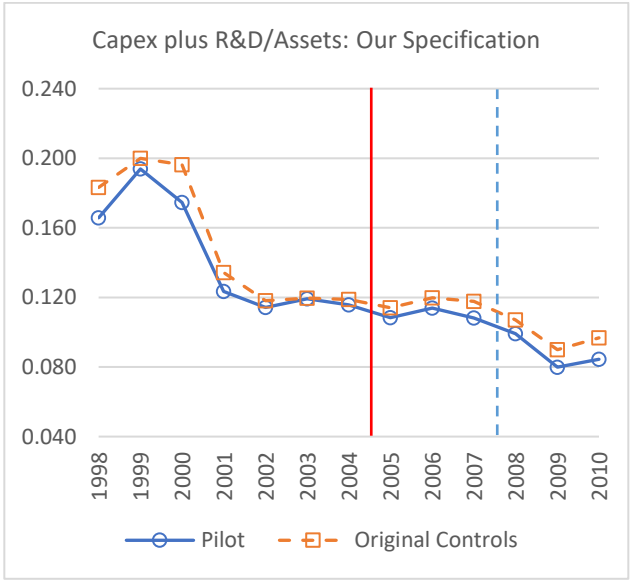
Figures show annual univariate means, separately for pilot and control firms, for each LLS outcome plus Capex/Assets and R&D/Sales, for our specification and sample. Solid and dashed vertical lines separate Pre, During, and Post periods.



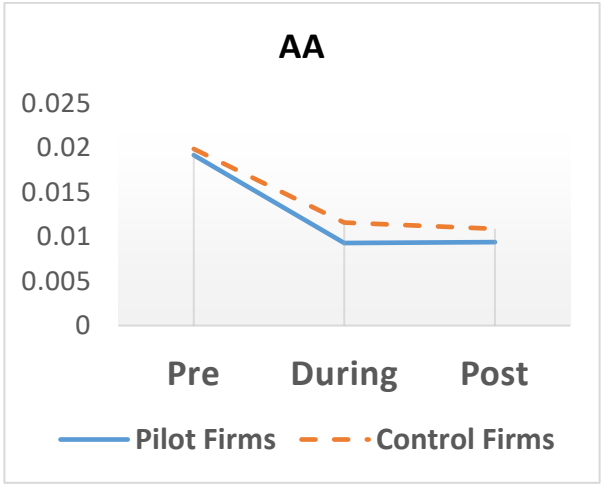



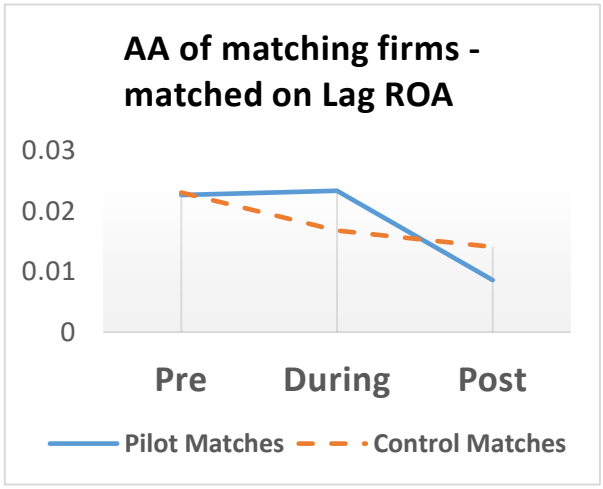



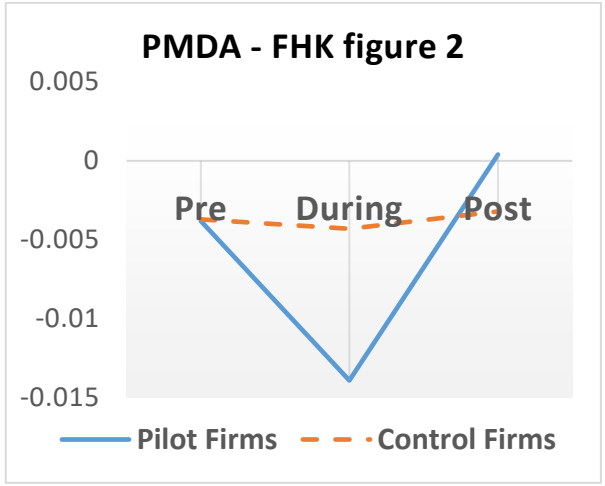



## Figure 7. (FHK) Accruals: Sensitivity to How One Finds Matching Firms

**Panel A. AA for pilot and control firms, AA for their matching firms (matching on lagged ROA), and PMDA using FHK exact sample and specification.** Figure presents sample means, averaged over the Pre, During, and Post periods, for AA and PMDA, separately for pilot and control firms, using the FHK exact, posted sample and sample periods. Left-hand figure shows AA for pilot and control firms. Middle figure shows AA for the matching firms for the pilot firms and the matching firms for the control firms. Right-hand figure shows PMDA (AA for sample firm minus AA for matching firm and the same for control firms) and replicates FHK Figure 2.

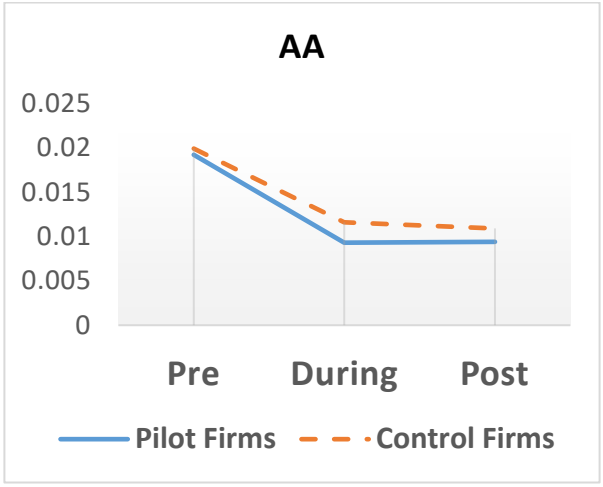


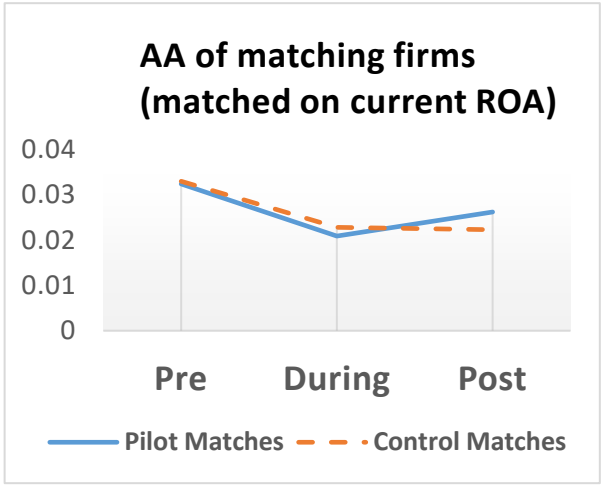



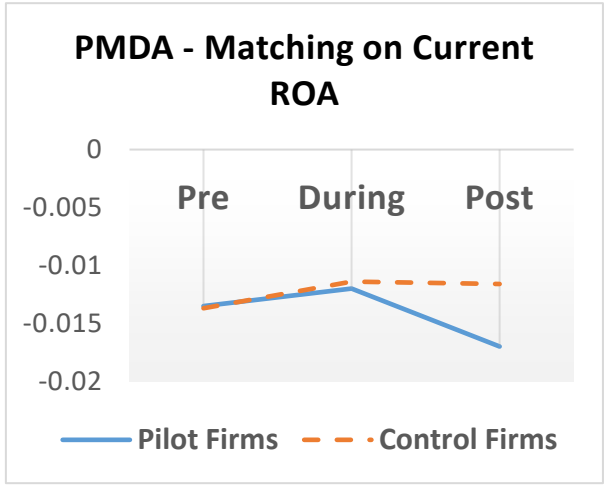




**Panel B. AA for Pilot and Control Firms, for their matching firms (matching on current ROA), and PMDA (matching on current ROA)**

Left hand figure for AA is same as in Panel A. Middle figure shows AA for matching firms, chosen using current year ROA. Right-hand figure shows PMDA (difference between the two)

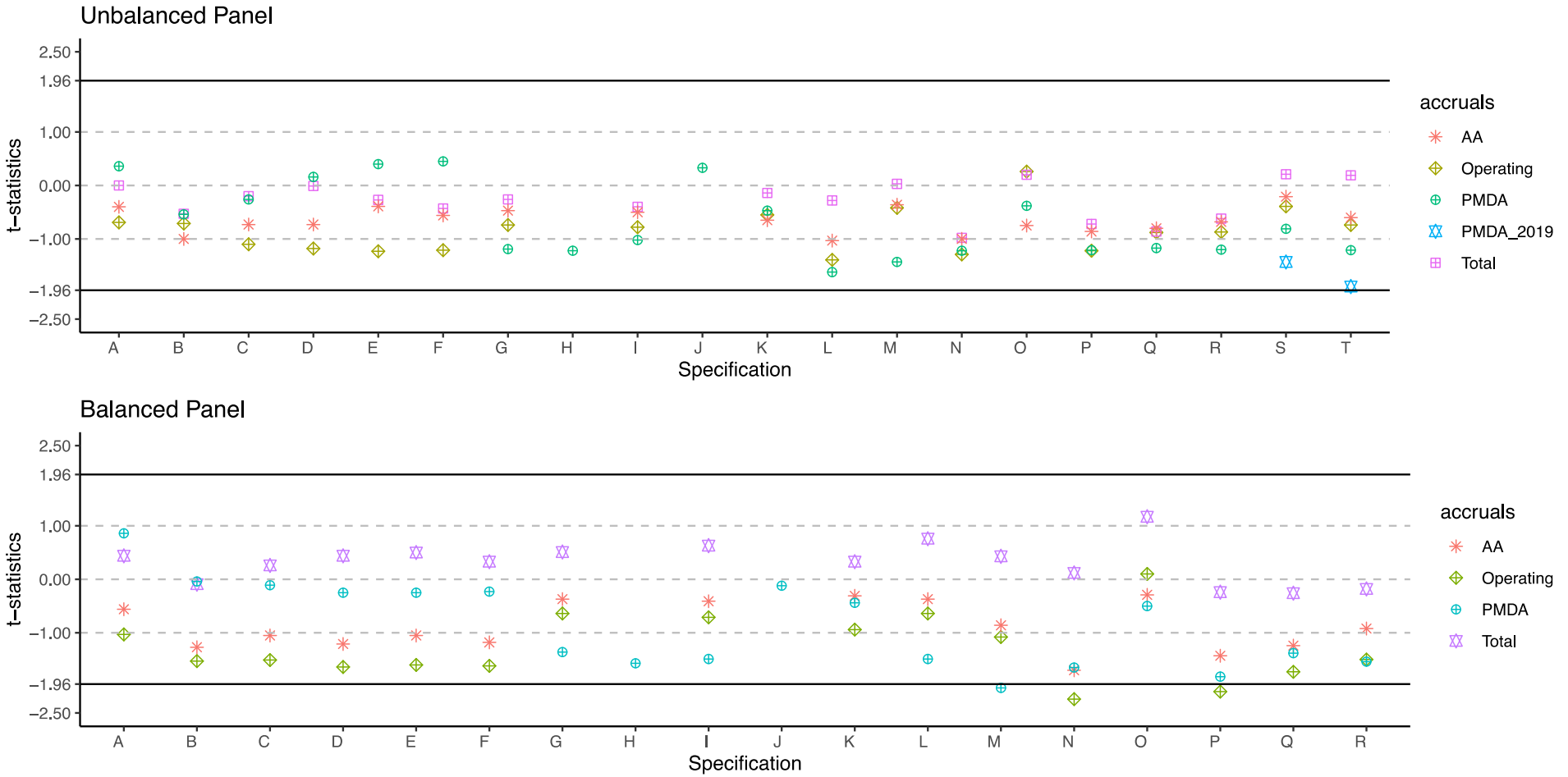


## Figure 8. (FHK) Accruals Measures and Statistical Significance Across Specifications

**Panel A.** Figure shows *t*-statistics, using s.e.'s clustered on firm, for coefficients on Pilot*During for indicated accruals measures, for unbalanced panel, Specifications are as follows (see these tables for additional details): **A**. Our specification (Table 5). ***B-E (from Table IA-10)***. B. A but use calendar year periods. **C.** B but remove 2004 from Pre period. **D.** C but use FHK sample. **E**. D but remove firm FE. **F.** FHK best-match specification (Table 5). **G**. FHK 2012 PMDA Specification (Table IA-11). ***H-M (from Table IA-12)***. **H**. G but correct FHK matching error. **I**. G, but correct FHK duplicates error. **J.** G but use current year ROA to find PMDA match. **K**. H, except winsorize operating accruals at 1%/99% within fiscal year instead of excluding outliers. **L**. I but use firm and fiscal year FE. **M.** G, except include 2004 in Pre period. ***N-R use FHK Operating Accruals Specification (Table IA-13)***. **N.** FHK Operating Accruals Specification. **O**. N but winsorize operating accruals at 1%/99% within fiscal year instead of excluding outliers. **P.** N but winsorize in second stage regressions. **Q.** N but impute covariate values in determining sample. **R.** N but both winsorize in second stage regressions and impute covariate values. **S.** specification from FHK Reply (2019), Table 15, but corrects FHK post-period error (Table IA-14). **T.** M but add firm FE. **Panel B**. Similar to Panel A except using balanced panel. Omits S and T, which apply only to unbalanced panel. **Both panels**. Solid horizontal lines show 5% significance (t = ± 1.96).



## Figure 9. (HHZ) Statistical Significance Across Specifications

Figure shows *t*-statistics, with standard errors clustered on firm, for the coefficients on Pilot*During for regressions of ln(audit fees), for the specifications in Tables 8, IA-17, IA-18, and IA-19. Specifications are as follows (see these tables for additional details). Regressions use unbalanced panel except as indicated: **A**. Our specification from Table 8. **B**. Our specification, but for balanced panel, from Table IA-18. ***C-H are from Table IA-17, Panels B-G.*** **C**. Same as A but remove fiscal 2004 from Pre period. **D.** Same as E but extend Post period to include 2011-2013**. E**. Same as D but switch to HHZ definition of Pre, During, and Post periods. **F**. Same as E, except replace fiscal year FE with During and Post dummies. **G.** Same as F but require firms to be in R3000 in June 2005. **H.** Same as G but require data on HHZ full covariates for all models. **I**. Best-match specification from Table 8. **J**. Best-match specification, but for balanced panel, from Table IA-18. ***K-T are from Table IA-19, Panels A-J.*** **K**. HHZ exact sample and specification, provided to us. **L**. Corrects HHZ data error for audit fees for 4 firms in 2000. **M.** Same as K, but map Audit Analytics fiscal years correctly to Compustat fiscal years. **N.** Same as M but add fiscal year FE. **O.** Same as M but use covariate values from Compustat. **P.** Start from O but obtain additional observations from Audit Analytics and Compustat. **Q**. Start from P but use historical CIK values to match to Audit Analytics **R**. Similar to Q but start with HHZ list of 1,899 firms before their matching to Audit Analytics and Compustat. **S.** Start from R but remove HHZ requirement that firms be included in R3000 in June 2005. **T**. Start from S but exclude firms that ceased trading before experiment start. **All panels**. Solid horizontal lines show 5% significance (t = ± 1.96)**.**