Reusing Natural Experiments*

Davidson Heath Matthew C. Ringgenberg

Mehrdad Samadi Ingrid M. Werner

September 2019

ABSTRACT

Natural experiments are used in empirical research to make causal inferences. After a natural experiment is first used, other researchers often reuse the setting, examining different outcomes based on causal chain arguments. Using simulation evidence combined with two extensively studied natural experiments, business combination laws and the Regulation SHO pilot, we show that the repeated use of a natural experiment significantly increases the likelihood of false discoveries. To correct this, we propose multiple testing methods which account for dependence across tests and we show evidence of their efficacy.

JEL classification: G1, G10

Keywords: False Positive, Identification, Multiple Hypothesis Testing, Natural Experiments

*Heath and Ringgenberg are with the University of Utah, Samadi is with SMU Cox, and Werner is with The Ohio State University. The authors thank Lucian Bebchuk, Stephen Brown, David De Angelis, Joey Engelberg, Campbell Harvey, Joost Impink, Florian Peters, Alessio Saretto, Noah Stoffman, Allan Timmermann, and Michael Wittry.

Over the last three decades, the credibility revolution has fundamentally altered empirical research in the field of economics, driven by a new-found emphasis on empirical research design. By exploiting conditions that resemble random assignment, researchers can better estimate the causal effect of one variable on another. In the last ten years approximately 15% of all papers published in *The Journal of Finance, Journal of Financial Economics*, and *Review of Financial Studies* use at least one of the following terms: "natural experiment(s)", "quasi(-) natural experiment(s)", or "regulatory experiment(s)" (see Figure 1).[1]

While the increased reliance on natural experiments has been praised for bolstering the credibility of empirical research in the social sciences (Angrist & Pischke, 2010), it is not a panacea. Often, after a natural experiment is first used, other researchers reuse the setting in order to examine different outcome variables. Examples of natural experiments that have been reused repeatedly in social science include: the German separation and reunification; the Vietnam war draft lottery; years of schooling; state-level changes in minimum wage, tax rates, corporate law, and regulation; and even the birth of twins.[2] In this paper, we show that the repeated use of a natural experiment significantly increases the likelihood of false discoveries.

While multiple hypothesis testing is potentially problematic in many settings, the problem is particularly acute for natural experiments in which the same source of exogenous variation is used to test many different null hypotheses. Within a set of studies

[1]Similarly, Bowen, Frésard, and Taillard (2016) estimate that 39 percent of empirical corporate finance articles between 2010 and 2012 use identification technology (they classify methods based on the following categories: Instrumental variables, difference-in-differences, selection models, regression discontinuity designs, and randomized experiments), compared to just 8 percent in the 1970s.

[2]See Meyer (1995), Rozenzwieg and Wolpin (2000), Angrist and Kreuger (2001), and Fuchs-Schündeln and Hassan (2017) for surveys of natural experiments in economics.

that examine related questions, it is often challenging to say whether different null hypotheses are really part of the same "family" of tests. For natural experiments, the issue is clear: all null hypotheses that are tested using the same natural experiment are examining the same theoretical question (i.e., what was the effect of the experiment?). As researchers examine more and more dependent variables using the same setting, the number of Type I errors (false positives) increases. Put differently, the reuse of natural experiments, without correcting for multiple testing, is undermining the credibility revolution.

In this paper, we argue that researchers should account for multiple hypothesis testing when reusing experiments. We start by providing simulation evidence. We simulate the passage of laws to generate placebo treatment effects where the date and state of passage for each law is randomly assigned, similarly to Bertrand, Duflo, and Mullainathan (2004). We then examine 200 different outcome variables, each simulated using independent and identical normal distributions. We initially simulate data with no true effects so that in the absence of Type I errors, there should be no statistically significant effects. We run difference-in-difference regressions on the placebo data. When we check the first ten simulated dependent variables, we find no statistically significant effect at the 5% level. When we check the first 50 simulated dependent variables, we find five statistically significant effects at the 5% level. When we check all 200 simulated dependent variables, we find nine statistically significant effects at the 5% level. In other words, the simulation shows that as we consider more dependent variables using the same experiment, we find more results that are false positives. This is the multiple testing problem.

We next examine a correction for multiple testing using the step-down procedure developed by Romano and Wolf (2005). The Romano and Wolf (2005) procedure controls the family-wise error rate (FWER), which is the probability of making one or more false rejections given all hypotheses considered. While other methods exist to control the FWER (Dunn, 1961; Holm, 1979), the Romano and Wolf (2005) procedure accounts for dependence across tests; as a consequence, the method has more power to reject false null hypotheses than other FWER methods.[3] To examine the performance of the Romano and Wolf (2005) procedure, we simulate data in which we know the number of true effects. On average, our simulation shows that the Romano and Wolf (2005) procedure is effective at controlling the FWER. In Table 1 we show the t-statistic that should be used to control the family-wise error rate at the 5% level as a function of the number of times an experiment is reused in this setting.

To further illustrate the multiple testing problem in natural experiments, we examine two real-world examples. Specifically, we re-examine the empirical evidence on the causal effects of two extensively studied experiments: the enactment of state business combination laws and the Regulation SHO pilot. To date, more than 100 papers have been written using these two settings.[4] We build a sample of 23 dependent variables that have been previously examined in each setting, but we use a uniform sampling frequency and observation window to enable our bootstrap. Hence, we do not attempt to replicate previous studies. Rather, we re-evaluate the effect of the experiments on a set of outcome variables that have been studied in the literature using a common methodology to examine each variable. We then apply the Romano and Wolf (2005)

[3]In the settings that we examine, outcomes are linked through financial statements and aggregate market forces.

[4]Karpoff and Wittry (2018) document more than 80 academic papers that use business combination laws and other state anti-takeover laws for identification. Similarly, Black, Desai, Litvak, Yoo, and Yu (2019) document more than 40 academic papers that use Regulation SHO for identification.

correction to the two settings using three different approaches.

The first approach uses a sequential ordering of the dependent variables: we apply the multiple testing adjustment to each paper using only the dependent variables for which results have been reported on the date the paper in question was written. This first approach effectively raises the bar for statistical significance *over time*, as more papers are written. For each paper, it answers the question "Can we reject the null in this paper given the existing evidence available at the time this study was written?". The second approach is based on causal chain arguments. The causal chain approach sequences the results such that the null hypotheses that are most likely to be rejected are examined first. This approach has been referred to as a "best foot forward policy" in the multiple testing literature (Foster & Stine, 2008). For business combination laws, the first hypothesis to be tested is whether treatment status affects the probability of a takeover, since this is the main intended effect of these laws. For Regulation SHO, the first hypothesis to be tested is whether treatment status affects short selling, since this is the main intended effect of the regulation. In a sense, this approach answers the question "Can we reject the null in this paper given the existing evidence that must be true for this result to make sense economically?". The third approach assumes that all 23 variables were explored, and addresses the question, "Can we reject the null that *nothing* changed as a result of the experiment?".

In all three approaches our evidence using Romano and Wolf (2005) suggests that many of the existing results on both business combination laws and Regulation SHO may be false positives.[5] When we examine the results using sequential ordering, we find only two significant results using business combination laws and four significant results using Regulation SHO. Similarly, when we examine the results using causal chains, we fail to reject 22 out of 23 of the null hypotheses for business combination laws and 19 out of 23 of the null hypotheses for Regulation SHO. Moreover, when we examine all 23 variables at the same time, we fail to reject the null hypotheses that business combination laws and Regulation SHO had no effect for all but one and three outcomes, respectively.

In order to put an upper bound on the magnitude of the multiple testing problem and to avoid data-snooping critiques regarding our choice of dependent variables, we also use a comprehensive approach that examines all possible variables in two popular databases: the Center for Research in Security Prices (CRSP) and Compustat. We construct 293 variables from CRSP and Compustat data items with pre-specified coverage. For business combination laws, we find that 60 of the 293 outcomes are statistically significant at the 5% level. For Regulation SHO, we find that 26 of the 293 outcomes are statistically significant at the 5% level. After applying the Romano and Wolf (2005) correction, no outcomes survive in either setting. Moreover, in both settings we find that the distribution of p-values increases at a roughly constant rate (from 0 to 100%), consistent with the idea that all of the observed variation in p-values is merely due to random chance (see Panels A and B of Figure 5). The results highlight the challenge that we face as a profession when experiments can be reused and not all tests are revealed

[5]While this may seem surprising given the large number of papers relying on causal chain arguments when reusing experiments, Cain, McKeon, and Solomon (2017) and Karpoff, Schonlau, and Wehrly (2019) find that business combination laws did not substantially change the probability of hostile takeovers. Similarly, the evidence in Diether, Lee, and Werner (2009) and Litvak and Black (2016) suggests that Regulation SHO did not significantly change short interest and did not substantially alter the dynamics of asset prices. We discuss these issues in Section III, below.

publicly.

Our results contribute to a growing literature on multiple testing in economics. Even Leamer (1983), who helped start the credibility revolution, notes that specification searches (in which researchers examine many dependent variables) can invalidate traditional inference methods. Accordingly, a growing literature explores ways to adjust for multiple testing. Early methods like those proposed in Dunn (1961) and Holm (1979) did not account for dependence across tests, and as a result, these methods have weak power to reject a false null hypothesis. White (2000) develops a reality-check bootstrap procedure that addresses this issue in order to improve the test's power. Building on the White (2000) procedure, Romano and Wolf (2005) develop a step-down procedure that controls the probability of one or more false rejections across multiple tests. Specifically, the Romano and Wolf (2005) step-down procedure provides adjusted p-values for each hypothesis while controlling the FWER.

Several papers now use these methods and their variants to address multiple testing issues in practice. List, Shaikh, and Xu (2016) address the problem of multiple hypothesis testing in field experiments, proposing a procedure based on Romano and Wolf (2005) for testing multiple hypotheses simultaneously that (i) asymptotically controls the family-wise error rate and (ii) is asymptotically balanced in that the marginal probability of rejecting any true null hypothesis is approximately equal in large samples. They illustrate their procedure by revisiting a field experiment about charitable giving conducted by Karlan and List (2007) which had multiple outcomes, multiple subgroups, and multiple treatments. Our paper is related to, but distinct from, the analysis in List et al. (2016). They show how to correct inferences when a researcher has control over the parameters of an experiment and tests multiple hypotheses at the same time. Like

List et al. (2016), we use the Romano and Wolf (2005) algorithm, but our focus is on the repeated use of natural experiments *across* studies as opposed to problems that arise when testing multiple hypotheses *within* one field experiment.

The problems we raise are related to the general problem of p-hacking discussed by Harvey (2017) in his American Finance Association Presidential address.[6] The topic of how selective publication - the bias against publishing insignificant results - leads to biased estimates and distorted inference has also been the focus of recent work in economics. Brodeur, Cook, and Heyes (2018) apply multiple methods to 13,440 hypothesis tests reported in 25 top economics journals in 2015, to show that selective publication is a substantial problem in research employing difference-in-differences and (in particular) instrumental variable. They study the distribution of p-values, and find suspicious bunching of p-values close to cutoffs.[7] Andrews and Kasy (2019) propose two approaches for identifying the conditional probability of publication as a function of a study's results, and then propose bias-corrected estimators and confidence sets. We capture the selective publication aspect of our candidate experiments in two ways. First, we examine treatment effects for all variables within a well-defined universe to examine the distribution of traditional p-values. Second, we simulate data to illustrate the frequency of false positives that naturally arises when testing multiple hypotheses for the same experiment.

Several recent papers adjust for multiple testing in settings that do not involve natural experiments. Harvey and Liu (2013) propose using a bootstrap method when conducting multiple testing and Harvey, Liu, and Zhu (2016) argue that researchers exam-

[6]Mulherin, Netter, and Poulsen (2018) also discuss similar issues in their observations from nineteen years as editors of the *Journal of Corporate Finance*.

[7]See also Brodeur, Lé, Sangnier, and Zylberberg (2016).

ining whether a new asset pricing factor explains the cross-section of expected returns should use a t-statistic greater than 3.0 to overcome issues with multiple testing. Harvey and Liu (2014) propose a method to correct for multiple testing when evaluating trading strategies. Chordia, Goyal, and Saretto (2017) conduct a data mining exercise of trading strategies, applying several multiple testing methods. Engelberg, McLean, Pontiff, and Ringgenberg (2019) examine whether cross-sectional variables that have been shown to predict stock returns can be aggregated to predict market returns, and they use the Romano and Wolf (2005) procedure to calculate adjusted p-values. In contrast to these studies, our paper is the first to examine the reuse of natural experiments.

The rest of the paper proceeds as follows. Section I describes our procedure for re-evaluating the existing results on business combination laws and Regulation SHO, including data sources and the construction of variables. It also provides an overview of the Romano and Wolf (2005) step-down procedure. Section II presents our main findings. Section III discusses key issues regarding the reuse of experiments and discusses how to account for multiple testing in practice. Section IV concludes.

I. Data and Methodology

To examine the practical importance of multiple testing in natural experiments, we re-evaluate two natural experiments that have been used over 100 times: business combination laws and Regulation SHO. We select these two experiments because they have gathered an exceptional following and illustrate two very different settings: a staggered introduction of state laws and a randomized control trial, respectively. However, our point is applicable to all settings that have been used repeatedly in academic studies (e.g., Vietnam war draft lottery; years of schooling; state level changes in minimum wage, tax rates, corporate law, and regulation; and regulatory experiments such as the U.S. tick size pilot).

We start by discussing our process for the construction of data in each setting. Given the variation in data availability, sample construction, and regression specifications across papers, our aim is not to replicate the sample and method in each individual paper, but rather, to examine the natural experiment more generally. In order to apply bootstrap-based multiple testing methods, we employ a common data frequency, observation window, and screening procedures to build a sample of 23 dependent variables that have been previously examined in each setting.

I.A. Business Combination Laws

U.S. states have adopted business combination laws at different points in time leading to plausibly exogenous variation in the threat of a corporate takeover. This variation has been used to examine a wide-variety of outcome variables including wages, corporate investment, corporate innovation, board size, and dividends. We follow the sample construction procedure in Karpoff and Wittry (2018).[8] This sample consists of annual Compustat data from 1976 through 1995, excluding financial firms, utilities and observations with missing/negative sales or total assets. Because some of the existing literature uses data that is not publicly available while others variables have limited sample periods, we examine a subset of 23 variables from the existing literature. These 23 dependent variables are listed in Table 2 and their construction is further detailed

[8]We thank Michael Wittry for sharing the data set. Our main inferences are qualitatively similar when we include the Karpoff and Wittry (2018) controls for institutional and legal context.

in Appendix Table A1. As in Karpoff and Wittry (2018), our final sample consists of 10,213 firms and 88,648 firm-year observations. We winsorize all continuous outcome variables at the 0.5% and 99.5% levels.

I.B. Regulation SHO

Regulation SHO was a randomized controlled trial designed by the SEC to examine whether the uptick rule affected short selling behavior and stock prices. We examine the sample of treatment and control firms in Diether et al. (2009). This sample excludes stocks that were added to the Russell 3000 index during June 2004 through June 2005. Stocks are also excluded if they underwent corporate events such as mergers, bankruptcies, etc., were added or eliminated in the June 2005 index reconstitution, underwent ticker changes, were listed on Nasdaq's small cap market, changed their listing venue, or they were acquired, merged, or privatized. Stocks with an average price above $100 or average quoted spread exceeding $1.00 are also excluded. We subsequently merge these data with the other sources of outcome variables detailed in Table 2. We further require the availability of annual Compustat data with fiscal years ending during 2002 through 2009, excluding observations with missing/negative sales or total assets. As with business combination laws, we examine a subset of 23 variables from the existing literature. These 23 dependent variables are listed in Table 2 and their construction is further detailed in Appendix Table A1. The final sample consists of 1,708 (576 pilot, 1,132 control) firms and 12,284 firm-year observations. Following Fang, Huang, and Karpoff (2016) and Grullon, Michenaud, and Weston (2015), all continuous outcome variables are winsorized at the 1% and 99% levels.

<Insert Tab. 2>

I.C. Outcome Data Mining

For both business combination laws and Regulation SHO, we also collect a comprehensive set of Compustat and CRSP variables, including commonly used transformations of each variable. In order to arrive to a set of Compustat outcome variables, we collect raw variables from financial statements which are non-missing for at least 70% of observations in a sample from January 1970 through June 2019.[9] For Compustat outcomes, we use the raw variable, raw variable scaled by total assets, and the percentage change of the raw variable scaled by total assets. This approaches results in 96 raw Compustat variables, generating 288 Compustat outcomes in total. We also use monthly CRSP stock data in order to calculate firm-year average trading volume, average share turnover, cumulative returns, average dollar bid-ask spreads, and average percentage bid-ask spreads using firms' fiscal years. The resulting sample contains 293 different dependent variables (See Appendix Table A2 for details).

I.D. Romano and Wolf Procedure

There is a large literature on correcting for multiple testing. Some methods control the FWER, or the probability of making one or more false rejections given all hypotheses considered. Other methods control the false discovery rate (FDR), defined as the expected value of the ratio of false rejections to rejections. Yet other methods control the

[9] We also exclude outcomes for which a treatment effect could not be estimated due to collinearity, since we use a common specification for all variables.

ratio of false rejections to rejections, or the false discovery proportion (FDP) directly. These different approaches have different merits. As the number of hypotheses being tested becomes larger, controlling the FWER becomes a more stringent criterion. Put differently, the more hypotheses tested, the more likely it is that there will be at least one false rejection of a null hypothesis. In some fields (e.g., genetics) researchers may examine tens of thousands of hypotheses; the FDR and FDP were developed to address these situations. Since the number of possible hypotheses is smaller in most natural experiments in economics, we use the FWER.[10]

The most powerful FWER procedures account for the dependence structure across hypotheses by re-sampling using bootstrapping or permutations and reject as many null hypotheses as possible by using a step-down approach. Specifically, we follow the step-down procedure developed in Romano and Wolf (2005) (see also Romano and Wolf (2016)). For a given natural experiment (e.g., business combination laws) with S possible dependent variables we proceed as follows:

1. For each of the S dependent variables, we run a regression using the experiment. For example, for our re-evaluation of business combination laws, we have 23 difference-in-difference regressions. We retain the coefficient estimate and t-statistic of the treatment effect for each dependent variable.

2. We then construct a bootstrap sample for each dependent variable by resampling

[10]See Harvey et al. (2016) for more on this issue; they write, "Both FWER and FDR are important concepts that are widely applied in many scientific fields. However, based on specific applications, one may be preferred over the other. When the number of tests is very large (e.g., a million), FWER controlling procedures tend to become very tough as they control for the occurrence of even a single false discovery among one million tests. As a result, they often lead to a very limited number of discoveries, if any. Conversely, FWER control is more desirable when the number of tests is relatively small, in which case more discoveries can be achieved and at the same time trusted."

the actual data using the stationary bootstrap procedure of Politis and Romano (1994) with 1000 replications and a mean block size of three.[11]

  (a) Because we want to evaluate the null hypothesis that the treatment effect for each dependent variable is zero, we center the actual data before resampling it by subtracting the fitted value from Step 1 from each observation.[12] We then create the bootstrap sample from these values.

3. For each dependent variable and replicant sample, we again run regressions using the experiment. For example, for the 23 dependent variables in our re-evaluation of business combination laws, we have $23 \times 1000 = 23{,}000$ difference-in-differences regressions. We retain the 1000 treatment effect t-statistics for each dependent variable to build a distribution of significance levels.

4. Finally, we perform the step-down procedure. We first sort the S dependent variables based on the absolute value of their actual t-statistics (t_S) from step 1. Then, for each draw of the bootstrap, we calculate the maximum of the absolute value of t-statistic across all dependent variables for that replicant sample ($t_S^{*,m}$).

  (a) Starting with the dependent variable with the largest actual t-statistic, we calculate the Romano and Wolf (2005) adjusted p-value as

$$p = \frac{\#\{t_S^{*,m} > t_S\} + 1}{M + 1} \tag{1}$$

[11]Sullivan, Timmermann, and White (1999) apply the White (2000) reality check to a set of trading strategies using the Politis and Romano (1994) bootstrap and find that the results are robust to different block sizes.

[12]We do not include the intercept in the calculation of the fitted value. Specifically, for each observation $y_{i,t}$ in the actual data we calculate $\tilde{y_{i,t}} = y_{i,t} - (\beta \cdot Treatment_{i,t})$, where β is the coefficient from Step 1.

where M is the number of bootstrap samples. The procedure counts the fraction of times the bootstrap t-statistics exceed the actual t-statistic.

5. Finally, we remove the most recently examined dependent variable from the sample (and bootstrap sample) and repeat step 4 above using the next most significant dependent variable. We proceed until we have examined each dependent variable.

The resulting procedure yields an adjusted p-value, for each dependent variable, that accounts for multiple testing.[13] We also perform two variations on the Romano and Wolf procedure: (i) sequential ordering and (ii) causal chains.

(i) For sequential ordering, we add an additional loop outside the steps discussed above. In other words, if S papers were written on the first date t, we perform the Romano and Wolf (2005) procedure as discussed above for each additional outcome variable from the papers written on the first date and save the resulting p-values. If, on date $t + \tau$ additional papers have been written, we rerun the Romano and Wolf (2005) procedure for each additional outcome variable available on date $t+\tau$ and we save the p-values for the papers that were added after date t (i.e., we do not overwrite the S adjusted p-values we calculated on date t). We cycle through all dates and outcomes until we have p-values for all outcomes.

(ii) For causal chains, we perform a similar procedure, except we add an additional loop based on groupings of variables instead of the date each paper was written. Specifically, if S dependent variables in a literature are examining first order effects, we first perform the Romano and Wolf (2005) procedure as discussed above using

[13] In order to calculate adjusted critical values, we use the 95% percentile of the maximum bootstrapped t-statistics across all draws when testing the first variable where we fail to reject the null.

those S papers and save the resulting p-values. If K dependent variables in a literature are examining second order effects, we then rerun the Romano and Wolf (2005) procedure using all $S + K$ variables and we save the p-values for the K papers (i.e., we do not overwrite the S adjusted p-values we calculated using first order effects). We cycle through all paper groupings until we have p-values for all papers.

II. Results

In this section, we show that the repeated used of natural experiments increases the likelihood of false positives. We start by providing simulation evidence. We then examine two real-world natural experiments that have been extensively studied: business combination laws and Regulation SHO.

II.A. Simulation

To examine the potential for multiple testing problems in natural experiments, we first simulate data. Similar to the exercise in Bertrand et al. (2004), we construct a natural experiment that simulates state-level variation in the adoption of a policy. We simulate the existence of corporations in 50 geographic states, with 60 firms per state and 20 years of monthly data. For each state, we assign a treatment date using a uniform distribution. The resulting database has 720,000 firm-month observations, and each firm is assigned to a state that receives a treatment shock, and these shocks are staggered over time.

We then construct dependent variables. We simulate 200 dependent variables, where D of the variables are manufactured to be a linear function of the treatment status of a firm in a particular state, and the remaining 200 - D dependent variables are simulated as pure noise using a normal distribution with mean zero and unit standard deviation.

In real-world data, it is not possible to know the number of true effects in any setting, however, our simulation allows us to control this parameter in order examine the effectiveness of multiple testing corrections. Accordingly, we simulate four different samples, where the number of true effects (D) is zero, five, ten, and fifty, respectively.

We first examine the sample with zero true effects; we run S difference-in-difference regressions of the form:

$$y_{i,t}^{s} = \alpha_i + \alpha_t + \beta \cdot Treatment_{i,t} + \epsilon_{i,t}, \tag{2}$$

where s indexes the different dependent variables $y_{i,t}^{s}$ (S = 200) for firm i on date t, $Treatment$ is an indicator variable that takes the value one if firm i is in a state that is treated on date t, and α_i and α_t are firm and date fixed effects, respectively. The results are shown in Table 1.

<Insert Tab. 1>

In the absence of type I errors, there should be no statistically significant effects. When we check the first 10 simulated dependent variables (y^1 to y^{10}), we find no statistically significant effect at the 5% level (Table 1 Panel A, column (4)). However, when we check the first 50 simulated dependent variables), we find five statistically significant effects at the 5% level. When we check all 200 simulated dependent variables, we find nine statistically significant effects at the 5% level. This is the multiple testing problem: as we consider more variables, we find more false positives.

We next examine a correction for multiple testing using the Romano and Wolf (2005) procedure. To examine the performance of the Romano and Wolf (2005) procedure, we simulate data with true effects. Panel B of Table 1 examines a sample where there are five true effects. Similarly, Panels C and D examine samples with 10 and 50 true effects, respectively. As with Panel A (where there are zero true effects), in Panels B, C, and D we find more false positives as we examine more variables.

We then examine the Romano and Wolf (2005) adjusted results. Columns (3) and (5) in each panel of Table 1 display the number of significant results using p-values adjusted for multiple testing. Column (3) shows the number of true effects that are statistically significant after the Romano and Wolf (2005) adjustment, while column (5) shows the number of false effects that are significant after the Romano and Wolf (2005) adjustment. On average, the simulation evidence shows the Romano and Wolf (2005) procedure is effective at controlling the family-wise error rate. In each case, the Romano and Wolf (2005) procedure correctly removes false positives and retains true positives.[14] For example, in row 3 of Panel C, we examine 100 dependent variables with 10 true effects. The raw (unadjusted) results find 16 variables that are statistically significant at the 5% level (10 true effects and 6 false effects). After the Romano and Wolf (2005) adjustment, the results show 10 statistically effects (and zero false discoveries).

The simulation also provides hurdles for reusing experiments as researchers examine

[14]It is not surprising that the true effects remain statistically significant after the Romano and Wolf (2005) adjustment (since we have control over their statistical significance when simulating the data). However, this exercise does show that false effects are likely to be removed using the Romano and Wolf (2005) procedure.

more dependent variables, the table shows the t-statistic necessary to control the family-wise error rate, for given assumptions about the number of true effects and the number of dependent variables considered. The critical value ranges from 3.35 if there are many true effects and few variables considered (row one in Panel D) to 6.38 if there are no true effects and many variables considered (row five in Panel A). Fortunately, while the number of true effects is not known in a real-world setting, the results show that the hurdle for significance is not very sensitive to this assumption. Put differently, the variation in critical vales is driven mostly by the number of candidate dependent variables: if 10 variables are examined, the critical values range from 3.35 to 4.70. If 200 candidate variables are considered, the critical values range from 6.16 to 6.38.

II.B. Business Combination Laws

The simulation shows the Romano and Wolf (2005) procedure is effective at controlling the FWER. Accordingly, we next apply it to real-world natural experiments that have been used in more than 100 academic studies. We start with business combination laws. U.S. states have adopted anti-takeover laws (also called business combination laws) at different points in time leading to plausibly exogenous variation in the threat of a corporate takeover. Following the pioneering work of Garvey and Hanka (1999) and Bertrand and Mullainathan (1999), the setting has been used more than 80 times to examine a wide-variety of outcome variables including wages, corporate investment, corporate innovation, board size, and dividends. To the best of our knowledge, none of the existing papers adjusts for multiple testing. Accordingly, we apply the Romano and Wolf (2005) correction to our sample of 23 dependent variables from existing business combination studies. Table 2 provides an overview of these 23 variables.[15]

Following Karpoff and Wittry (2018) we estimate panel regressions of the form:

$$y_{i,j,l,s,t} = \alpha_i + \alpha_{l,t} + \alpha_{j,t} + \beta \cdot BC_{s,t} + \theta' \mathbf{x_{i,t}} + \epsilon_{i,j,l,s,t}, \tag{3}$$

where $y_{i,j,l,s,t}$ is the outcome variable of interest for firm i in year t in industry j, located in state l, and incorporated in state s. BC is an indicator variable which is equal to one if second-generation business combination laws had been adopted in state s by year t and equal to zero otherwise. Further following Karpoff and Wittry (2018), $\mathbf{x}$ is a vector control variables including the natural log of book value of assets (size), size squared, firm age, and firm age squared. Firm, state of location-year, and industry-year fixed effects are also included. Standard errors are clustered at the state of location level. The results of this estimation are reported in Table 3, Panel A. Of the 23 variables we re-examine, 8 of the variables (roughly thirty-five percent) are statistically significant at the 5% level based on annual data and our observation window. Before adjusting for multiple hypothesis testing, BC laws are associated with a reduction in $AMIHUD$, an increase in $CAPEX$, a reduction in $CASHSEC$, an increase in $LEVERAGE$, a reduction in $PPEGROWTH$, a reduction in $SALESGROWTH$, an increase in SGA, and a reduction in STI (proportion of cash holdings in short-term investment).

<Insert Tab. 3>

We then apply the Romano and Wolf (2005) step-down procedure as discussed in

[15]While there are more than 80 existing papers, some examine dependent variables that are not publicly available and some examine dependent variables that were already examined in the literature, so we focus on a subset of 23 variables.

Section I, above. Specifically, we build a bootstrap sample of 1,000 replicants by randomly sampling with replacement from the 10,213 firms in the business combination law sample. In order to draw years for the bootstrap, we use the stationary bootstrap of Politis and Romano (1994) by drawing random blocks with an average block size of three years. In order to preserve cross-sectional and time-series correlation in our bootstrapped panels, we apply the same firms and dates to all outcomes for a given sample.

We apply the Romano and Wolf (2005) procedure using three different approaches. As previously discussed, the first approach uses a sequential ordering of the dependent variables: we apply the multiple testing adjustment to each paper using only the dependent variables that had already been examined on the date the paper in question was written. It answers the question, "can we reject the null in this paper given the existing evidence available at the time this study was written?". The second approach adds an economic component by examining causal chains. The causal chain approaches sequences the results such that the null hypotheses most likely to be rejected are examined first. This approach has been referred to as a "best foot forward policy" in the multiple testing literature (Foster & Stine, 2008). This approach answers the question, "can we reject the null in this paper given the existing evidence that must be true for this result to make sense economically?". The third approach assumes that all 23 variables were explored, and addresses the question, "can we reject the null that nothing changed as a result of the experiment?".

We start with a detailed description of the first approach, which uses sequential ordering based on the date each study was written. Similar to Harvey et al. (2016), our first approach involves manually searching SSRN, Google Scholar, and academic journals for the earliest reported draft date of each paper. The draft dates are reported in Appendix Table A1. We apply an iteration of the Romano and Wolf (2005) procedure for each additional outcome variable. If multiple outcomes share the same date, we sort alphabetically on variable name within each date. Panels A of Figure 2 presents p-values under single hypothesis testing and multiple hypothesis testing using the Romano and Wolf (2005) procedure. Panel C presents adjusted critical values as function of the number of outcomes examined in each setting.

<Insert Fig. 2>

The results from sequential ordering suggests that many of the existing results on business combination laws may be false positives. In Panel A, we find that only two dependent variables, $LEVERAGE$ and STI, are statistically significant after computing adjusted p-values. Panel C provides additional information on the severity of the multiple testing problem in this setting. We need a t-statistic exceeding 3.50 after the 10th variable is examined. The far-right observation in Panel C shows that we need a t-statistic exceeding 4.25 to control the FWER at the usual level when considering all 23 outcomes.

We next examine the causal chain approach. For business combination laws, we group outcomes as follows: we apply a single hypothesis testing critical value to the direct effect outcome, the probability of a takeover ($TAKEOVER$). This is the main effect; effects on all other variables, if they exist, are a result of changes to the probability of a takeover.[16] We then group outcomes related to corporate investment and disclosure decisions as second order outcomes, since theory suggests these are likely related to managerial entrenchment (and therefore, the threat of a takeover). Finally, we group outcome variables related to external parties as third order outcomes.

<Insert Fig. 3>

The results are shown in Figure 3. Panel A presents p-values under single hypothesis testing and multiple hypothesis testing using the Romano and Wolf (2005) procedure with causal chain ordering. Panel C presents adjusted critical values as function of the causal chain order of the outcome. The results immediately highlight a serious concern with business combination laws: the probability of a takeover is not statistically significant.[17] This finding agrees with recent evidence in Cain et al. (2017) and Karpoff et al. (2019), who provide evidence that business combination laws do not substantially alter the likelihood of takeovers. In the sequential chains procedure, this fact alone casts doubt on all other dependent variables that have been examined in the literature. The Romano and Wolf (2005) results confirm this: only STI is statistically significant after applying the multiple testing correction.

<Insert Fig. 4>

[16] It is theoretically possible that other variables change if corporate managers and/or investors believe there is a change in takeover probability, even if none occurs. Even so, takeover probability is still the first variable in the causal chain argument.

[17] Because we sequence the main effect first, the raw and adjusted p-values are identical for this variable.

Finally, we apply the Romano and Wolf (2005) approach using all 23 outcome variables at the same time. The results are shown in Panel A of Figure 4. Only one variable, STI, is statistically significant after adjusting for multiple testing. Overall, the evidence suggests that many of the existing results on business combination laws are likely false positives owing to the large number of candidate dependent variables examined by the existing literature. To explore the severity of this problem, we examine the critical values that would be required, assuming that researchers explored *all* dependent variables available in two widely used databases: CRSP and Compustat. As discussed in Section I, we examine 293 different dependent variables, including raw and popular transformations of each variable. The results are shown in Figure 5.

<Insert Fig. 5>

Panel A shows that while 60 of the data mined outcomes are statistically significant before adjusting for multiple testing (some of which have already been documented in the literature), none survive the adjusted critical value of 5.98. We also note that the distribution increases from left to right at a roughly constant rate, consistent with the idea that observed variation in p-values is due to random chance. Put differently, if researchers using the business combination law setting did collectively examine all variables in CRSP and Compustat, then we cannot reject the null that business combination laws had no effect on firm-level outcomes.

II.C. Regulation SHO

We also examine the Regulation SHO pilot, which has been examined in more than 40 academic studies. While business combination laws represent a natural experiment, in which a researcher can exploit quasi-random variation to generate an exogenous shock, Regulation SHO represents a real experiment in which researchers had control over the parameters. In a now famous paper called "The credibility revolution in empirical economics: How better design is taking the con out of econometrics," Angrist and Pischke (2010) discuss causal inference in economics and argue that randomized control trials (RCTs) represent the ideal setting. Unfortunately, in economics researchers rarely have the ability to conduct an RCT. Regulation SHO was, however, the rare case of an RCT in economics. It was conducted by the Securities and Exchange Commission (SEC) and established a procedure to temporarily suspend Rule 10a-1 "the uptick rule" as well as any short-sale price test for short sales for a stratified sample of 1,000 of the stocks in the Russell 3000 index. The SEC staff sorted all Russell 3000 securities by volume, and designated every third security as a treatment firm, leaving the remaining 2,000 securities as control firms. Treatment began on May 2, 2005 and the experiment continued until July 6, 2007 at which point the removal of the uptick rule was applied to all firms. While the Regulation SHO study was setup as an RCT, the study is now effectively being used as a natural experiment: more than 40 papers have now reused the setting to examine hypotheses that were not part of the original experiment design.

The Regulation SHO experiment was designed by the SEC to examine whether short-sale price tests affected short selling behavior, and as a result, the dynamics of stock prices. The first paper to examine the experiment, Diether et al. (2009), examined these variables. However, in subsequent years the setting has been reused to examine a wide-variety of outcome variables including corporate investment, innovation, M&A, managerial myopia, payout policies, incentive contracts, corporate governance, SEO under pricing, CEO turnover, CEO compensation, employee relations, workplace safety, voluntary disclosure, reporting conservatism, disclosure of bad news, disclosure readability, analyst forecast precision, analysts rounding of forecasts, analyst forecast quality, banks' loan monitoring, and banks' loss recognition. Again, to the best of our knowledge, none of the existing papers adjusts for multiple testing. Accordingly, we apply the Romano and Wolf (2005) correction to our sample of 23 dependent variables from existing Regulation SHO studies.[18]

We estimate panel regressions of the form:

$$y_{i,t} = \alpha_i + \alpha_t + \beta \cdot Treatment_{i,t} + \theta' \mathbf{x_{i,t}} + \epsilon_{i,t}, \tag{4}$$

where $y_{i,t}$ is the outcome variable of interest for firm i in year t; $Treatment_{i,t}$ is an indicator variable equal to one if the firm is in the pilot group and the fiscal year ends on July 31, 2005 or later, equal to one if the firm is in the control group and the firm's fiscal year ends on July 31, 2008 or later, and equal to zero otherwise. This ensures that pilot firms' entire fiscal year is after the pilot announcement date on July 28, 2004 and that control firms' entire fiscal year is after the repeal of the Regulation SHO price tests for all firms on July 6, 2007. $\mathbf{x}$ is a vector control variables including the natural log of book value of assets (size), size squared, firm age, and firm age squared. Firm and year fixed

[18]Even though there are more than 40 papers on Regulation SHO, some of the dependent variables in the literature are not publicly available and some papers examine dependent variables that were already examined in the literature.

effects are also included. Standard errors are clustered at the firm level. The results of this estimation are reported in Table 3, Panel B. Of the 23 variables we re-examine, 7 or roughly thirty percent are statistically significant at the 5% level based on annual data and our sample window. Before adjusting for multiple hypothesis testing, Reg SHO is associated with an increase in $CITE$, a reduction in DA_MJONES, a reduction in PIN_EOH, a reduction in $PROPOSALS$, a reduction in $REPO$, an increase in $SPREAD$, and an increase in $STOCKVOL$.

We apply the Romano and Wolf (2005) step-down procedure using the same process we used for business combination laws. Again, we apply the Romano and Wolf (2005) procedure in three ways: (i) using sequential ordering, (ii) using causal chains, and (iii) examining all 23 variables at the same time. For the first approach (sequential ordering based on the date each paper was written), the raw and adjusted p-values are shown in Panel B of Figure 2. In Panel B, we find that only four of the 23 dependent variables from existing papers, $SPREAD$, $STOCKVOL$, PIN_EOH, and $CITE$ are statistically significant after computing adjusted p-values. The far-right observation in Panel C shows that we need a t-statistic of approximately 2.82 to control the FWER at the usual level when considering all 23 outcomes.

We then examine the causal chain approach. The Regulation SHO pilot was intended to loosen restrictions on short selling under certain circumstances. This could, potentially change short selling activity (the main effect). In turn, changes in short selling activity could have implications for the price formation process. Changes to the price formation process could then affect corporate decisions, such as investment and disclosure.[19] Finally, corporate investment and disclosure decisions could affect external parties, including auditors, analysts, and other firms' behavior. Accordingly, for Regulation SHO we group outcomes as follows: we apply a single hypothesis testing critical value to the direct effect outcome, short interest (SIR). We group outcomes related to the price formation process as second order outcomes. We group outcomes related to corporate investment and disclosure decisions as third order outcomes. Finally, we group outcome variables related to external parties as fourth order outcomes. We apply an iteration of the Romano and Wolf (2005) procedure for each causal chain grouping.

The results are shown in Panels B and D of Figure 3. Panel B presents p-values under single hypothesis testing and multiple hypothesis testing using the Romano and Wolf (2005) procedure with sequential chain ordering. Panel D presents adjusted critical values as function of the number of outcomes examined. Once again, the results immediately highlight a serious concern: Regulation SHO did not significantly alter the level of short selling (SIR). This finding agrees with the evidence in Diether et al. (2009), yet this has not prevented more than 40 other papers from claiming that Regulation SHO changes other dependent variables because it facilitated short selling. In other words, just as we saw with business combination law, the sequential chain argument fails with the main effect. The Romano and Wolf (2005) results confirm this: only four of the remaining dependent variables in Panel B are statistically significant after applying the multiple testing correction (PIN_EOH, $CITE$, $STOCKVOL$, and $REPO$).

When we consider all outcome variables in Panel B of Figure 4, only three outcomes survive ($REPO$ is no longer significant). Once again, all three approaches suggest that

[19] It has also been argued that the threat of a firm being shorted can influence managerial behavior.

many of the existing results on Regulation SHO are likely false positives owing to the large number of candidate dependent variables examined by the existing literature. To explore the severity of this problem, we next look at the critical values that would be required, assuming that researchers explored all 293 dependent variables we get from CRSP and Compustat.

The results are shown in Panel D of Figure 5. Before multiple hypothesis corrections are applied, we find that 26 of the 293 outcomes are statistically significant at the 5% level. However, after we adjust for multiple testing, no outcomes survive the adjusted critical value of 3.48. We also again find that the distribution of p-values across the possible dependent variables increases from left to right at a roughly constant rate, consistent with the idea that observed variation in p-values is due to random chance.

III. Discussion

Overall, the results in the previous section suggest that many of the findings in two widely-studied experiments may be false positives. For both business combination laws and Regulation SHO, there is little evidence of a first order effect from the shock, yet, many studies have been published claiming second, or even third and fourth order effects.

Our paper highlights several key issues that should be addressed when using natural experiments.

III.A. First Stage

First, researchers should verify the necessary conditions for the first step of the causal chain. In studying the effects of a natural experiment, there is a natural division between direct treatment effects and effects further down the causal chain. Direct effects are effects that follow directly from the experiment itself. For Regulation SHO, the experiment was designed to weaken short sale constraints by removing price-tests such as the uptick rule. Thus, the direct effect is short selling activity, which might change as a result of the experiment. For business combination laws the experiment was expected to increase the expected costs of hostile takeovers. Thus, the direct effect is measured by the likelihood of a hostile takeover.

Investigations of the direct effects amount to checking the first stage of an instrumental variables design for relevance. Put differently, checking that the treatment produces a significant shift in the direct-effect variable, both economically and statistically, helps to guard against spurious findings. For both Regulation SHO and business combination laws, recent studies have raised concerns about the direct effects of the experiment, calling into doubt the subsequent findings in these studies. In other words, both settings appear to suffer from a weak instrument problem.

Moreover, even if the relevance condition holds, settings in which treatment status is as good as randomly assigned may still fail the exogeneity requirement. For Regulation SHO, Boehmer, Jones, and Zhang (2019) show that lifting the uptick rule did have some significant direct effect on treated firms, but that control firms were also affected through spillovers, a finding which violates the stable unit treatment assumption (SUTVA). Similarly, for business combination laws Karpoff and Wittry (2018) show that the size and direction of a law's effect on a firm's takeover protection depends on (i) other state anti-takeover laws, (ii) preexisting firm-level takeover defenses, and (iii) the legal regime as reflected by important court decisions. Yet large literatures make use of both settings, even though there is evidence that they fail both the relevance and exogeneity conditions.

III.B. Compound Exclusion Restrictions

In a related point, we also note that researchers reusing an experimental setting should reconcile their exclusion restrictions and their new findings with existing empirical evidence available when their study is written. As a hypothetical example, suppose that a research team discovers a natural experiment that changes variable Y_1 because it changes variable X. Suppose another research team later examines the same setting, and finds a statistically significant result for variable Y_2. The typical exclusion restriction states that the experiment affects Y_2 only through X, but there is already evidence that Y_1 changes too. Accordingly, the researchers should reconcile their exclusion restriction with this existing evidence. In practice, few of the business combination and Regulation SHO papers reconcile their exclusion restriction with the large existing literature. While this requirement is necessarily situation-specific and subjective, we direct the reader to more formal prescriptions for causal inference from the statistics literature (Pearl, 1995, 2009).

III.C. Multiple Testing

Finally, our study highlights that multiple testing is a crucial issue in natural experiments. Indeed, the probability of a false positive in natural experiments may even be higher than the unconditional probability in other settings because natural experiments are likely to be examined by many researchers examining many dependent variables. In this sense, the reuse of natural experiments, without correcting for multiple testing, may actually undermine the credibility revolution. We advocate the use of multiple testing methods which can account for dependence across tests every time a natural experiment is reused.

In sum, we argue that the use (and reuse) of a natural experiment should require the following steps:

1. Researchers should verify the relevance and exclusion restrictions of the main effect before examining higher order effects.

2. If reusing a setting, the researcher should reconcile their exclusion restrictions with the existing findings in the literature.

3. Finally, the researcher should adjust for multiple testing in order to control the FWER.

IV. Conclusion

Natural experiments have become an important tool for identifying the causal relationships between variables. While the use of natural experiments has increased the credibility of empirical economics in many dimensions (Angrist & Pischke, 2010), we show that the repeated reuse of a natural experiment significantly increases the number of false discoveries. As a result, the reuse of natural experiments, without correcting for multiple testing, is undermining the credibility of empirical research. While we are the first to provide direct evidence on this point, we are not the first to acknowledge the issue. For example, Leamer (2010) writes, *"[some researchers] may come to think that it is enough to wave a clove of garlic and chant "randomization" to solve all our* *problems...”* Our results confirm this point; randomization by itself does not solve all inference problems.

Using simulation evidence, we show that the repeated use of a natural experiment to test different hypotheses leads to Type I errors. We then demonstrate the effectiveness of the Romano and Wolf (2005) procedure in adjusting p-values to account for this problem. We also provide guidance on evaluating the statistical significance of empirical results when the same setting is used multiple times.

To demonstrate the practical importance of the issues we raise, we examine two extensively studied real-world examples: business combination laws and Regulation SHO. Combined, these two natural experiments have been used in well over 100 different academic studies. We re-evaluate 46 outcome variables that were found to be significantly affected by these experiments, using common data frequency and observation window. Our analysis suggests that many of the existing findings in these studies may be false positives.

We also note that business combination laws and Regulation SHO are not alone. There are many other frequently re-used natural experiments for which our arguments apply, including staggered changes of state-level laws and taxes, and regulatory experiments implemented (the U.S. tick size pilot) and planned (the SEC transaction fee pilot and the FINRA corporate bond block trade transparency pilot) by U.S. regulators. Our arguments also apply to regression discontinuity settings such as the extensively studied Russell Index reconstitution, which to date has been used for identification in more than 80 academic studies.[20]

[20]For early work in this literature see Mullins (2014) and Chang, Hong, and Liskovich (2015). See Wei and Young (2017) for a critique of this literature.

The repeated use of natural experiments without accounting for multiple hypothesis testing is likely leading to many false discoveries in real-world studies. Overall, we hope that our study contributes to the credibility revolution, not by dissuading the use of natural experiments, but rather by helping researchers account for multiple testing when natural experiments are reused.

References

Alexander, G. J., & Peterson, M. A. (2008). The effect of price tests on trader behavior and market quality: An analysis of reg sho. *Journal of Financial Markets*, *11*(1), 84–111.

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766–94.

Angrist, J. D., & Kreuger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69-85.

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30.

Armstrong, C. S., Balakrishnan, K., & Cohen, D. (2012). Corporate governance and the information environment: Evidence from state antitakeover laws. *Journal of Accounting and Economics*, *53*(1-2), 185–204.

Atanassov, J. (2013). Do hostile takeovers stifle innovation? evidence from antitakeover legislation and corporate patenting. *The Journal of Finance*, *68*(3), 1097–1131.

Babenko, I., Choi, G., & Sen, R. (2018). Management (of) proposals. *Available at SSRN 3155428*.

Bebchuk, L., Cohen, A., & Ferrell, A. (2008). What matters in corporate governance? *The Review of financial studies*, *22*(2), 783–827.

Bennett, B., & Wang, Z. (2018). The real effects of financial markets: Do short sellers cause ceos to be fired? *Fisher College of Business Working Paper*(2018-03), 006.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275.

Bertrand, M., & Mullainathan, S. (1999). Is there discretion in wage setting? A test using takeover legislation. *The Rand Journal of Economics*, 535–554.

Bertrand, M., & Mullainathan, S. (2003). Enjoying the quiet life? corporate governance and managerial preferences. *Journal of political Economy*, *111*(5), 1043–1075.

Bhargava, R., Faircloth, S., & Zeng, H. (2017). Takeover protection and stock price crash risk: Evidence from state antitakeover laws. *Journal of Business Research*, *70*, 177–184.

Black, B. S., Desai, H., Litvak, K., Yoo, W., & Yu, J. J. (2019). Pre-Analysis Plan for the REG SHO Reanalysis Project. *Available at SSRN 3415529*.

Blau, B. M., & Griffith, T. G. (2016). Price clustering and the stability of stock prices. *Journal of Business Research*, *69*(10), 3933–3942.

Boehmer, E., Jones, C. M., & Zhang, X. (2019). Potential pilot problems: Treatment spillovers in financial regulatory experiments. *Journal of Financial Economics*.

Bowen, D. E., Frésard, L., & Taillard, J. P. (2016). Whats your identification strategy? innovation in corporate finance research. *Management Science*, *63*(8), 2529–2548.

Brodeur, A., Cook, N., & Heyes, A. G. (2018). Methods matter: P-hacking and causal inference in economics. *IZA Discussion Paper*.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1–32.

Brown, S., Hillegeist, S. A., & Lo, K. (2004). Conference calls and information asymmetry. *Journal of Accounting and Economics*, *37*(3), 343–366.

Cain, M. D., McKeon, S. B., & Solomon, S. D. (2017). Do takeover laws matter? evidence from five decades of hostile takeovers. *Journal of Financial Economics*, *124*(3), 464–485.

Campello, M., Matta, R., & Saffi, P. A. (2018). The rise of the equity lending market: Implications for corporate policies. *Available at SSRN 2703318*.

Cardella, L., Fairhurst, D. J., & Klasa, S. (2018). What determines the composition of a firm's cash reserves? *Available at SSRN 2391467*.

Chang, Y.-C., Hong, H. G., & Liskovich, I. (2015). Regression discontinuity and the price effects of stock market indexing. *Review of Financial Studies*, *28*(8), 212–246.

Chang, Y.-C., Huang, M., Su, Y.-S., & Tseng, K. (2018). Short-termist ceo compensation in speculative markets: A controlled experiment. *Working Paper*.

Chen, H., Zhu, Y., & Chang, L. (2017). Short-selling constraints and corporate payout policy. *Accounting & Finance*.

Chordia, T., Goyal, A., & Saretto, A. (2017). p-hacking: Evidence from two million trading strategies. *Working Paper*.

De Angelis, D., Grullon, G., & Michenaud, S. (2017). The effects of short-selling threats on incentive contracts: Evidence from an experiment. *The Review of Financial Studies*, *30*(5), 1627–1659.

Deng, X., Gao, L., & Kim, J.-B. (2017). Short selling and stock price crash risk: Causal evidence from a natural experiment. *Available at SSRN 2782559*.

Diether, K. B., Lee, K.-H., & Werner, I. M. (2009). It's sho time! short-sale price tests and market quality. *The Journal of Finance*, *64*(1), 37–73.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64.

Easley, D., Kiefer, N. M., & O'Hara, M. (1997). One day in the life of a very common stock. *The Review of Financial Studies*, *10*(3), 805–835.

Engelberg, J., McLean, R. D., Pontiff, J., & Ringgenberg, M. C. (2019). Are cross-sectional predictors good market-level predictors? *Working Paper*.

Fang, V. W., Huang, A. H., & Karpoff, J. M. (2016). Short selling and earnings management: A controlled experiment. *The Journal of Finance*, *71*(3), 1251–1294.

Foster, D. P., & Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(2), 429–444.

Francis, B., Hasan, I., John, K., & Song, L. (2011). Corporate governance and dividend payout policy: A test using antitakeover legislation. *Financial Management*, *40*(1), 83–112.

Francis, B., Hasan, I., & Song, L. (2009). Agency problem and investment-cash flow sensitivity: Evidence from antitakeover legislation. *Working pape*.

Fuchs-Schündeln, N., & Hassan, T. A. (2017). Natural experiments in macroeconomics. *Handbook of Macroeconomics*, *2*, 923-2012.

Garvey, G. T., & Hanka, G. (1999). Capital structure and corporate control: The effect of antitakeover statutes on firm leverage. *The Journal of Finance*, *54*(2), 519–546.

Giroud, X., & Mueller, H. M. (2010). Does corporate governance matter in competitive industries? *Journal of financial economics*, *95*(3), 312–331.

Gormley, T. A., & Matsa, D. A. (2016). Playing it safe? managerial preferences, risk, and agency conflicts. *Journal of Financial Economics*, *122*(3), 431–455.

Grullon, G., & Michaely, R. (2014). The impact of product market competition on firms payout policy. *Unpublished working paper, Rice University*.

Grullon, G., Michenaud, S., & Weston, J. P. (2015). The real effects of short-selling constraints. *The Review of Financial Studies*, *28*(6), 1737–1767.

Harvey, C. R. (2017). The scientific outlook in financial economics: Transcript of the presidential address and presentation slides. *Duke I&E Research Paper*(2017-06).

Harvey, C. R., & Liu, Y. (2013). Multiple testing in economics. *Available at SSRN 2358214*.

Harvey, C. R., & Liu, Y. (2014). Evaluating trading strategies. *The Journal of Portfolio Management*, *40*(5), 108–118.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5–68.

He, J., & Tian, X. (2016). Do short sellers exacerbate or mitigate managerial myopia? evidence from patenting activities. *Working Paper*.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

Jenter, D., & Kanaan, F. (2015). Ceo turnover and relative performance evaluation. *the Journal of Finance*, *70*(5), 2155–2184.

John, K., Li, Y., & Pang, J. (2016). Does corporate governance matter more for high financial slack firms? *Management Science*, *63*(6), 1872–1891.

John, K., & Litov, L. P. (2010). Corporate governance and financing policy: New evidence. *Unpublished working paper*.

Jones, J. J. (1991). Earnings management during import relief investigations. *Journal of accounting research*, *29*(2), 193–228.

Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, *97*(5), 1774–1793.

Karpoff, J. M., Schonlau, R. J., & Wehrly, E. W. (2019). Which antitakeover provisions matter? *Available at SSRN 3142195*.

Karpoff, J. M., & Wittry, M. D. (2018). Institutional and legal context in natural experiments: The case of state antitakeover laws. *The Journal of Finance*, *73*(2), 657–714.

Ke, Y., Lo, K., Sheng, J., & Zhang, J. L. (2018). Does short selling improve analyst forecast quality? *Working Paper*.

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics*, *132*(2), 665–712.

Leamer, E. E. (1983). Lets take the con out of econometrics. *Modelling Economic Series*, *73*, 31–43.

Leamer, E. E. (2010). Tantalus on the road to asymptopia. *Journal of Economic Perspectives*, *24*(2), 31–46. doi: 10.1257/jep.24.2.31

List, J. A., Shaikh, A. M., & Xu, Y. (2016). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 1–21.

Litvak, K., & Black, B. (2016). The secs busted randomized experiment: What can and cannot be learned. *Northwestern Law & Econ Research Paper Forthcoming*.

Massa, M., Qian, W., Xu, W., & Zhang, H. (2015). Competition of the informed: Does the presence of short sellers affect insider selling? *Journal of Financial Economics*, *118*(2), 268–288.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, *13*(2), 151–161.

Mulherin, J. H., Netter, J. M., & Poulsen, A. B. (2018). Observations on research and publishing from nineteen years as editors of the journal of corporate finance. *Journal of Corporate Finance*, *49*, 120–124.

Mullins, W. (2014). The governance impact of indexing: Evidence from regression discontinuity. *Working Paper*.

Pasquariello, P. (2017). Agency costs and strategic speculation in the us stock market. *Ross School of Business Paper*(1284).

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys*, *3*, 96–146.

Peters, F. S., & Wagner, A. F. (2014). The executive turnover risk premium. *The Journal of Finance*, *69*(4), 1529–1563.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical association*, *89*(428), 1303–1313.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237–1282.

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

Rozenzwieg, M. R., & Wolpin, K. I. (2000). Natural "natural experiments" in economics. *Journal of Economic Literature*, *38*(4), 827-274.

Sauvagnat, J. (2013). Corporate governance and asset tangibility. *Working Paper*.

Sullivan, R., Timmermann, A., & White, H. (1999). Data-snooping, technical trade rule performance, and the bootstrap. *The Journal of Finance*, *54*, 16471691.

Wald, J. K., & Long, M. S. (2007). The effect of state laws on capital structure. *Journal of Financial Economics*, *83*(2), 297–319.

Wang, Z. (2018). Short sellers, institutional investors, and corporate cash holdings. *Available at SSRN 2410239*.

Wei, W., & Young, A. (2017). Selection bias or treatment effect? a re-examination of russell 1000/2000 index reconstitutions. *Available at SSRN 2780660*.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.

Yun, H. (2008). The choice of corporate liquidity and corporate governance. *The Review of Financial Studies*, *22*(4), 1447–1475.

Zeng, H. (2015). The antitakeover laws and corporate cash holdings. *Academy of Accounting and Financial Studies Journal*, *19*(1), 25.

Zhao, Y., Chen, K. H., Zhang, Y., & Davis, M. (2012). Takeover protection and managerial myopia: Evidence from real earnings management. *Journal of Accounting and Public Policy*, *31*(1), 109–135.

Table 1: Simulation Evidence

We simulate the existence of corporations in 50 geographic states, with 60 firms per state and 20 years of monthly data. For each state, we assign a treatment date using a uniform distribution. The resulting database has 720,000 firm-month observations, and each firm is assigned to a state that receives a treatment shock, and these shocks are staggered over time. We simulate 200 dependent variables, where D of the variables are manufactured to be a linear function of the treatment status of a firm in a particular state, and the remaining 200 - D dependent variables are simulated using a normal distribution with mean zero and unit standard deviation. Panel A presents results when there are zero true effects (D = 0). Panel B examines presents results when there are 5 true effects. Similarly, Panels C and D examine samples with 10 and 50 true effects, respectively. We then run difference-in-difference regressions on the simulated data of the form:

$$y^s_{i,t} = \alpha_i + \alpha_t + \beta \cdot Treatment_{i,t} + \epsilon_{i,t},$$

where s indexes the different dependent variables (S = 200) for firm i on date t, $Treatment$ is an indicator variable that takes the value one if firm i is in a state that is treated on date t, and α_i and α_t are firm and date fixed effects. The first column displays the number of dependent variables considered, the second column counts the number of true effects in the data, the third column counts the true effects after the Romano and Wolf (2005) (RW) adjustment, the fourth column counts the false effects found before RW adjustment, the fifth column counts any false effects found after the RW adjustment, and the final column reports the adjusted critical value implied by the Romano and Wolf (2005) procedure.

(1) # Variables Examined	(2) # Significant Unadjusted True Effects	(3) # Significant RW Adjusted True Effects	(4) # Significant Unadjusted False Effects	(5) # Significant RW Adjusted False Effects	(6) RW Adjusted Critical Value
Panel A: 0 True Effects					
10	0	0	0	0	4.70
50	0	0	5	0	5.66
100	0	0	6	0	6.04
150	0	0	8	0	6.28
200	0	0	9	0	6.38
Panel B: 5 True effects					
10	5	5	0	0	4.46
50	5	5	1	0	5.66
100	5	5	3	0	6.03
150	5	5	3	0	6.18
200	5	5	5	0	6.21
Panel C: 10 True Effects					
10	10	10	0	0	3.49
50	10	10	3	0	5.46
100	10	10	6	0	5.94
150	10	10	7	0	6.29
200	10	10	10	0	6.38
Panel D: 50 True Effects					
10	10	10	0	0	3.35
50	50	50	0	0	3.61
100	50	50	2	0	5.48
150	50	50	5	0	5.96
200	50	50	8	0	6.16

Table 2: Outcome Variable Re-evaluations

This table presents the list of re-evaluated outcome variables examined. Panel A presents outcomes from the business combination law literature. Panel B presents outcomes from the Regulation SHO literature. *Outcome* is the name of the outcome variable. *Description* describes the outcome variable. *Related Paper(s)* lists papers that have used a related outcome variable. *Source(s)* provides the source of the outcome variable. All outcomes are transformed into annual frequency using firms' fiscal years. A more detailed explanation of the construction of these outcomes is provided in Appendix Table A1.

Panel A: Business Combination Laws

Outcome	Description	Related Paper(s)	Source(s)
AF_ERROR	Absolute value of the difference between the mean of the most recent annual EPS forecasts and actual annual EPS, scaled by natural log of total assets	Armstrong et al. (2012)	IBES
AMIHUD	Firm-year average of absolute value of daily returns divided by daily dollar volume	Pasquariello (2017)	CRSP daily stock file
ASSETGROWTH	Percentage change in total assets	Giroud and Mueller (2010); Sauvagnat (2013); Karpoff and Wittry (2018)	Compustat annual fundamentals
CAPEX	Capital expenditures scaled by total assets	Francis et al. (2009); Giroud and Mueller (2010); Karpoff and Wittry (2018)	Compustat annual fundamentals
CASHSEC	Cash and marketable securities scaled by total assets	Yun (2008); Zeng (2015); Gormley and Matsa (2016); Karpoff and Wittry (2018)	Compustat annual fundamentals
DA_MJONES_ABS	Discretionary accruals using the modified Jones (1991) approach	Zhao et al. (2012)	Compustat annual fundamentals
DISP	Standard deviation of the most recent annual EPS forecasts scaled by the mean of the most recent annual EPS forecasts	Armstrong et al. (2012)	IBES
DIV	Dividends-common scaled by total assets	Grullon and Michaely (2014)	Compustat annual fundamentals
DIVIDENDPAYOUT	Dividends-common scaled by income before extraordinary items	Francis et al. (2011)	Compustat annual fundamentals
EQCH	Equity issuance minus purchase of common and preferred stock scaled by lagged total assets	Sauvagnat (2013)	Compustat annual fundamentals
EQISSUE	Sale of common and preferred stock scaled by lagged total assets	Sauvagnat (2013)	Compustat annual fundamentals
LEVERAGE	Sum of debt in current liabilities and long-term debt scaled by total assets	Wald and Long (2007); John and Litov (2010)	Compustat annual fundamentals
LOG_CITPAT	Natural log of one plus the firm-year number citations per patent in three years, divided by the annual total number of citations per patent in three years	Atanassov (2013)	Kogan et al. (2017)
LOG_PATENTS	Natural log of one plus the firm-year number of patents granted in three years divided by the annual mean number of patents across all firms in three years	Atanassov (2013)	Kogan et al. (2017)
NUMEST	Firm-year number of analyst estimates	Armstrong et al. (2012)	IBES
PPEGROWTH	Percentage growth of property, plant, and equipment scaled by total assets	Giroud and Mueller (2010); Karpoff and Wittry (2018)	Compustat annual fundamentals
ROA	Earnings before interest, taxes, depreciation, and amortization scaled by total assets	Bertrand and Mullainathan (2003); Giroud and Mueller (2010); Gormley and Matsa (2016); Karpoff and Wittry (2018)	Compustat annual fundamentals
SALESGROWTH	Percentage change in sales	Sauvagnat (2013)	Compustat annual fundamentals
SGA	Selling, general, and administrative expenses scaled by total assets	Giroud and Mueller (2010); John et al. (2016); Karpoff and Wittry (2018)	Compustat annual fundamentals
SKEW	Firm-year skewness of daily returns	Bhargava et al. (2017)	CRSP daily stock file
STI	Proportion of cash holdings in short term investments	Cardella et al. (2018)	Compustat annual fundamentals
STOCKVOL	Firm-year standard deviation of daily returns	Gormley and Matsa (2016)	CRSP daily stock file
TAKEOVER	Indicator variable equal to one if the firm is the target of a takeover in a fiscal year and equal to zero otherwise	Cain et al. (2017); Karpoff et al. (2019)	SDC

Panel B: Regulation SHO			
Outcome	Description	Related Paper(s)	Source(s)
ASSETGROWTH	Percentage change in total assets	Grullon et al. (2015)	Compustat annual fundamentals
CAPEXRND	Capital expenditures plus R&D expenses scaled by lagged total assets	Grullon et al. (2015); Campello et al. (2018)	Compustat annual fundamentals
CASH	Cash and and short-term investment scaled by total assets	Campello et al. (2018); Wang (2018)	Compustat annual fundamentals
CITE	Natural logarithm of one plus the firm-year total number of citations in one year, scaled by the firm-year number of patents granted in one year	He and Tian (2016)	Kogan et al. (2017)
DA_MJONES	Discretionary accruals using the modified Jones (1991) approach	Fang et al. (2016)	Compustat annual fundamentals
DISSUE	Long term debt issuance scaled by lagged total assets	Grullon et al. (2015); Campello et al. (2018)	Compustat annual fundamentals
DIV	Dividends-common scaled by total assets	Chen et al. (2017)	Compustat annual fundamentals
EINDEX	The entrenchment index of Bebchuk et al. (2008)	De Angelis et al. (2017)	Bebchuk et al. (2008)
EQISSUE	Sale of common and preferred stock scaled by lagged total assets	Grullon et al. (2015)	Compustat annual fundamentals
FBIAS	Firm-Year average of quarterly mean forecast errors where signed forecast error is defined as the the difference of an analyst quarterly EPS estimate and actual EPS scaled by price	Ke et al. (2018)	IBES
FORCED	An indicator variable equal to one if a firm experienced a forced CEO turnover and equal to zero otherwise	Bennett and Wang (2018)	Peters and Wagner (2014); Jenter and Kanaan (2015)
INACCURACY	Firm-Year average of quarterly mean unsigned forecast error where forecast error is defined as the the absolute value of difference of an analyst quarterly EPS estimate and actual EPS scaled by price	Ke et al. (2018)	IBES
INSIDEDUM	An indicator equal to one if any officer or director make an open market sale of stock in a firm-year and equal to zero otherwise	Massa et al. (2015)	Thomson Reuters Insider Filings
OP_EQ_DOL	Ratio of the value of stock options granted to the CEO to the total value of equity grants	De Angelis et al. (2017)	Execucomp and Compustat annual fundamentals
OP_EQ_NUM	Ratio of the number of stock options granted to the CEO to the total number of stock options and shares of restricted stock granted	De Angelis et al. (2017)	Execucomp and Compustat annual fundamentals
PIN_EOH	Firm-year Easley et al. (1997) probability of informed trade	De Angelis et al. (2017)	Brown et al. (2004)
PROPOSALS	The firm-year number of all management-sponsored proposals	Babenko et al. (2018)	ISS
REPO	Purchase of common and preferred stock scaled by lagged total assets	Campello et al. (2018); Chang et al. (2018)	Compustat annual fundamentals
SIR	Firm-Year average of short interest divided by shares outstanding	Diether et al. (2009); Grullon et al. (2015)	Nasdaq and Compustat
SKEW	Firm-year skewness of daily returns	Deng et al. (2017)	CRSP daily stock file
SPREAD	Firm-year average of daily average dollar effective spreads	Alexander and Peterson (2008); Diether et al. (2009)	TAQ
STOCKVOL	Firm-year standard deviation of daily returns	Alexander and Peterson (2008); Diether et al. (2009); Blau and Griffith (2016)	CRSP daily stock file
VALUE	Natural log of one plus the firm-year average of real citation value for patents granted in one year	He and Tian (2016)	Kogan et al. (2017); FRED

Table 3: Outcome Variable Re-evaluation Estimates

This table presents treatment coefficients of the re-evaluated outcome variables examined. Panel A presents outcomes from the business combination law literature. In order to re-evaluate the enactment of business combination laws, we estimate panel regressions of the form:

$$y_{i,j,l,s,t} = \alpha_i + \alpha_{l,t} + \alpha_{j,t} + \beta \cdot BC_{s,t} + \theta' \mathbf{x_{i,t}} + \epsilon_{i,j,l,s,t},$$

where $y_{i,j,l,s,t}$ is the outcome variable of interest for firm i in year t in industry j, located in state l, and incorporated in state s. BC is an indicator variable which is equal to one if second-generation business combination laws had been adopted in state s by year t and equal to zero otherwise, $\mathbf{x}$ is a vector control variables including the natural log of book value of assets (size), size squared, firm age, and firm age squared. Firm, state of location-year, and industry-year fixed effects are also included. Standard errors are clustered at the state of location level. Panel B presents outcomes from the Regulation SHO literature. In order to re-evaluate the Regulation SHO pilot, we estimate panel regressions of the form:

$$y_{i,t} = \alpha_i + \alpha_t + \beta \cdot Treatment_{i,t} + \theta' \mathbf{x_{i,t}} + \epsilon_{i,t},$$

where $y_{i,t}$ is the outcome variable of interest for firm i in year t; $Treatment_{i,t}$ is an indicator variable equal to one if the firm is in the pilot group and the fiscal year ends on July 31, 2005 or later, equal to one if the firm is in the control group and the firm's fiscal year ends on July 31, 2008 or later, and equal to zero otherwise. $\mathbf{x}$ is a vector control variables including the natural log of book value of assets (size), size squared, firm age, and firm age squared. Firm and year fixed effects are also included. Standard errors are clustered at the firm level. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

Panel A: Business Combination Laws			Panel B: Regulation SHO Pilot		
Outcome	BC	t-statistic	Outcome	$Treatment$	t-statistic
AF_ERROR	−0.061	−1.01	ASSETGROWTH	−0.012	−1.00
AMIHUD	−0.000**	−2.10	CAPEXRND	−0.004*	−1.66
ASSETGROWTH	−0.044*	−1.87	CASH	−0.002	−0.41
CAPEX	0.003**	2.01	CITE	0.057***	3.09
CASHSEC	−0.008***	−2.83	DA_MJONES	−0.007**	−2.04
DA_MJONES_ABS	−0.003*	−1.80	DISSUE	−0.014*	−1.77
DISP	0.009	0.53	DIV	−0.000	−0.41
DIV	0.000	−1.13	EINDEX	0.023	1.34
DIVIDENDPAYOUT	−0.016	−1.59	EQISSUE	0.003	0.83
EQCH	−0.009	−0.81	FBIAS	0.000	0.55
EQISSUE	−0.009	−0.77	FORCED	−0.002	−0.34
LEVERAGE	0.023**	2.35	INACCURACY	0.001	1.24
LOG_CITPAT	0.000	0.01	INSIDEDUM	−0.026*	−1.83
LOG_PATENTS	0.002	0.68	OP_EQ_DOL	1.044	0.69
NUMEST	−0.060	−0.98	OP_EQ_NUM	0.091	0.06
PPEGROWTH	−0.016**	−2.13	PIN_EOH	−0.006***	−3.66
ROA	−0.017*	−1.75	PROPOSALS	−0.060**	−1.98
SALESGROWTH	−0.273**	−2.53	REPO	−0.006***	−2.74
SGA	0.018**	2.21	SIR	0.000	0.06
SKEW	0.008	0.38	SKEW	−0.040	−0.88
STI	−0.042***	−5.49	SPREAD	0.002**	2.16
STOCKVOL	0.000	0.22	STOCKVOL	0.002***	4.32
TAKEOVER	0.002	1.61	VALUE	0.004	0.21

Figure 1: Fraction of Top Finance Publications Using Quasi-natural, Natural, and Regulatory Experiments. This figure presents the annual fraction of publications in *The Journal of Finance*, *Journal of Financial Economics* (JFE), and *Review of Financial Studies* (RFS) using the terms "natural experiment(s)", "quasi(-)natural experiment(s)", and "regulatory experiment(s)" from 1968 through 2018. The Electronic Journal Center (EJC) is used to count the number of published articles using the terms of interest while ISI Web of Science is used to obtain the annual total number of published articles by year in the same journals.

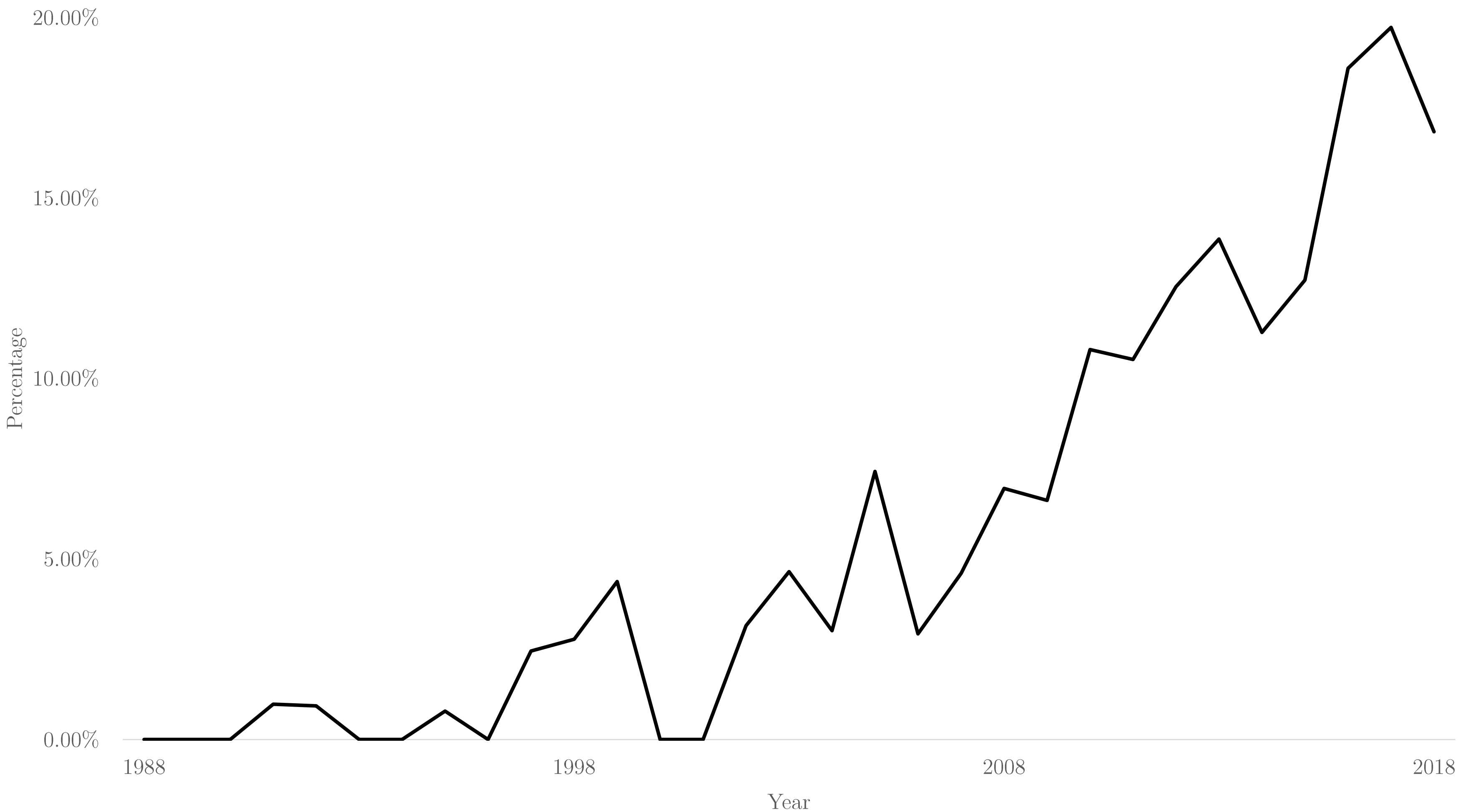

Figure 2: Outcome Re-evaluations, First Draft Date Ordering of Outcomes. This figure presents multiple testing corrected p-values and critical values for a set of outcomes previously examined in studies using business combination laws and Regulation SHO. Statistics are obtained by sequencing outcome variables using their first available draft date and applying an iteration of the Romano and Wolf (2005) procedure for each additional outcome variable. Single hypothesis testing p-values and critical values are used for the first outcome in each setting. Panels A and B present raw and corrected p-values for business combination laws and Regulation SHO, respectively. Panels C and D present adjusted critical values where the horizontal axis is the number of outcomes included in the Romano and Wolf (2005) procedure. Red dashed lines represent the five percent level of statistical significance. These 23 dependent variables are listed in Table 2 and their construction is further detailed in Appendix Table A1.

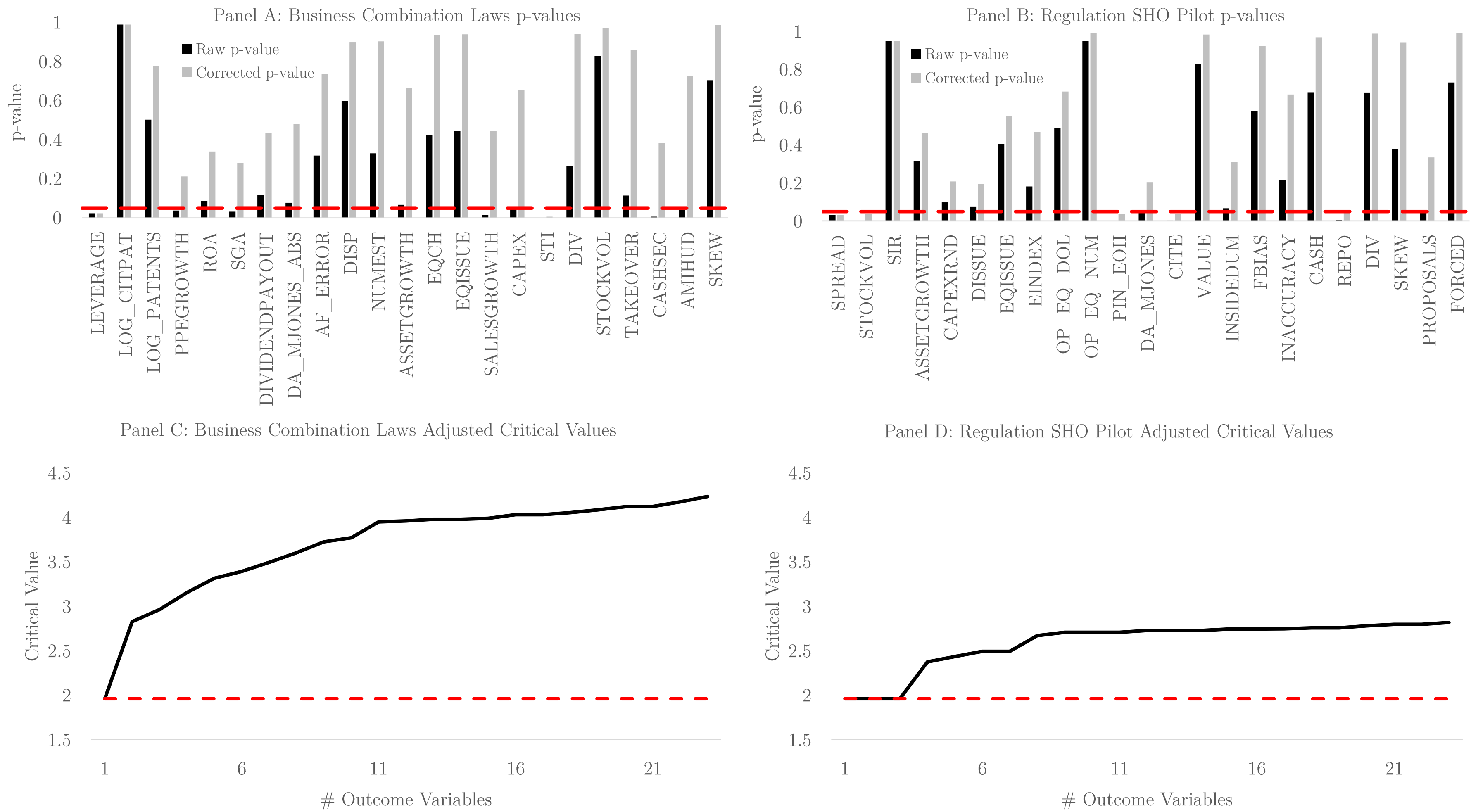

Figure 3: Outcome Re-evaluations, Causal Chain Ordering of Outcomes. This figure presents multiple testing corrected p-values and critical values for a set of outcomes previously examined in studies using business combination laws and Regulation SHO. Statistics are obtained by sequencing outcome variables using causal chain arguments and applying an iteration of the Romano and Wolf (2005) procedure for each additional causal chain order. Single hypothesis testing p-values and critical values are used for the first outcome in each setting. Panels A and B present raw and corrected p-values for business combination laws and Regulation SHO, respectively. Panels C and D present adjusted critical values where the horizontal axis is the number of causal chain orders included in the Romano and Wolf (2005) procedure. Red dashed lines represent the five percent level of statistical significance. These 23 dependent variables are listed in Table 2 and their construction is further detailed in Appendix Table A1.

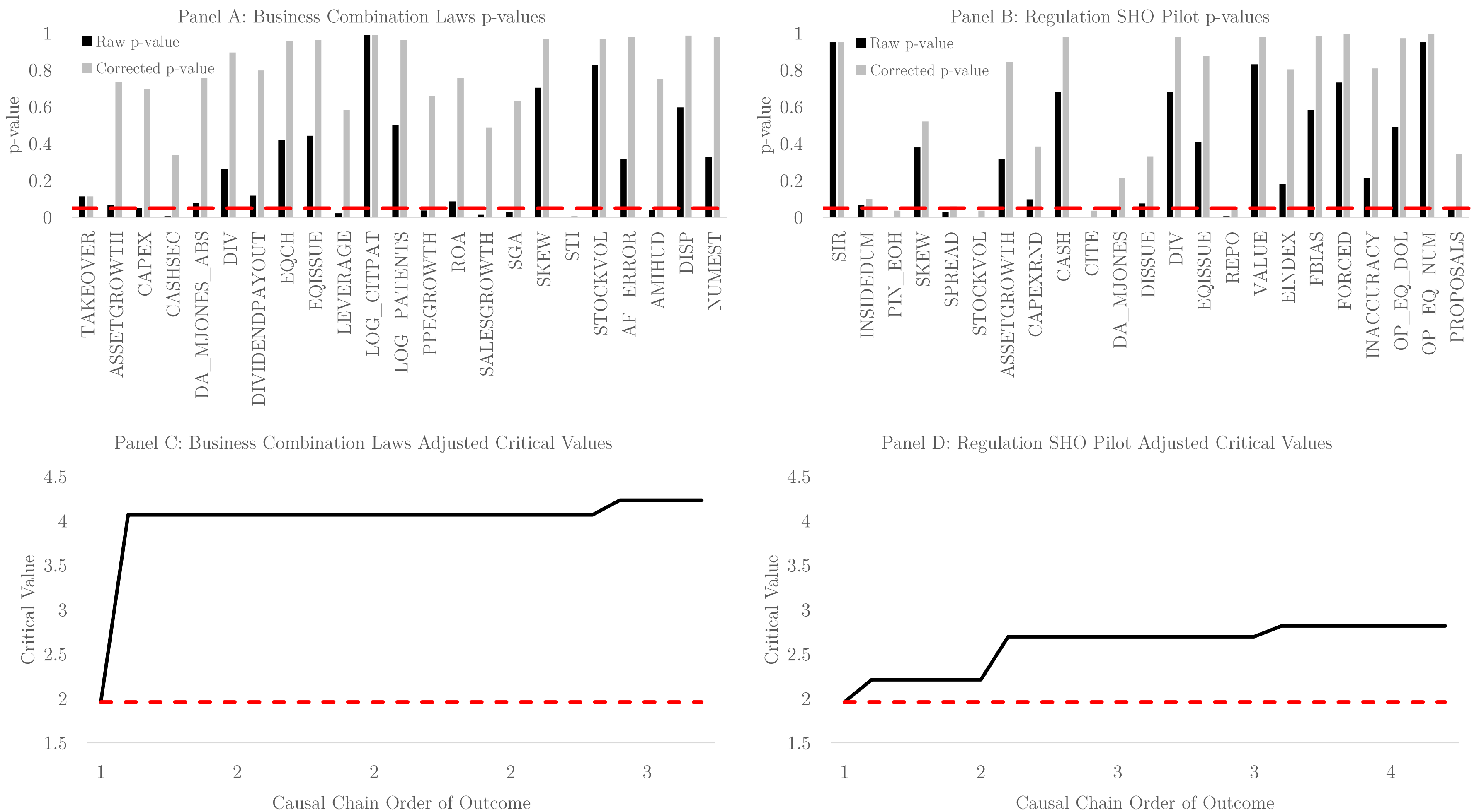

Figure 4: Outcome Re-evaluations. This figure presents multiple testing corrected p-values when all 23 outcomes we re-examine are considered. We apply one iteration of the Romano and Wolf (2005) procedure to the 23 outcomes altogether, Panel A presents results for the outcomes from the business combination laws literature. Panel B presents results for the outcomes from the Regulation SHO pilot literature. Red dashed lines represent the five percent level of statistical significance. These 23 dependent variables are listed in Table 2 and their construction is further detailed in Appendix Table A1.

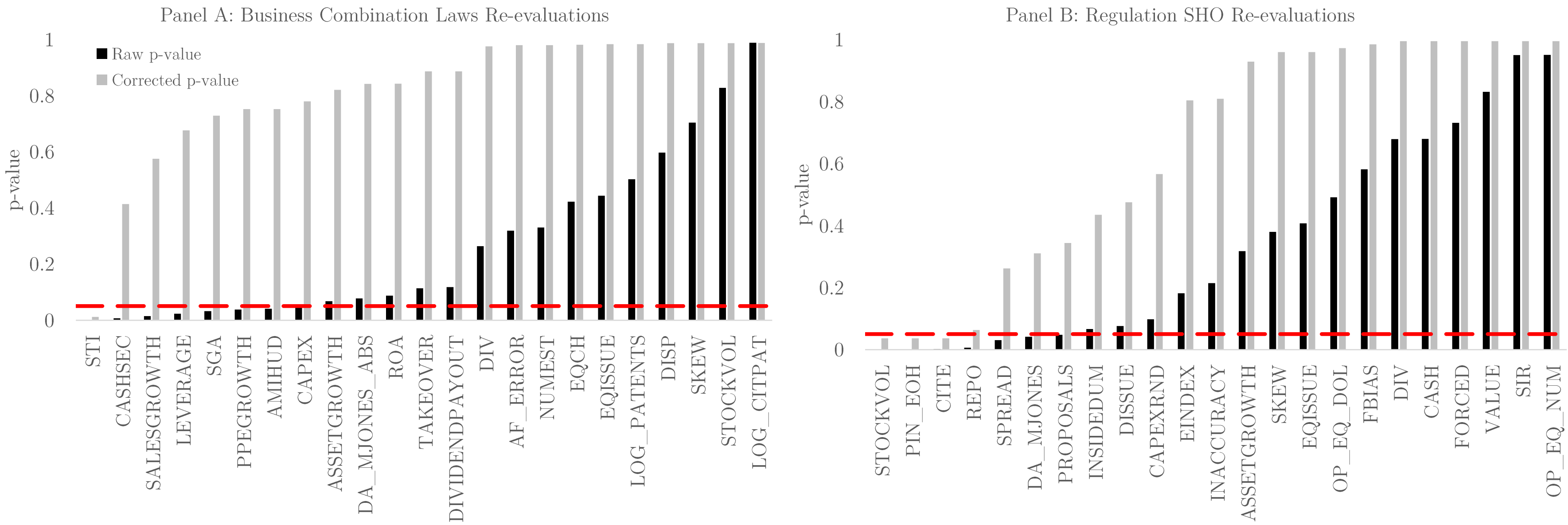

Figure 5: Outcome Data Mining. This figure presents multiple testing corrected *p*-values when a set of 293 outcomes drawn from Compustat and CRSP with pre-specified coverage at the annual frequency are considered. We apply one iteration of the Romano and Wolf (2005) procedure to the 293 outcomes altogether (Panels A and B). In Panels C and D, we present results for the top 20 data mined outcomes in terms of *p*-values. Red dashed lines represent the five percent level of statistical significance. The data mined outcomes are listed in Appendix Table A2.

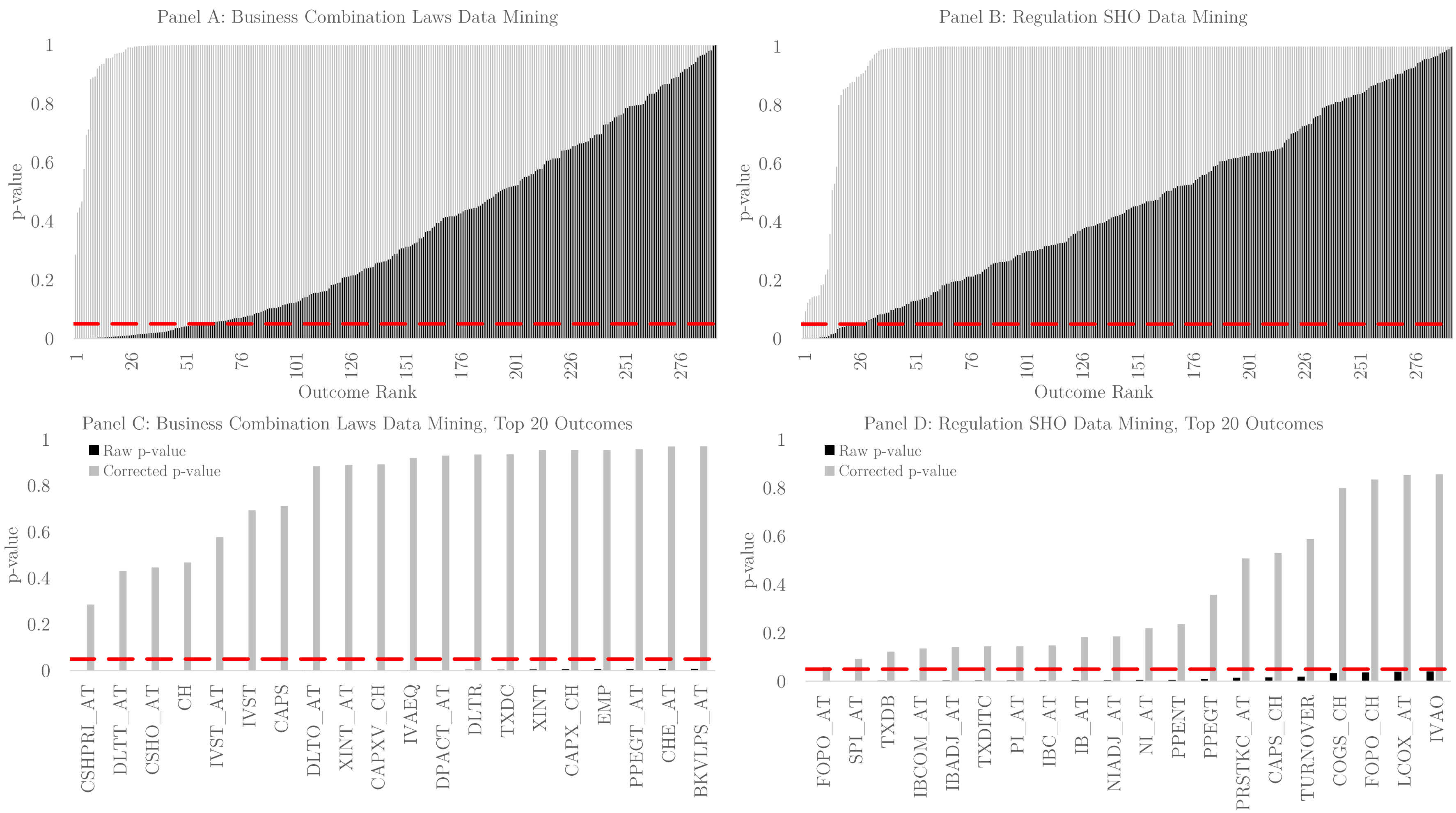

Table A1: Outcome Re-evaluations Construction

This table details the construction of re-evaluated outcome variables. Panel A presents outcomes from the business combination law literature. Panel B presents outcomes from the Regulation SHO literature. *Outcome* is the name of the outcome variable. *Draft Date* is the earliest reported draft date on SSRN, Google Scholar, and journals. *Construction* details the construction of the outcome variable. All outcomes are transformed into annual frequency using firms' fiscal years.

Panel A: Business Combination Laws		
Outcome	Draft Date	Construction
AF_ERROR	3/10/2010	Using the most recent forecast summary of annual EPS from IBES summary data, we calculate the absolute value of the difference between the mean forecast and actual EPS. We scale this difference by the absolute value of total book assets from COMPUSTAT annual fundamentals data. We merge IBES with CRSP using the WRDS IBES/CRSP linkage table, then subsequently merge these data with the sample of Karpoff and Wittry (2018) using the WRDS CRSP/Compustat linkage table.
AMIHUD	6/9/2015	Using CRSP daily stock data, we calculate the firm-year average of the the daily absolute value of returns divided by daily dollar volume. We merge these data with the sample of Karpoff and Wittry (2018) using the WRDS CRSP/Compustat linkage table.
ASSETGROWTH	10/22/2011	Using Compustat annual fundamentals data, we calculate the percentage change in total assets.
CAPEX	7/21/2013	Using Compustat annual fundamentals data, we calculate capital expenditures scaled by total assets.
CASHSEC	1/1/2015	Using Compustat annual fundamentals data, we calculate cash and marketable securities scaled by total assets.
DA_MJONES_ABS	3/1/2009	Using the earnings management models code of Joost Impink (https://github.com/JoostImpink), we calculate the absolute value of the modified Jones (1991) measure. Since cash flow statement information was not required during the sample, we calculate total accruals using the balance sheet approach.
DISP	3/10/2010	Using the most recent forecast summary of annual EPS from IBES summary data, we scale the standard deviation of forecasts by the absolute value of the mean forecast. We merge IBES with CRSP using the WRDS IBES/CRSP linkage table, then subsequently merge these data with the sample of Karpoff and Wittry (2018) using the WRDS CRSP/Compustat linkage table.
DIV	5/1/2014	Using Compustat annual fundamentals data, we calculate dividends-common scaled by total assets.
DIVIDENDPAYOUT	1/1/2009	Using Compustat annual fundamentals data, we calculate dividends-common scaled by income before extraordinary items.
EQCH	10/22/2011	Using Compustat annual fundamentals data, we calculate equity issuance minus purchase of common and preferred stock scaled by lagged total assets.
EQISSUE	10/22/2011	Using Compustat annual fundamentals data, we calculate the sale of common and preferred stock scaled by lagged total assets.
LEVERAGE	8/2/2005	Using Compustat annual fundamentals data, we calculate the sum of debt in current liabilities and long-term debt, scaled by total assets.
LOG_CITPAT	3/2/2007	Using the patent data of Noah Stoffman (https://iu.app.box.com/v/patents), we calculate the natural log of one plus the firm-year number citations per patent in three years, divided by the total number of citations per patent in three years. We merge these data with the sample of Karpoff and Wittry (2018) using the WRDS CRSP/Compustat linkage table.
LOG_PATENTS	3/2/2007	Using the patent data of Noah Stoffman (https://iu.app.box.com/v/patents), we calculate the natural log of one plus the firm-year number of patents granted in three years divided by the year mean number of patents of all firms in three years. We merge these data with the sample of Karpoff and Wittry (2018) using the WRDS CRSP/Compustat linkage table.
NUMEST	3/10/2010	Using the most recent forecast summary of annual EPS from IBES summary data, we use the number of analyst estimates. We merge IBES with CRSP using the WRDS IBES/CRSP linkage table, then subsequently merge these data with the sample of Karpoff and Wittry (2018) using the WRDS CRSP/Compustat linkage table.
PPEGROWTH	3/2/2007	Using Compustat annual fundamentals data, we calculate the percentage growth of property, plant, and equipment scaled by total assets.
ROA	8/7/2007	Using Compustat annual fundamentals data, we calculate earnings before interest, taxes, depreciation, and amortization scaled by total assets.
SALESGROWTH	10/22/2011	Using Compustat annual fundamentals data, we calculate percentage growth in sales.
SGA	8/7/2007	Using Compustat annual fundamentals data, we calculate selling, general, and administrative expenses scaled by total assets.
SKEW	11/15/2015	Using CRSP daily stock data, we calculate the firm-year skewness of daily returns based on firms' fiscal years. We merge these data with the sample of Karpoff and Wittry (2018) using the WRDS CRSP/Compustat linkage table.
STI	2/7/2014	Using Compustat annual fundamentals data, we calculate the difference of cash and short term investments and cash, scaling by cash and short term investments.
STOCKVOL	7/14/2014	Using CRSP daily stock data, we calculate the firm-year standard deviation of daily returns. We merge these data with the sample of Karpoff and Wittry (2018) using the WRDS CRSP/Compustat linkage table.
TAKEOVER	11/2/2014	Using SDC platinum data, we examine mergers and acquisitions of US targets with deal form M, AM, or AA and a completed status. We merge these data with CRSP using historical CUSIPS, then subsequently merge these data with the sample of Karpoff and Wittry (2018) using the WRDS CRSP/Compustat linkage table. We construct an indicator variable that is equal to one if a firm was the target of a takeover in a given year and is equal to zero otherwise.

Panel B: Regulation SHO		
Outcome	Draft Date	Construction
ASSETGROWTH	11/16/2011	Using CRSP/Compustat merged annual fundamentals data, we calculate the percentage change in total assets.
CAPEXRND	11/16/2011	Using CRSP/Compustat merged annual fundamentals data, we calculate capital expenditures plus R&D expenses, scaled by lagged total assets.
CASH	12/15/2015	Using CRSP/Compustat merged annual fundamentals data, we calculate cash and and short-term investment, scaled by total assets.
CITE	1/18/2014	Using the patent data of Noah Stoffman (https://iu.app.box.com/v/patents) we calculate the natural logarithm of one plus firm-year total number of citations, scaled by the firm-year number of patents granted. We subsequently merge these data with the CRSP/Compustat merged annual fundamentals data.
DA_MJONES	6/29/2013	Using the earnings management models code of Joost Impink (https://github.com/JoostImpink), we caculate the modified Jones (1991) measure.
DISSUE	11/16/2011	Using CRSP/Compustat merged annual fundamentals data, we calculate long term debt issuance scaled by lagged total assets.
DIV	10/28/2016	Using CRSP/Compustat merged annual fundamentals data, we calculate dividends-common scaled by assets.
EINDEX	3/24/2013	Using the E-Index data of Lucian Bebchuk (http://www.law.harvard.edu/faculty/bebchuk/data.shtml), we merge the E-Index data with the CRSP stock reference file on historical cusips. We subsequently merge these data with the CRSP/Compustat merged annual fundamental file. We use the most recent E-Index score for a given firm-year.
EQISSUE	11/16/2011	Using CRSP/Compustat merged annual fundamentals data, we calculate the sale of common and preferred stock, scaled by lagged total assets.
FBIAS	12/14/2014	Using IBES detail data and the most recent quarterly forecats of EPS, we calculate the firm-year average of quarterly mean forecast error where signed forecast error is defined as the the difference of an analyst estimate and actual EPS scaled by price. We subsequently merge IBES with the CRSP/Compustat merged annual fundamentals data using the WRDS IBES/CRSP linkage table.
FORCED	5/14/2018	Using the forced CEO turnover data of Florian Peters, we merge forced turnovers with the CRSP/Compustat merged annual fundamentals data using gvkeys. We construct an indicator variable that equal to one if a firm undergoes forced CEO turnover and equal to zero otherwise.
INACCURACY	12/14/2014	Using IBES detail data and the most recent quarterly forecasts of EPS, we calculate the firm-year average of quarterly mean forecast error where signed forecast error is defined as the the difference of an analyst estimate and actual EPS scaled by price. We subsequently merge IBES with the CRSP/Compustat merged annual fundamentals data using the WRDS IBES/CRSP linkage table.
INSIDEDUM	12/10/2014	Using Thomson Reuters insider filings, we collect open market sales with role codes equal to one or more of the following: "CB", "D", "DO", "H", "OD", "VC", "AV", "CEO", "CFO", "CI", "CO", "CT", "EVP", "O", "OB", "OP", "OS","OT", "OX", "P", "S", "SVP", and "VP". We merge these data with the CRSP stock reference data using historical cusips. We subsequently merge these data with the CRSP/Compustat merged annual fundamental data. We construct an indicator variable that is equal to one if an insider sale took place in a given firm-year and is equal to zero otherwise.
OP_EQ_DOL	3/24/2013	Using Execucomp, we calculate the firm-year ratio of the value of stock options granted to the CEO to the total value of equity grants in percentage points. Before 2006, we use the variables option_awards_blk_value and rstkgrnt to determine the value of stock options and stock grants, respectively. For 2006 and later, we use the variables option_awards_fv and stock_awards_fv to determine the value of stock options and stock grants, respectively. We subsequently merge these data with the CRSP/Compustat merged annual fundamentals data using gvkeys.
OP_EQ_NUM	3/24/2013	Using Execucomp, we calculate the firm-year ratio of the number of stock options granted to the CEO to the total number of stock options and shares of restricted stock granted in percentage points. In order to determine the number of shares of restricted stock, we scale the value of restricted stock by the Compustat annual fundamental stock price. We subsequently merge these data with the CRSP/Compustat merged annual fundamentals data using gvkeys.
PIN_EOH	3/24/2013	Using the firm-year probability of informed trade data of Stephen Brown (http://scholar.rhsmith.umd.edu/sbrown/probability-informed-trade-easley-et-al-model), we merge with the CRSP/Compustat merged annual fundamentals data using permnos.
PROPOSALS	4/20/2018	Using ISS corporate vote data, we collect management sponsored proposals with a vote requirement greater than one percent, excluding court and proxy contest meeting types. We also exclude the following agenda ids: "M0201", "M0296", "M0299", "M0101", "M0040", "M0136", "M0020", "M0105", "M0104", "M0617", and 'M0010". we merge these proposal data with the CRSP stock reference data using historical cusips. We subsequently merge these data with the CRSP/Compustat merged annual fundamental data. We construct a count variable for the firm-year number of management sponsored proposals.
REPO	12/15/2015	Using the CRSP/Compustat merged annual fundamentals file, we calculate purchase of common and preferred stock scaled by lagged total assets.
SIR	6/22/2006	Using short interest data from NASDAQ and Compustat and shares outstanding from the CRSP monthly stock data, we calculate the firm-year average of short interest scaled by shares outstanding. We subsequently merge these data with the CRSP/Compustat merged annual fundamentals data.
SKEW	9/20/2017	Using the CRSP daily stock data, we calculate the firm-year skewness of daily returns based on firms' fiscal years. We subsequently merge these data with the CRSP/Compustat merged annual fundamentals data.
SPREAD	3/15/2006	Using the WRDS TAQ intraday indicators data derived from monthly TAQ, we merge stock-day equally weighted average dollar effective spreads with the REG SHO Pilot list of Diether et al. (2009) using stock tickers. We calculate firm-year averages of this spread measure and subsequently merge these data with the CRSP/Compustat merged annual fundamentals data.
STOCKVOL	3/15/2006	Using the CRSP daily stock file, we caculate the firm-year standard deviation of daily returns based on firms' fiscal years. subsequently merge these data with the CRSP/Compustat merged annual fundamentals file.
VALUE	1/18/2014	Using the patent data of Noah Stoffman (https://iu.app.box.com/v/patents) and CPI data from FRED, we calculate the natural log of one plus firm-year average real citation value for patents granted one year from now. We subsequently merge these data with the CRSP/Compustat merged annual fundamentals file.

Table A2: Data Mined Outcomes

This table presents the list of CRSP/Compustat outcomes examined. *Outcome* is the name of the outcome variable. *Description* provides the description, or details the construction of the outcome variable. *Source* provides the source of the outcome variable. All outcomes transformed into annual frequency using firms' fiscal years. In order for a Compustat variable to be included, we require that financial statement variables be non-missing for at least 70% of observations in a sample from January 1970 through June 2019. For Compustat outcomes, we use the raw variable (variable names below), raw variable scaled by total assets (suffix "_AT"), and the percentage change of the raw variable scaled by total assets (suffix "_CH").

Outcome	Description	Source
ACO	Current Assets - Other - Total	Compustat Annual Fundamentals
ACOX	Current Assets - Other - Sundry	Compustat Annual Fundamentals
AO	Assets - Other	Compustat Annual Fundamentals
AOX	Assets - Other- Sundry	Compustat Annual Fundamentals
AP	Accounts Payable - Trade	Compustat Annual Fundamentals
AQC	Acquisitions	Compustat Annual Fundamentals
BKVLPS	Book Value Per Share	Compustat Annual Fundamentals
CAPS	Capital Surplus/Share Premium Reserve	Compustat Annual Fundamentals
CAPX	Capital Expenditures	Compustat Annual Fundamentals
CAPXV	Capital Expend Property, Plant, and Equipment Schd V	Compustat Annual Fundamentals
CEQ	Common/Ordinary Equity - Total	Compustat Annual Fundamentals
CEQL	Common Equity - Liquidation Value	Compustat Annual Fundamentals
CEQT	Common Equity - Tangible	Compustat Annual Fundamentals
CH	Cash	Compustat Annual Fundamentals
CHE	Cash and Short-Term Investments	Compustat Annual Fundamentals
COGS	Cost of Goods Sold	Compustat Annual Fundamentals
CSHO	Common Shares Outstanding	Compustat Annual Fundamentals
CSHPRI	Common Shares Used to Calculate Earnings Per Share - Basic	Compustat Annual Fundamentals
CSTK	Common/Ordinary Stock (Capital)	Compustat Annual Fundamentals
DCLO	Debt - Capitalized Lease Obligations	Compustat Annual Fundamentals
DCPSTK	Convertible Debt and Preferred Stock	Compustat Annual Fundamentals
DCVT	Debt - Convertible	Compustat Annual Fundamentals
DD1	Long-Term Debt Due in One Year	Compustat Annual Fundamentals
DLC	Debt in Current Liabilities - Total	Compustat Annual Fundamentals
DLTIS	Long-Term Debt - Issuance	Compustat Annual Fundamentals
DLTO	Other Long-term Debt	Compustat Annual Fundamentals
DLTR	Long-Term Debt - Reduction	Compustat Annual Fundamentals
DLTT	Long-Term Debt - Total	Compustat Annual Fundamentals
DP	Depreciation and Amortization	Compustat Annual Fundamentals
DPACT	Depreciation, Depletion and Amortization (Accumulated)	Compustat Annual Fundamentals
DPC	Depreciation and Amortization (Cash Flow)	Compustat Annual Fundamentals
DV	Cash Dividends (Cash Flow)	Compustat Annual Fundamentals
DVC	Dividends Common/Ordinary	Compustat Annual Fundamentals
DVP	Dividends - Preferred/Preference	Compustat Annual Fundamentals
DVPSP_C	Dividends per Share - Pay Date - Calendar	Compustat Annual Fundamentals
DVPSP_F	Dividends per Share - Pay Date - Fiscal	Compustat Annual Fundamentals
DVPSX_C	Dividends per Share - Ex-Date - Calendar	Compustat Annual Fundamentals
DVPSX_F	Dividends per Share - Ex-Date - Fiscal	Compustat Annual Fundamentals
DVT	Dividends - Total	Compustat Annual Fundamentals
EBIT	Earnings Before Interest and Taxes	Compustat Annual Fundamentals
EBITDA	Earnings Before Interest	Compustat Annual Fundamentals
EMP	Employees	Compustat Annual Fundamentals
EPSFI	Earnings Per Share (Diluted) - Including Extraordinary Items	Compustat Annual Fundamentals
EPSFX	Earnings Per Share (Diluted) - Excluding Extraordinary Items	Compustat Annual Fundamentals
EPSPI	Earnings Per Share (Basic) - Including Extraordinary Items	Compustat Annual Fundamentals
EPSPX	Earnings Per Share (Basic) - Excluding Extraordinary Items	Compustat Annual Fundamentals
FOPO	Funds from Operations - Other	Compustat Annual Fundamentals
GP	Gross Profit (Loss)	Compustat Annual Fundamentals
IB	Income Before Extraordinary Items	Compustat Annual Fundamentals

IBADJ	Income Before Extraordinary Items - Adjusted for Common Stock Equivalents	Compustat Annual Fundamentals
IBC	Income Before Extraordinary Items (Cash Flow)	Compustat Annual Fundamentals
IBCOM	Income Before Extraordinary Items - Available for Common	Compustat Annual Fundamentals
ICAPT	Invested Capital - Total	Compustat Annual Fundamentals
INTAN	Intangible Assets - Total	Compustat Annual Fundamentals
INVT	Inventories - Total	Compustat Annual Fundamentals
IVAEQ	Investment and Advances - Equity	Compustat Annual Fundamentals
IVAO	Investment and Advances - Other	Compustat Annual Fundamentals
IVST	Short-Term Investments - Total	Compustat Annual Fundamentals
LCO	Current Liabilities - Other - Total	Compustat Annual Fundamentals
LCOX	Current Liabilities - Other - Sundry	Compustat Annual Fundamentals
LO	Liabilities - Other - Total	Compustat Annual Fundamentals
LT	Liabilities - Total	Compustat Annual Fundamentals
NI	Net Income (Loss)	Compustat Annual Fundamentals
NIADJ	Net Income Adjusted for Common/Ordinary Stock (Capital) Equivalents	Compustat Annual Fundamentals
NOPI	Nonoperating Income (Expense)	Compustat Annual Fundamentals
NOPIO	Nonoperating Income (Expense) - Other	Compustat Annual Fundamentals
NP	Notes Payable - Short-Term Borrowings	Compustat Annual Fundamentals
OIADP	Operating Income After Depreciation	Compustat Annual Fundamentals
OIBDP	Operating Income Before Depreciation	Compustat Annual Fundamentals
PI	Pretax Income	Compustat Annual Fundamentals
PPEGT	Property, Plant and Equipment - Buildings (Net)	Compustat Annual Fundamentals
PPENT	Property, Plant and Equipment - Total (Net)	Compustat Annual Fundamentals
PRSTKC	Purchase of Common and Preferred Stock	Compustat Annual Fundamentals
PSTK	Preferred/Preference Stock (Capital) - Total	Compustat Annual Fundamentals
PSTKL	Preferred Stock - Liquidating Value	Compustat Annual Fundamentals
PSTKN	Preferred/Preference Stock - Nonredeemable	Compustat Annual Fundamentals
PSTKRV	Preferred Stock - Redemption Value	Compustat Annual Fundamentals
RE	Retained Earnings	Compustat Annual Fundamentals
RECCO	Receivables - Current - Other	Compustat Annual Fundamentals
RECT	Receivables - Total	Compustat Annual Fundamentals
RETURN	Annual cumulative return	CRSP Monthly Stock File
REVT	Revenue - Total	Compustat Annual Fundamentals
SALE	Sales/Turnover (Net)	Compustat Annual Fundamentals
SEQ	Stockholders Equity - Parent	Compustat Annual Fundamentals
SPI	Special Items	Compustat Annual Fundamentals
SPREAD	Firm-Year Average Spread Between Bid and Ask	CRSP Monthly Stock File
SPREAD_PERC	Firm-Year Average of Percentage Spread Between Bid and Ask	CRSP Monthly Stock File
SSTK	Sale of Common and Preferred Stock	Compustat Annual Fundamentals
TSTK	Treasury Stock - Total (All Capital)	Compustat Annual Fundamentals
TSTKC	Treasury Stock - Common	Compustat Annual Fundamentals
TURNOVER	Firm-Year Average Volume divided by shares outstanding	CRSP Monthly Stock File
TXDB	Deferred Taxes (Balance Sheet)	Compustat Annual Fundamentals
TXDC	Deferred Taxes (Cash Flow)	Compustat Annual Fundamentals
TXDITC	Deferred Taxes and Investment Tax Credit	Compustat Annual Fundamentals
TXP	Income Taxes Payable	Compustat Annual Fundamentals
TXT	Income Taxes - Total	Compustat Annual Fundamentals
VOL	Firm-Year Average Trading Volume	CRSP Monthly Stock File
XIDO	Extraordinary Items and Discontinued Operations	Compustat Annual Fundamentals
XIDOC	Extraordinary Items and Discontinued Operations (Cash Flow)	Compustat Annual Fundamentals
XINT	Interest and Related Expense - Total	Compustat Annual Fundamentals
XOPR	Operating Expenses - Total	Compustat Annual Fundamentals