DISCUSSION PAPER SERIES

DP14710

REUSING NATURAL EXPERIMENTS

Davidson Heath, Matthew Ringgenberg, Mehrdad Samadi and Ingrid M Werner

FINANCIAL ECONOMICS

ISSN 0265-8003

REUSING NATURAL EXPERIMENTS

Davidson Heath, Matthew Ringgenberg, Mehrdad Samadi and Ingrid M Werner

Discussion Paper DP14710
Published 04 May 2020
Submitted 30 April 2020

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Financial Economics

REUSING NATURAL EXPERIMENTS

Abstract

After a natural experiment is first used, other researchers often reuse the setting, examining different outcome variables. We examine the consequences of reusing an experimental setting using two extensively studied natural experiments, business combination laws and the Regulation SHO pilot. We apply multiple hypothesis corrections and our findings suggest many results in the existing literature are false positives. We provide guidelines for inference when an experiment is reused using simulation evidence for several popular empirical settings including difference-in-differences regressions, instrumental variables regressions, and regression discontinuity designs.

JEL Classification: G1, G10

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

Davidson Heath - davidsontheath@gmail.com
University of Utah

Matthew Ringgenberg - matthew.ringgenberg@eccles.utah.edu
University of Utah

Mehrdad Samadi - msamadi@mail.smu.edu
Southern Methodist University

Ingrid M Werner - werner.47@osu.edu
Fisher College of Business, The Ohio State University and CEPR

Acknowledgements
The authors thank Lucian Bebchuk, Thorsten Beck, Stephen Brown, Andrew Chen, Yong Chen, David De Angelis, Joey Engelberg, Benjamin Gillen, Campbell Harvey, Joost Impink, Jonathan Karpoff, Ye Li, Florian Peters, Peter Reiss, Alessio Saretto, Sophie Shive, Elvira Sojli, Holger Spamann, Noah Stoffman, Allan Timmermann, Michael Wittry, Michael Wolf, participants at the 15th Annual Central Bank Conference on the Microstructure of Financial Markets, the 2019 FRA-Vegas conference, the Chapman University Behavioral and Experimental Finance conference, and seminar participants at Rutgers University, SMU Cox, The Ohio State University, Texas A&M University, Tulane University, the University of Utah, and the Virtual Finance Seminar (hosted by Michigan State and the University of Illinois at Chicago).

Reusing Natural Experiments*

Davidson Heath Matthew C. Ringgenberg

Mehrdad Samadi Ingrid M. Werner

April 2020

ABSTRACT

After a natural experiment is first used, other researchers often reuse the setting, examining different outcome variables. We examine the consequences of reusing an experimental setting using two extensively studied natural experiments, business combination laws and the Regulation SHO pilot. We apply multiple hypothesis corrections and our findings suggest many results in the existing literature are false positives. We provide guidelines for inference when an experiment is reused using simulation evidence for several popular empirical settings including difference-in-differences regressions, instrumental variables regressions, and regression discontinuity designs.

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 and CEPR. The authors thank Lucian Bebchuk, Thorsten Beck, Stephen Brown, Andrew Chen, Yong Chen, David De Angelis, Joey Engelberg, Benjamin Gillen, Campbell Harvey, Joost Impink, Jonathan Karpo, Ye Li, Florian Peters, Peter Reiss, Alessio Saretto, Sophie Shive, Elvira Sojli, Holger Spamann, Noah Stoman, Allan Timmermann, Michael Wittry, Michael Wolf, participants at the 15th Annual Central Bank Conference on the Microstructure of Financial Markets, the 2019 FRA-Vegas conference, the Chapman University Behavioral and Experimental Finance conference, and seminar participants at Rutgers University, SMU Cox, The Ohio State University, Texas A&M University, Tulane University, the University of Utah, and the Virtual Finance Seminar (hosted by Michigan State and the University of Illinois at Chicago).

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 five years approximately 17% 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] Similarly, roughly 10% of published papers in the top three Accounting journals[2] and 12% of published papers in the top five Economics journals[3] also mention these terms (see Figure 1).

While the increased reliance on natural experiments has been praised for bolstering the credibility of empirical research in the social sciences (e.g., Angrist and Pischke (2010)), it is not a panacea. Natural experiments that can be used to answer research questions are difficult to find. As a result, after an experiment is first used, other researchers often reuse the setting to examine different outcome variables. Examples of natural experiments that have been reused include state-level changes in rules or laws (e.g., minimum wages, tax rates, corporate laws, contract laws, and regulations); discontinuities in membership to a particular group (e.g., Russell 3000 index membership, credit ratings, and FICO scores); and randomized controlled trials (RCTs) (e.g., the

[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] *Journal of Accounting Research, Journal of Accounting and Economics*, and *The Accounting Review.*

[3] *American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics*, and *Review of Economic Studies.*

Regulation SHO and U.S. Tick Size Pilot programs).[4] In this paper, we provide the first evidence on the consequences of reusing natural experiments; the reuse of a setting increases the likelihood of erroneously inferring that it caused an outcome variable to change (i.e., it increases the likelihood of false positives). We then show how to use multiple hypothesis testing corrections in a variety of settings, and we provide guidelines for inference when a setting is reused.

While multiple hypothesis testing is potentially an issue 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. The reason is that each separate test, while typically conducted by different researchers, is in effect part of a "family" of tests all asking the same question: What was the effect of the treatment? As researchers expand the family of tests by examining more and more dependent variables in the same setting, the likelihood of making at least one Type I error (i.e., false positive) increases. For example, assuming all null hypotheses are true, the probability of making at least one Type I error is equal to $1-(1-\alpha)^S$, where S is the number of hypotheses examined.[5] At the α=5% level when one hypothesis is examined, the probability of at least one Type I error is 5%. However, when 50 different hypotheses are examined the probability of at least one Type I error is 92.3% – this is the multiple testing problem. Put differently, the reuse of natural experiments, without correcting for multiple hypothesis testing, is undermining the credibility revolution.

In order to address the multiple testing problem, we propose using corrections that account for dependence across tests. We use the step-down procedure developed by

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

[5]Assuming the tests are mutually independent.

Romano and Wolf (2005, 2016) (henceforth Romano-Wolf). The Romano-Wolf procedure provides asymptotic control of the family-wise error rate (FwER), which is the probability of making one or more false rejections given all hypotheses considered in a family of tests.[6] While other methods exist to control the FwER (Dunn, 1961; Holm, 1979), Clarke, Romano, and Wolf (2019) show that because the Romano-Wolf procedure accounts for dependence across tests it has more power to reject false null hypotheses than other FwER methods. In other words, it is more forgiving than other methods that control the FwER.

To illustrate the multiple testing problem in natural experiments, we examine two real-world settings. Specifically, we re-examine the empirical evidence on the causal effects of treatments in two extensively studied experiments: the enactment of state business combination laws and the Regulation SHO pilot. To date, more than 120 papers have been written using these two settings.[7] 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 in order to conduct the required bootstrap for the Romano-Wolf procedure. Hence, we do not attempt to replicate previous studies. Rather, we re-evaluate the effect of the treatments on a set of outcome variables that have been studied in the literature using a common methodology across outcomes.

We propose and examine three different ways to apply the Romano-Wolf correction when natural experiments are reused. The first approach sequentially orders outcomes:

[6]For convenience of terminology, we equate control of the FwER to asymptotic control of the FwER.

[7]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. These represent lower bounds on the number of outcome variables that have actually been examined in each setting.

in other words, when we apply the multiple testing adjustment to a given outcome, we consider the results that had been previously reported at the time that the paper in question was written. This first approach effectively raises the bar for statistical significance *over time*, as more papers are written. The second approach is based on causal chain arguments. The approach sequences the results such that the null hypotheses that are most likely to be rejected (as predicted by economic theory) 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 SHO experiment. The third approach assumes that all 23 variables were explored. In a sense, this approach assumes researchers examined all of the variables at once.

In all three approaches, our evidence using Romano-Wolf suggests that many of the existing results on both business combination laws and Regulation SHO may be false positives.[8] 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 find only one significant result using business combination laws and three significant results

[8]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 3, below.

using 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.

It is possible that published papers include only a subset of the true variables examined – researchers may have examined variables and not included them in a paper. In order to get a sense for the possible severity of the multiple testing problem and to avoid data-snooping critiques regarding our choice of dependent variables, we next 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 66 of the 293 outcomes are statistically significant at the 5% level before multiple testing adjustments. For Regulation SHO, we find that 24 of the 293 outcomes are statistically significant at the 5% level before multiple testing adjustments. After applying the Romano-Wolf correction, only eight business combination laws outcomes survive while no Regulation SHO outcomes survive. Moreover, in both settings we find that the corrected p-values increase 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). These results highlight the challenge that we face as a profession when experiments can be reused by different researchers and not all tests are revealed publicly.

Finally, we provide guidance for researchers who reuse a setting. We use simulations to derive adjusted critical values as a function of the number of previously examined variables. To ensure that our findings apply to a wide range of research designs, we examine three popular econometric techniques: difference-in-differences regressions, in-

strumental variables regressions, and regression discontinuity designs. Moreover, within the difference-in-differences regressions, we examine two sub-cases based on popular research designs: (i) the staggered introduction of a state-level shock, which has variation across firms and across time and (ii) an RCT in which firms are randomly selected for treatment at one point in time. For each of these settings and techniques, we simulate the exogenous independent variables and then we examine the same 293 CRSP-Compustat outcome variables that we use in the data-mining exercises for business combination laws and Regulation SHO.[9]

For the commonly studied outcomes and corresponding dependence structures that we examine in this paper, we find that adjusted t-statistic critical values evolve at a similar rate across a broad range of empirical settings and methods. Consequently, we provide adjusted t-statistic critical values that can be used by future researchers reusing natural experiments across a wide-variety of settings and econometric techniques. In order to address the multiple testing problem, a good rule of thumb is that a new hypothesis should have a t-statistic of at least 2.5 with 5 prior findings and 3.0 with 20 prior findings using the same natural experiment.

Our results contribute to a growing literature on multiple testing in economics. Even Edward 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

[9]Because the Romano-Wolf critical values may be influenced by the dependence structure of the data, we use real data for the outcome variables to ensure they are representative of data commonly used in academic studies.

weak power to reject false null hypotheses. White (2000) develops a bootstrap-based 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. In a follow-up paper, Romano and Wolf (2016) show how to use the step-down procedure to calculate 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 (2019) propose using a procedure based on Romano and Wolf (2010) to address the problem of multiple hypothesis testing in field experiments. In their setting, a researcher has control over the parameters of an experiment and tests multiple hypotheses at the same time. By contrast, in our setting, researchers reuse a particular natural experiment without considering prior (or subsequent) work on that same experiment.

Several recent papers adjust for multiple testing in other finance settings, including asset pricing tests (Harvey and Liu (2013), Harvey, Liu, and Zhu (2016)), trading strategies (Harvey and Liu (2014), Chordia, Goyal, and Saretto (2017)), the study of anomalies (Hou, Xue, and Zhang (2018)), predicting aggregate stock returns (Engelberg, McLean, Pontiff, and Ringgenberg (2019)), and fund performance (Giglio, Liao, and Xiu (2019), Andrikogiannopoulou and Papakonstantinou (2019)).

The issues we raise are related to the general problem of p-hacking discussed by Harvey (2017) in his American Finance Association Presidential address.[10] The topic

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

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) find suspicious bunching of p-values close to cutoffs.[11] Andrews and Kasy (2019) note that certain empirical results are more likely to be published, leading to publication bias. They propose bias-corrected estimators and confidence sets that take the conditional probability of publication as a function of a study's results into account. In contrast to these existing studies, our paper is the first to examine the reuse of natural experiments.

The rest of the paper proceeds as follows. Section 1 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-Wolf step-down procedure. Section 2 presents our main findings. Section 3 discusses key issues regarding the reuse of experiments and discusses how to account for multiple testing in practice. Section 4 concludes.

1. Data and Methodology

In this section, we discuss the construction of key variables used in our analyses. We then provide a step-by-step overview of the Romano-Wolf procedure. Finally, we explain the bootstrap process we use to implement the Romano-Wolf procedure.

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

1.1. Data

To examine the practical importance of multiple testing in natural experiments, we first re-evaluate two natural experiments that have been used over 120 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: the business combination setting uses the staggered introduction of state laws to generate variation across firms and across time. In contrast, Regulation SHO was conceived as an RCT in which firms are randomly selected for treatment at one point in time.[12] While we examine these two settings in detail, our point is applicable to all settings that have been used repeatedly in academic studies (e.g., years of schooling; state level changes in minimum wage, tax rates, corporate law, and regulation; and other regulatory experiments such as the U.S. tick size pilot). To illustrate this point, in Section 2.3 we also provide simulation evidence for a broader range of empirical settings including the staggered introduction of a state-level shock, a randomized control trial, instrumental variables regressions, and a regression discontinuity design.

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 re-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 dependent variables. Because some of the existing literature uses data that is not publicly available while

[12] As we discuss below, while regulation SHO was conceived as an RCT, the study is now effectively being used as a natural experiment: more than 40 papers have been written using the setting to examine hypotheses that were not intended to be part of the original experiment.

other variables have limited sample periods, we examine a subset of 23 variables from the existing literature. These 23 dependent variables are listed in Table 1 and their construction is further detailed in Appendix Table A1.

1.1.1. 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).[13] This sample consists of annual Compustat data from 1976 through 1995, excluding financial firms, utilities and observations with missing/negative sales or total assets. As in Karpoff and Wittry (2018), our final sample consists of 10,213 firms and 88,648 firm-year observations. We follow the existing literature and winsorize all continuous outcome variables at the 0.5% and 99.5% levels.

1.1.2. Regulation SHO

Regulation SHO was a randomized control 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,

[13]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.

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 if 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 1. 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. 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), we winsorize all continuous outcome variables at the 1% and 99% levels.

<Insert Tab. 1>

1.1.3. Outcome Data Mining

To examine the severity of the multiple testing issue, we also collect a sample of all variables in Compustat and CRSP, including commonly used transformations of each variable. In order to obtain 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.[14] 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 approach results in 96 raw Compustat variables,

[14]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.

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). We winsorize all dependent variables at the 2.5% and 97.5% levels.[15]

1.2. Romano and Wolf Procedure

Using the data discussed above, we examine several applications of the Romano-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 total rejections. Yet other methods control the ratio of false rejections to rejections, known as the false discovery proportion (FDP). These different approaches have different merits. As the number of hypotheses being tested becomes larger, controlling the FwER becomes a relatively 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.[16]

[15]In the business combination law and regulation SHO settings, we winsorize using cutoffs common to the papers in those literatures. For the data mining section, we winsorize at the 2.5% and 97.5% levels due to extreme outliers. Our conclusions are robust to alternate winsorization choices.

[16]See Harvey et al. (2016) for more on this issue.

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. To control the FwER, we employ the Romano-Wolf procedure. For a given natural experiment 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, in the 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 all dependent variables by resampling the actual data with 1,000 replications.

   (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.[17] 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, in our re-evaluation of business combination laws, we have $23 \times 1{,}000 = 23{,}000$ difference-in-differences regressions. We retain the 1,000 treatment effect t-statistics for each dependent variable to build a distribution of significance levels.

[17] 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. Alternatively, Romano and Wolf (2005) propose to resample from the "raw" data and, afterwards, center the bootstrap "null statistics".

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 bootstrapped t-statistics across all dependent variables ($t_S^{*,m}$).

   (a) Starting with the dependent variable with the largest actual absolute t-statistic, we calculate the Romano and Wolf (2016) adjusted p-value as

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

   where M is the number of bootstrap samples (in our case M = 1,000). The procedure counts the fraction of times the absolute bootstrap t-statistics are greater than or equal to the actual absolute t-statistic.

   (b) We impose a monotonocity condition such that the p-value in each iteration must be greater than or equal to the p-value calculated in the last iteration.

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.[18] We also propose and examine two variations on the Romano-Wolf procedure: (i) sequential ordering and (ii) causal chains.

[18] 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.

(i) For sequential ordering, we add a loop outside the steps discussed above. Specifically, if S papers were written on the first date t, we perform the Romano-Wolf 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 $t+\tau$ more papers have been written, we rerun the Romano-Wolf 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 a 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-Wolf procedure as discussed above using 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-Wolf 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.[19]

[19]These two approaches have different characteristics. The causal chain approach uses economic theory to order outcomes. While we do not believe there should be much disagreement about first order effects in most settings, there may be more disagreement about higher order effects. Put differently, this approach is inherently subjective because it is an economic approach, rather than a purely statistical approach. In contrast, the sequential ordering approach is objective, but it ignores economic information. It is possible that the first outcomes examined in the literature may not be the effects predicted by economic theory.

1.3. Bootstrap

The Romano-Wolf procedure uses a bootstrap to re-sample the data. Importantly, the bootstrap procedure should preserve the underlying dependence structure in the data. To do this, we build bootstrap samples of 1,000 replicants by randomly sampling firms with replacement from each sample. Firm draws are stratified by treatment status, (for example, state of incorporation for business combination laws) in order to preserve the number of treated firms relative to control firms. After drawing firms, we generate a new firm index for the purpose of preserving degrees of freedom when absorbing fixed effects. In order to preserve the time series properties of the raw data, we draw all dates for each firm. To account for the dependence structure of tests, a common bootstrap sample is used for all outcomes for a given replicant (for example, once we draw a set of firms and dates using the bootstrap, we examine *all* outcome variables using that set of firm and dates).

2. Results

In this section, we examine the consequences of reusing natural experiments. We first examine two real-world natural experiments that have been used in more than 120 academic studies: business combination laws and Regulation SHO. We then use simulation evidence to provide critical values for use in future academic studies that reuse natural experiments.

2.1. Business Combination Laws

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-Wolf correction to our sample of 23 dependent variables from existing business combination studies. Table 1 provides an overview of these 23 variables.[20]

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

$$y_{i,j,l,s,t} = \alpha_i + \alpha_{l,t} + \alpha_{j,t} + \beta \cdot BC_{s,t} + \theta' \mathbf{x_{i,t}} + \epsilon_{i,j,l,s,t}, \tag{2}$$

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), $\mathbf{x_{i,t}}$ is a vector control variables including the natural log of book value of assets (size), size squared, firm age, and firm age squared. We include firm, state of location-year, and industry-year fixed effects and standard errors are clustered at the firm level. The results are reported in Table 2, Panel A. Of the 23 variables we re-examine, seven of

[20]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.

the variables are statistically significant at the 10% level based on annual data and our observation window and five are statistically significant at the 5% level. Before adjusting for multiple hypothesis testing, business combination laws are associated with an increase in leverage ($LEVERAGE$) and selling, general, and administrative expenses (SGA) and a reduction in asset growth ($ASSETGROWTH$), cash and marketable securities ($CASHSEC$), return on assets (ROA), sales growth ($SALESGROWTH$), and proportion of cash holdings in short-term investments (STI).

<Insert Tab. 2>

We then apply the Romano-Wolf procedure using three different approaches: (i) sequential ordering of the outcome variables; (ii) sorting outcome variables based on a causal chain; and (iii) examining all outcome variables. We first examine the sequential ordering approach which is based on the date each study was written. It answers the question: "Can we reject the null in this paper, while controlling the FwER, given the existing evidence available at the time this study was written?" Similar to Harvey et al. (2016), we implement this approach by 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-Wolf procedure for each additional outcome variable. If multiple outcomes share the same date, we sort alphabetically on variable name within each date. Panel A of Figure 2 presents p-values under single hypothesis testing and multiple hypothesis testing using the Romano-Wolf 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. Researchers should use an adjusted t-statistic critical value exceeding 2.76 after the 10th variable is examined. The far-right observation in Panel C shows that researchers should use an adjusted t-statistic critical value of approximately 3.0 to control the FwER at the 5% level when considering all 23 outcomes.

We next examine the causal chain approach, where we sequence the results such that the null hypotheses most likely to be rejected are examined first. The first outcome in the causal chain is the variable that should be directly effected by the natural experiment; we then rely on predictions from economic theory to sort other outcome variables into higher order effects. This approach has been referred to as a "best foot forward policy" in the multiple testing literature (Foster & Stine, 2008). 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. 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.[21]

[21] Appendix Figure A1, Panel A, illustrates the causal chain for business combination laws.

<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-Wolf 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.[22] 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-Wolf results confirm this: only one variable, STI, is statistically significant after applying the multiple testing correction.

<Insert Fig. 4>

Finally, we apply the Romano-Wolf approach using all 23 outcome variables at the same time. This approach addresses the question: "Can we reject the null that nothing changed as a result of the experiment?" 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

[22]Because we sequence the main effect first, the raw and adjusted p-values are identical.

all dependent variables available in two widely used databases: CRSP and Compustat. As discussed in Section 1, 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 66 of the data mined outcomes are statistically significant before adjusting for multiple testing (some of which have already been documented in the literature), only eight outcomes survive the adjusted critical value of 3.67, several of which do not have a known economic foundation. We also note that the corrected p-values increase from left to right at a roughly constant rate, consistent with the idea that observed variation in p-values is due to random chance. Overall, the results in this section suggest that many of the existing results in the business combination literature do not survive after adjusting for multiple hypothesis testing.

2.2. 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, 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) to examine the "the uptick rule" which restricted short selling so that it occurs only when the price is above the last traded price of the security. The experiment established a procedure to temporarily suspend Rule 10a-1 "the uptick rule" as well as any short-sale price test 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 price tests were removed for 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 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-Wolf correction to our sample of 23 dependent variables from existing Regulation SHO

studies.[23]

We estimate panel regressions of the form:

$$y_{i,t} = \alpha_i + \alpha_t + \beta \cdot Treatment_{i,t} + \theta' \mathbf{x_{i,t}} + \epsilon_{i,t}, \tag{3}$$

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. $\mathbf{x_{i,t}}$ is a vector control variables including the natural log of book value of assets (size), size squared, firm age, and firm age squared. We include firm and year fixed effects and standard errors are clustered at the firm level.

The results of this estimation are reported in Table 2, Panel B. Of the 23 variables we re-examine, ten are statistically significant at the 10% level based on annual data and our sample window and seven are statistically significant at the 5% level. Before adjusting for multiple hypothesis testing, Regulation SHO is associated with an increase in the number of patent citations ($CITE$), dollar effective spreads ($SPREAD$), and stock volatility ($STOCKVOL$) and it is associated with a reduction in capital expenditures plus R&D ($CAPEXRND$), discretionary accruals (DA_MJONES), long-term debt issuance ($DISSUE$), insider sales ($INSIDEDUM$), probability of informed trading

[23]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.

(PIN_EOH), number of management-sponsored proposals ($PROPOSALS$), and stock repurchases ($REPO$).

Again, we apply the Roman-Wolf 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, 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, $SPREAD$, $STOCKVOL$, PIN_EOH, and $CITE$, are statistically significant after computing adjusted p-values. The far-right observation in Panel C shows that researchers should use an adjusted t-statistic critical value of approximately 3.0 to control the FwER at the 5% 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. 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. 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.[24] We apply an iteration of the Romano-Wolf procedure for each causal chain grouping.

[24]Appendix Figure A1, Panel B, illustrates the causal chain for Regulation SHO.

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-Wolf procedure with causal 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 causal chain argument fails with the main effect. The Romano-Wolf results confirm this: only three of the remaining dependent variables in Panel B are statistically significant after applying the multiple testing correction (PIN_EOH, $CITE$, and $STOCKVOL$).

When we consider all outcome variables in Panel B of Figure 4, the same three outcomes survive. Once again, all three approaches suggest that 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 24 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 3.68. 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.

Once again, the results suggest that many of the existing results in the literature do not survive after adjusting for multiple hypothesis testing.

2.3. Simulations

Our results on business combination laws and Regulation SHO suggest that multiple testing corrections change the inferences from many existing studies. Accordingly, future papers should adjust for multiple testing when reusing a natural experiment. Ideally, researchers should replicate all existing studies and then apply the Romano-Wolf procedure. However, in practice, it is difficult to replicate all original papers. Moreover, existing papers often do not have common sample sizes, making it difficult to apply the bootstrap procedure. Accordingly, in this section we use simulation evidence to construct critical values to aid future research. Our adjusted critical values can be used to provide cutoffs for statistical significance at the 5% level even when researchers are not able to replicate all of the papers in the existing literature.

We conduct simulations for three popular econometric techniques: difference-in-differences regressions (diff-in-diff), instrumental variables regressions (IV), and regression discontinuity designs (RDD). Within the difference-in-differences regressions, we also examine two particular research designs: (i) the staggered introduction of a state-level shock, which has variation across firms and across time and (ii) an RCT in which firms are randomly selected for treatment at one point in time. For each technique and setting, we conduct repeated simulations using the 293 outcome variables from the data mining exercise (see Table A2 in the Appendix). Again, we use the real CRSP-Compustat data for the outcome variables in order to reflect real-world dependence structures between tests. We simulate the exogenous treatment variables for each tech-

nique and setting, as discussed below. By construction, the exogenous independent variables are placebos; this implies that any significant outcomes in our analysis are false positives.

2.3.1. Simulating the Reuse of Natural Experiments

In order to simulate the reuse of each natural experiment, we repeatedly sample the 293 CRSP-Compustat outcome variables without replacement for a given simulation. In each simulation, we retain all outcomes that are significant at the 5% level before correcting for multiple testing, as the set of "prior findings." This approach assumes that the set of prior findings is known to researchers conducting subsequent studies, while the samples of outcomes that were previously tested and found insignificant are not.

We then examine how adjusted t-statistic critical values evolve when taking into account the set of prior findings. We repeatedly sample a new candidate outcome variable. We compute the critical t-statistic for the new candidate outcome to be ruled as significant at the 5% level *after* applying the Romano-Wolf correction for multiple testing. Thus, the simulations are designed to resemble our sequential application of Romano-Wolf in the re-evaluations exercise, effectively raising the bar for statistical significance as more significant outcomes are discovered.[25]

[25]These simulations follow the standard practice of evaluating statistical significance under the assumption that the null hypothesis is true for all outcomes. In the Appendix we show that our conclusions are the same when we add true treatment effects to the simulation – *i.e.*, when the null hypothesis is not true for all outcomes.

2.3.2. Difference-in-differences regression using a Staggered Shock

The sample consists of Compustat firm-level data with fiscal years ending between 1974 and 2004. To simulate the staggered introduction of state-level shocks, we randomly assign the enactment of business combination laws, without replacement, to the 50 states of incorporation in the sample. We then estimate panel regressions of the form:

$$y_{i,s,t} = \alpha_i + \alpha_t + \beta \cdot Treat_{s,t} + \epsilon_{i,s,t}, \tag{4}$$

where $y_{i,s,t}$ is the outcome variable of interest for firm i in year t incorporated in state s. $Treat_{s,t}$ is an indicator variable which is equal to one if the shock has occurred in state s by year t and equal to zero otherwise. We include firm and year fixed effects and we cluster standard errors at the firm level. On average across multiple simulations, 25 of 293 outcomes have a "treatment effect" that is statistically significant at the 5% level.[26]

2.3.3. Difference-in-differences regression using a Randomized Controlled Trial

To construct the simulated RCT sample, we first randomly select a treatment year between 1984 and 1994, then collect 10 years of annual Compustat firm-level data before the treatment year and 10 years of data after the treatment year. In order to simulate treatment status, one third of the firms are randomly assigned as treated while the others are assigned as controls. We then estimate panel regressions of the form:

$$y_{i,t} = \alpha_i + \alpha_t + \beta \cdot Treat_{i,t} + \epsilon_{i,t}, \tag{5}$$

[26]This is consistent with Bertrand, Duflo, and Mullainathan (2004) who argue that difference-in-differences designs may be prone to overstate statistical significance.

where $y_{i,t}$ is the outcome variable of interest for firm i in year t; $Treat_{i,t}$ is an indicator variable equal to one if the firm is in the treated stock group and the treatment has taken effect, and equal to zero otherwise. We include firm and year fixed effects and we cluster standard errors at the firm level. On average across multiple simulations, 15 of 293 outcomes have a "treatment effect" that is statistically significant at the 5% level.

2.3.4. Instrumental Variables Regression

To construct the simulated IV sample, we simulate an endogenous independent variable (X) for the 1984 to 2004 sample period and then simulate the instrument (Z) so that it is a function of the endogenous independent variable (so that we do not have a weak instrument) and a noise term. We then estimate two-stage least-squares regressions of the form:

$$X_{i,t} = \kappa_i + \kappa_t + \gamma \cdot Z_{i,t} + \eta_{i,t}, \tag{6}$$

$$y_{i,t} = \alpha_i + \alpha_t + \beta \cdot X_{i,t} + \epsilon_{i,t}, \tag{7}$$

where $y_{i,t}$ is the outcome variable of interest for firm i in year t; $X_{i,t}$ is the endogenous independent variable, and $Z_{i,t}$ is an instrumental variable. We include firm and year fixed effects with standard errors clustered at the firm level. On average across multiple simulations, 18 of 293 outcomes have a "treatment effect" that is statistically significant at the 5% level.

2.3.5. Regression Discontinuity Design

To construct the simulated RDD sample, we randomly draw a threshold and construct an indicator variable ($Treat_{i,t}$) that takes the value one above the threshold. We then

estimate panel regressions of the form:

$$y_{i,t} = \alpha_i + \alpha_t + \beta \cdot Treat_{i,t} + \lambda \cdot X_{i,t} \cdot Treat_{i,t} + \epsilon_{i,t}, \tag{8}$$

where $y_{i,t}$ is the outcome variable of interest for firm i in year t, $Treat_{i,t}$ is an indicator variable, and $X_{i,t}$ is a linear control function that is fitted separately above and below the threshold. We include firm and year fixed effects and we use a bandwidth of 100 firms on either side of each yearly simulated threshold.[27] On average across multiple simulations, 15 of 293 outcomes have a "treatment effect" that is statistically significant at the 5% level.

2.3.6. Adjusted t-statistic Critical Values

Figure 6 shows adjusted t-statistic critical values in relation to the number of prior findings. In all four settings, one or less findings are statistically significant after multiple testing correction – that is, the Romano-Wolf procedure successfully controls the family-wise error rate (FwER).

<Insert Fig. 6>

Strikingly, the Romano-Wolf cutoff is similar in all four settings. Table 3 shows that as the set of prior findings grows, the corrected critical value also rises at similar rate across settings. With five prior findings, the average critical value is 2.50 in the staggered shock setting, 2.51 in the RCT setting, and 2.55 in the RDD and IV settings. With ten

[27] In the Appendix we show that the results are similar for other choices of the bandwidth and control function.

prior findings, the critical values become 2.81, 2.74, 2.82 and 2.87, respectively. With twenty prior findings, the new critical values are 3.01, 2.96, 3.01 and 3.05 respectively. Ideally, researchers who are reusing a natural experiment should replicate all existing studies that use the setting and apply the Romano-Wolf correction. However, if this is not possible, the results in Table 3 provide a heuristic for statistical inference when reusing a setting.

<Insert Tab. 3>

3. Discussion

Our results suggest that many of the findings in widely-studied experiments may be false positives. In this section, we discuss caveats and robustness checks. We also discuss best practices for reusing natural experiments.

3.1. Alternate Multiple Hypothesis Testing Methods

Our analyses use the Romano-Wolf procedure to control the FwER. While we believe the Romano-Wolf procedure is the most appropriate for our setting, we show in Appendix Table A3 that our conclusions are robust to using alternate multiple testing corrections such as the Bonferroni FwER procedure, the Benjamini and Hochberg (1995) FDR procedure, the Benjamini and Yekutieli (2001) FDR procedure, and the Romano and Wolf (2007) FDP procedure. The table illustrates that, regardless of the correction procedure, many of the outcome variables from business combination laws and Regulation SHO may be false positives.

Another way to address the multiple hypothesis testing problem would be to gather more data, that is, to repeat the experiment on a different universe of firms or over a non-overlapping period of data.[28] However, in practice, this is often not possible. We are careful to note that we are not saying natural experiments should never be reused; rather, our goal is to provide guidelines for improving inference when reusing a setting. In the rest of this section, we discuss best practices for reusing a natural experiment.

3.2. 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. For Regulation SHO, the experiment was designed to weaken short sale constraints by removing price-tests. Thus, the direct effect is short selling activity, which might change as a result of the experiment. For business combination laws the law changes were 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, the relevance condition requires that the treatment produces an economically and statistically significant shift in the direct-effect variable. 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

[28]Note that simply adding more observations surrounding the same experiment does not solve the problem as the source of the exogenous variation would still be the same and thus, multiple testing still applies.

from a weak instrument problem.[29]

Moreover, even if the relevance condition holds, settings may still fail the exclusion restriction which requires that a shock affects an outcome variable only through a particular mechanism. For Regulation SHO, Boehmer, Jones, and Zhang (2019) argue that lifting the uptick rule did have some significant direct effect on treated firms, but it also affected control firms through spillovers, which violates the stable unit treatment value 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. Before using a setting, researchers should provide evidence that the relevance and exclusion conditions hold *in their sample*; the fact that other papers examined a setting is not sufficient.

3.3. Compound Exclusion Restrictions

In a related point, we also note that researchers reusing an experimental setting should reconcile their exclusion restrictions with existing empirical evidence available when their study is written.[30] 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

[29]It remains possible that business combination laws and/or Regulation SHO had an effect on corporate outcomes if corporate managers *believed* that the changes would affect hostile takeovers or short selling (even though ex-post they did not). However, researchers must establish this before using the setting, and we are not aware of any such evidence.

[30]A similar point is made in Morck and Yeung (2011) who note that "each successful use of an instrument creates an additional latent variable problem for all other uses of that instrument."

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.[31] 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).

3.4. 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 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 – either by directly applying the Romano-Wolf correction or using the critical values provided in Table 3. However, we also caution that multiple testing correction methods are not a panacea; simply passing the Romano-Wolf adjusted critical values shown in Table 3 does not mean a setting is valid. Rather, multiple hypothesis adjusted p-values are just one of many inputs that should be used to assess a finding.

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

[31] It is possible to interpret the new finding as a reduced form estimate, but at a minimum, the authors need to discuss the existing evidence.

1. Researchers should verify the relevance and exclusion restrictions *in their sample* before examining higher order effects. They should also formalize the economic mechanism using causal chain arguments.

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

3. Finally, researchers should adjust for multiple testing in order to control the FwER.

4. Conclusion

Natural experiments have become an important tool for identifying the causal relation between variables. While the use of natural experiments has increased the credibility of empirical economics in many dimensions (Angrist & Pischke, 2010), we find evidence that the repeated reuse of these settings 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.

To examine the consequences of reusing a natural experiment, we re-examine two extensively studied settings: business combination laws and Regulation SHO. Combined, these settings have been used in over 120 different academic studies. We re-evaluate 46 outcome variables that were found to be significant in existing studies – our findings

suggests that many of the existing findings in the literature may be false positives due to the reuse of these settings.

We also note that business combination laws and Regulation SHO are not alone. There are many other frequently re-used natural experiments in social sciences for which our arguments apply. For example, Baldwin and Bhavnani (2015) find that the Vietnam war draft lottery has been reused 16 times in different studies (as of 2012) and Universal Demand Laws have been used in more than 30 papers to date.[32] Further, the planned SEC transaction fee pilot and the FINRA corporate bond block trade transparency pilot will likely generate much future research.

To aid future research, we provide guidelines for inference when an experiment is reused. We use simulation evidence to construct adjusted critical values as a function of the number of times a setting is examined. Our adjusted critical values cover many popular settings and designs, including difference-in-differences regressions, instrumental variables regressions, and regression discontinuity designs. Researchers should use these critical values when assessing statistical significance. We also discuss best practices for reusing a natural experiment.

Overall, the repeated use of natural experiments without accounting for multiple hypothesis testing is likely leading to many false discoveries. We hope 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.

[32]See, for example, Appel (2019).

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.

Andrikogiannopoulou, A., & Papakonstantinou, F. (2019). Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? *Journal of Finance*, *74*(5), 2667-2688.

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.

Appel, I. (2019). Governance by litigation. *Available at SSRN 253227*.

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*.

Baldwin, K., & Bhavnani, R. R. (2015). Ancillary studies of experiments: Opportunities

and challenges. *Journal of Globalization and Development*, *6*(1), 113–146.

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

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., 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*.

Clarke, D., Romano, J. P., & Wolf, M. (2019). The romano-wolf multiple hypothesis correction in stata. *Available at SSRN Abstract 3513687*.

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.

Giglio, S., Liao, Y., & Xiu, D. (2019). Thousands of alpha tests. *Available at SSRN Abstract 3259268*.

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.

Hou, K., Xue, C., & Zhang, L. (2018). Replicating anomalies. *Review of Financial*

Studies, *forthcoming*.

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.

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. (2019). 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.

Morck, R., & Yeung, B. (2011). Economics, history, and causation. *Business History Review*, *85*(Spring), 39–63.

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.

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.

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

Romano, J. P., & Wolf, M. (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics*, *35*(4), 1378–1408.

Romano, J. P., & Wolf, M. (2010). Balanced control of generalized error rates. *The Annals of Statistics*, *38*(1), 598–633.

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*.

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*.

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: Outcome Variable Re-evaluations

This table presents the list 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. *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 examined at the annual frequency by firm fiscal years. See Appendix Table A1 for a more detailed explanation of the construction of these variables.

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 2: Outcome Variable Re-evaluation Estimates

This table presents treatment coefficients of the re-evaluated outcome variables. Panel A presents outcomes from the business combination law literature. 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. $BC_{s,t}$ 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, $\mathbf{x_{i,t}}$ is a vector control variables including the natural log of book value of assets (size), size squared, firm age, and firm age squared. We include firm, state of location-year, and industry-year fixed effects and standard errors are clustered at the firm level. Panel B presents outcomes from the Regulation SHO literature. 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. $\mathbf{x_{i,t}}$ is a vector control variables including the natural log of book value of assets (size), size squared, firm age, and firm age squared. We include firm and year fixed effects and 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		
Outcome	*BC*	*t*-statistic	Outcome	*Treatment*	*t*-statistic
AF_ERROR	−0.061	−0.84	ASSETGROWTH	−0.012	−1.00
AMIHUD	0.000	−1.63	CAPEXRND	−0.004*	−1.66
ASSETGROWTH	−0.044*	−1.83	CASH	−0.002	−0.41
CAPEX	0.003	1.63	CITE	0.057***	3.09
CASHSEC	−0.008**	−2.00	DA_MJONES	−0.007**	−2.04
DA_MJONES_ABS	−0.003	−1.44	DISSUE	−0.014*	−1.77
DISP	0.009	0.39	DIV	0.000	−0.41
DIV	0.000	−1.05	EINDEX	0.023	1.34
DIVIDENDPAYOUT	−0.016	−0.96	EQISSUE	0.003	0.83
EQCH	−0.009	−0.50	FBIAS	0.000	0.55
EQISSUE	−0.009	−0.49	FORCED	−0.002	−0.34
LEVERAGE	0.023**	2.09	INACCURACY	0.001	1.24
LOG_CITPAT	0.000	0.01	INSIDEDUM	−0.026*	−1.83
LOG_PATENTS	0.002	0.76	OP_EQ_DOL	1.044	0.69
NUMEST	−0.060	−0.78	OP_EQ_NUM	0.091	0.06
PPEGROWTH	−0.016	−1.22	PIN_EOH	−0.006***	−3.66
ROA	−0.017*	−1.68	PROPOSALS	−0.060**	−1.98
SALESGROWTH	−0.273**	−2.23	REPO	−0.006***	−2.74
SGA	0.018**	1.99	SIR	0.000	0.06
SKEW	0.008	0.24	SKEW	−0.040	−0.88
STI	−0.042***	−4.00	SPREAD	0.002**	2.16
STOCKVOL	0.000	0.20	STOCKVOL	0.002***	4.32
TAKEOVER	0.002	1.19	VALUE	0.004	0.21

Table 3: Multiple Testing Corrected Critical Values by Econometric Method

This table presents adjusted t-statistic critical values at the 5% level for three different econometric methods: difference-in-differences regressions ("Diff-in-Diff" shown in columns (2) and (3)), instrumental variables regressions ("IV" shown in column (4)), and regression discontinuity designs ("RDD" shown in column (5)). Within the difference-in-differences analysis, we examine two different settings: the staggered introduction of a state-level shock ("Staggered Shock" shown in column (2)) and a randomized controlled trial ("RCT" shown in column (3)). In each row, we present Romano-Wolf adjusted critical values for the 5% significance level given the number of prior results shown in Column (1). For each method and setting, we simulate the exogenous independent variable. We then examine 293 different dependent variables drawn from Compustat and CRSP with pre-specified coverage at the annual frequency. For each econometric method, we develop the set of "prior results" by repeatedly sampling the 293 outcome variables without replacement and we retain outcomes that are significant at the 5% level before correcting for multiple testing. For each new finding, we then apply an iteration of the Romano-Wolf correction including the new finding and prior findings and then compute adjusted t-statistic critical values. For each econometric method, we conduct 20 simulations using the procedure described above. The 293 dependent variables are listed in Appendix Table A2.

	Diff-in-Diff		IV	RDD
# Prior Results	Staggered Shock	RCT		
(1)	(2)	(3)	(4)	(5)
1	2.09	2.14	2.09	2.13
2	2.19	2.26	2.20	2.27
3	2.29	2.38	2.30	2.40
4	2.40	2.46	2.43	2.49
5	2.50	2.52	2.55	2.59
6	2.60	2.58	2.64	2.66
7	2.68	2.62	2.72	2.71
8	2.74	2.65	2.78	2.76
9	2.79	2.68	2.83	2.80
10	2.83	2.72	2.87	2.83
11	2.86	2.74	2.90	2.86
12	2.88	2.77	2.92	2.89
13	2.91	2.79	2.94	2.91
14	2.93	2.82	2.96	2.93
15	2.95	2.84	2.98	2.95
16	2.97	2.87	3.00	2.97
17	2.98	2.88	3.01	2.99
18	3.00	2.90	3.02	3.00
19	3.01	2.92	3.03	3.01
20	3.02	2.94	3.05	3.03

Figure 1: Top Finance, Accounting and Economics Publications Mentioning Experiments. This figure presents the annual fraction of publications in top Finance (*The Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies*), top Accounting (*Journal of Accounting and Economics*, *Journal of Accounting Research*, and *The Accounting Review*), and top Economics (*American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*) journals using the terms "natural experiment(s)", "quasi(-)natural experiment(s)", and "regulatory experiment(s)" from 1988 through 2019. The Electronic Journal Center (EJC), JSTOR, and EconLit, plus journal web-sites are searched to find the number of published articles per year using the terms of interest while ISI Web of Science is used to obtain the annual total number of published articles by year in each journal.

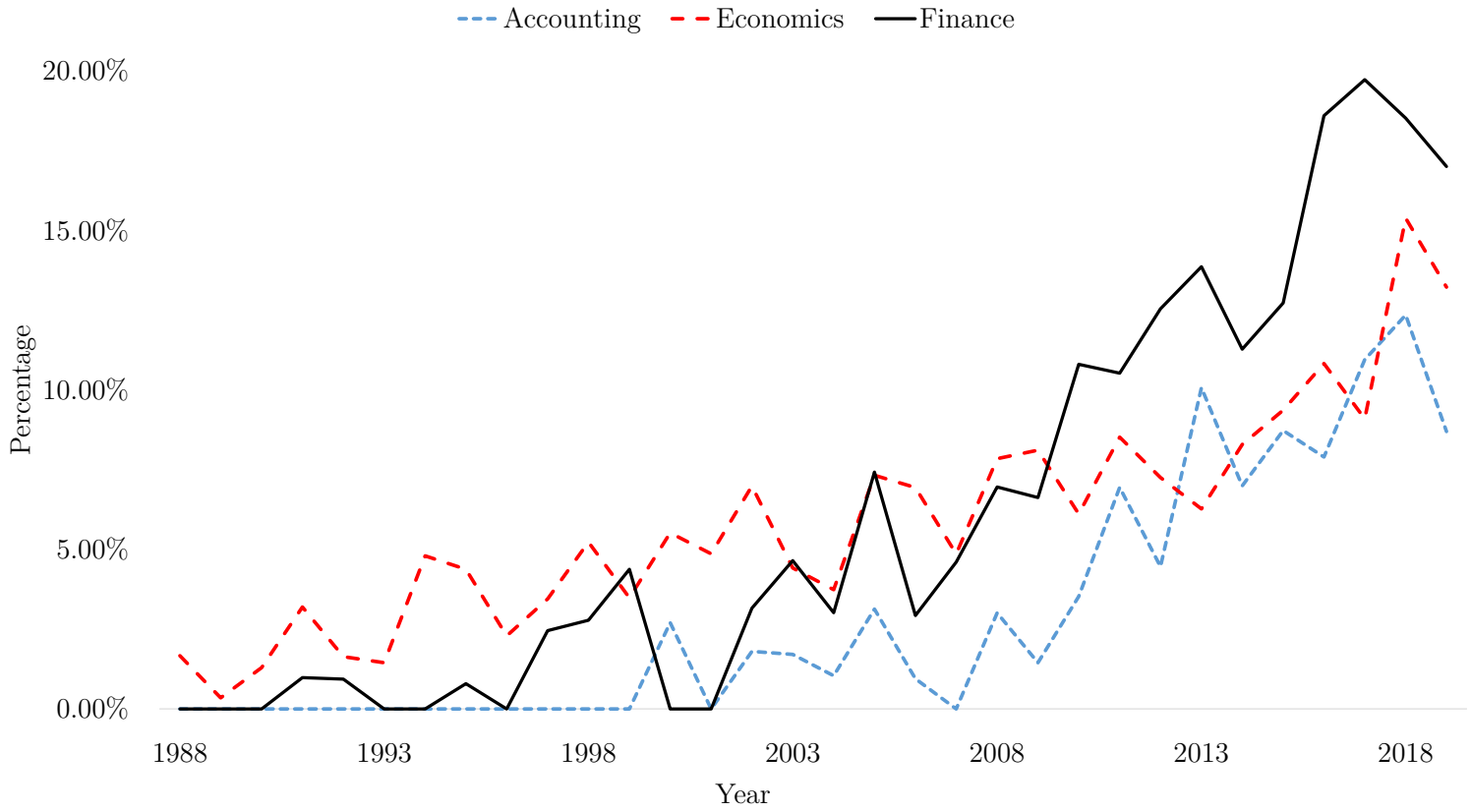


Figure 2: Outcome Re-evaluations, First Draft Date Ordering of Outcomes. This figure presents multiple testing corrected p-values and critical values for a set of outcomes previously examined in studies using business combination laws and Regulation SHO. Statistics are obtained by sequencing outcome variables using their first available draft date and applying an iteration of the Romano-Wolf 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 t-statistic critical values where the horizontal axis is the number of outcomes included in the Romano-Wolf procedure. Red dashed lines represent the five percent level of statistical significance. These 23 dependent variables are listed in Table 1 and their construction is further detailed in Appendix Table A1.

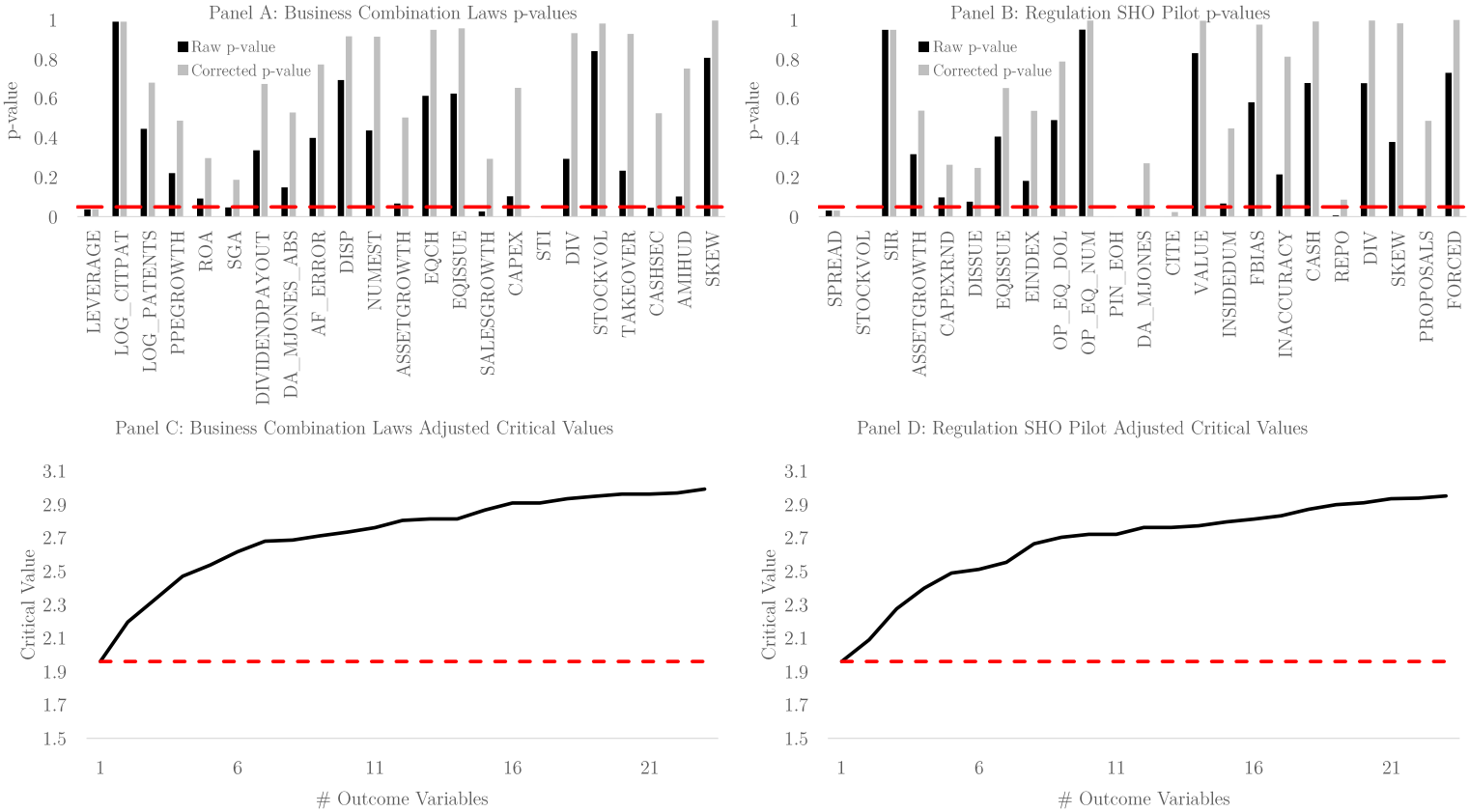


Figure 3: Outcome Re-evaluations, Causal Chain Ordering of Outcomes. This figure presents multiple testing corrected p-values and critical values for a set of outcomes previously examined in studies using business combination laws and Regulation SHO. Statistics are obtained by sequencing outcome variables using causal chain arguments and applying an iteration of the Romano-Wolf 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 t-statistic critical values where the horizontal axis is the number of causal chain orders included in the Romano-Wolf procedure. Red dashed lines represent the five percent level of statistical significance. These 23 dependent variables are listed in Table 1 and their construction is further detailed in Appendix Table A1.

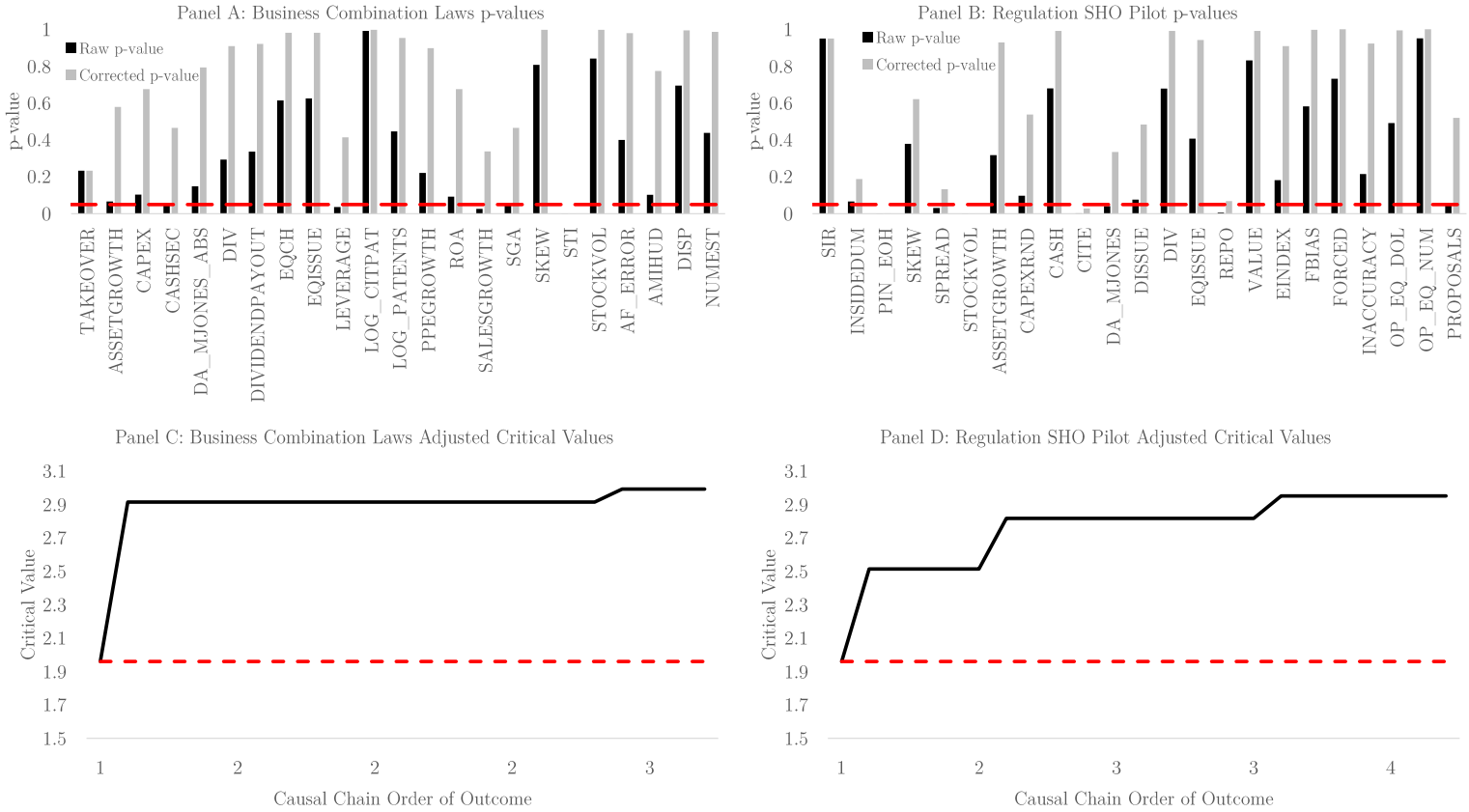


Figure 4: Outcome Re-evaluations, All Outcomes. This figure presents multiple testing corrected p-values when all 23 outcomes we re-examine are considered. We apply one iteration of the Romano-Wolf procedure to the 23 outcomes altogether, Panel A presents results for the outcomes from the business combination 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 1 and their construction is further detailed in Appendix Table A1.

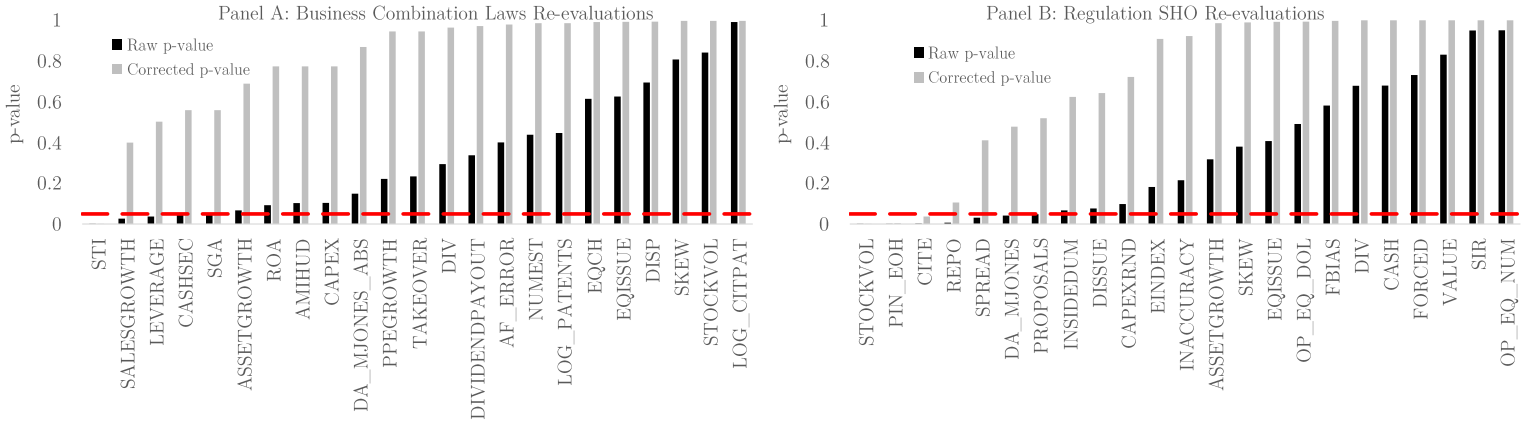


Figure 5: Outcome Re-evaluations, Data Mining. This figure presents multiple testing corrected p-values when a set of 293 outcomes drawn from Compustat and CRSP with pre-specified coverage at the annual frequency are considered. We apply one iteration of the Romano-Wolf procedure to the 293 outcomes 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.

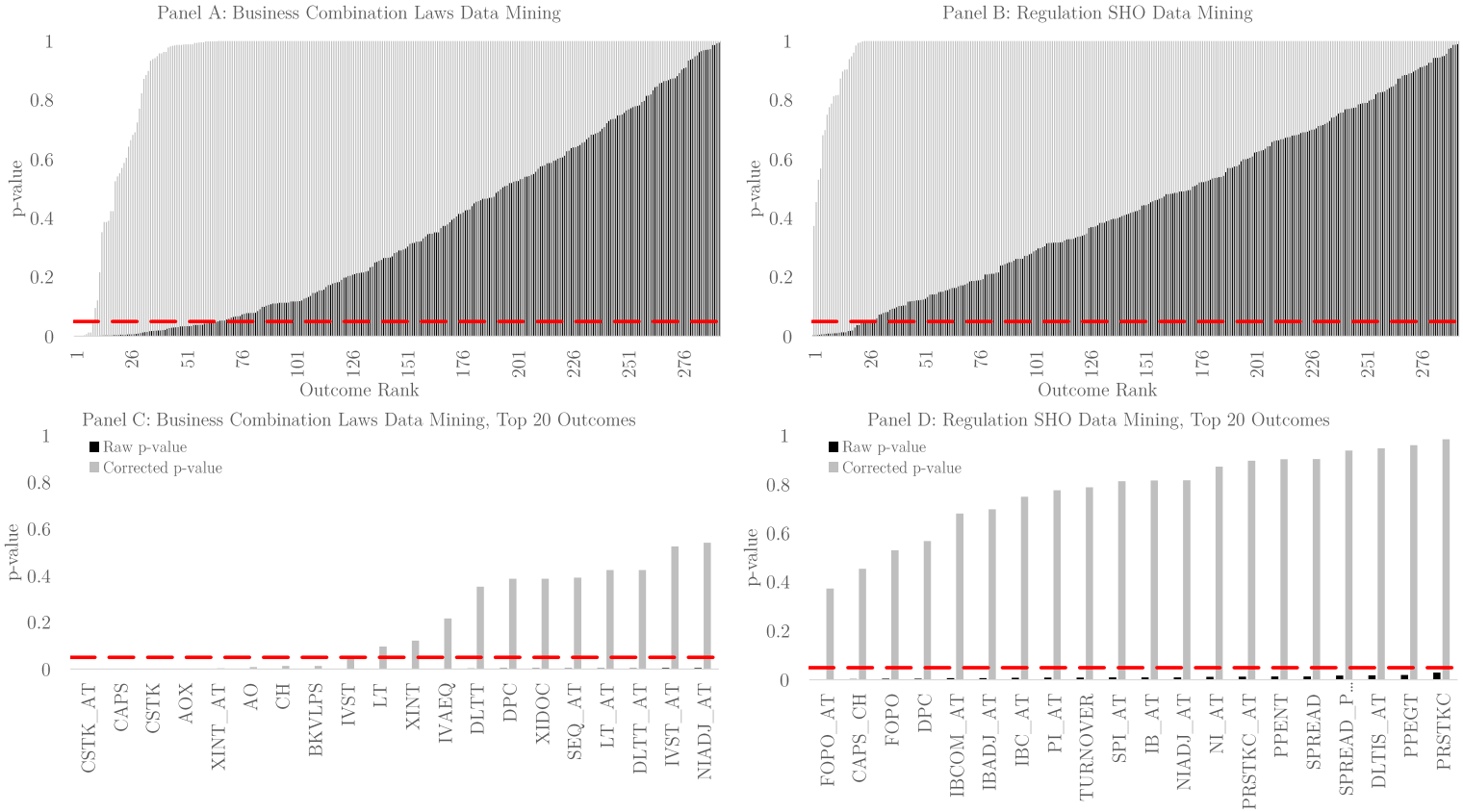
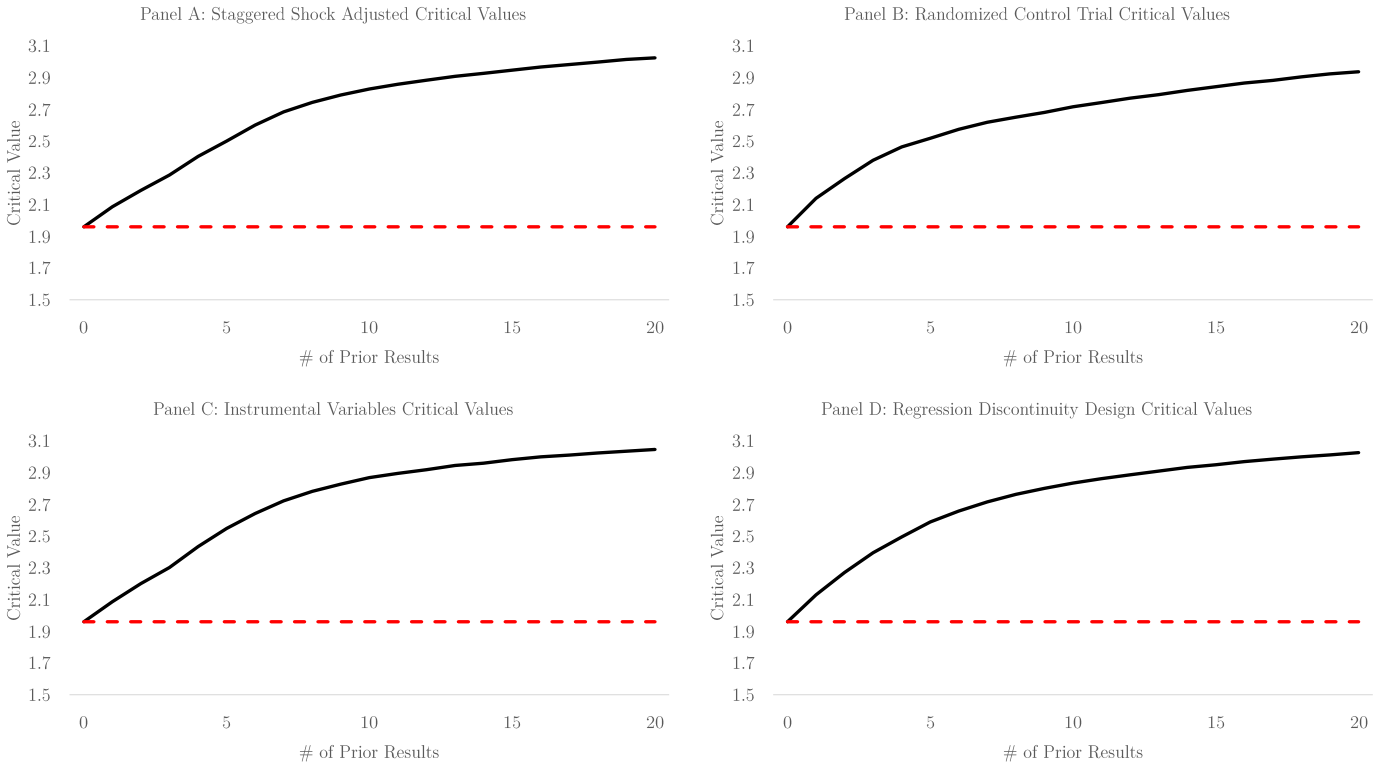


Figure 6: Multiple Testing Corrected Critical Values by Econometric Method. This figure presents adjusted t-statistic critical values for four simulated settings: the staggered introduction of a state-level shock, an RCT, an IV, and an RDD setting. We examine a set of 293 outcomes drawn from Compustat and CRSP with pre-specified coverage at the annual frequency, conducting 20 simulations in each setting and report averages. In each simulation, we repeatedly sample the 293 outcome variables without replacement, retaining outcomes that are significant at the 5% level before correcting for multiple testing, as the set of "prior findings." For each new finding, we apply an iteration of the Romano-Wolf correction including the new finding and prior findings, computing adjusted t-statistic critical values. Panel A presents results for the staggered shock, Panel B presents results for RCTs, Panel C presents results for IVs, and Panel D presents results for RDDs. The data mined outcomes are listed in Appendix Table A2.



5. Appendix to "Reusing Natural Experiments"

This appendix provides additional empirical evidence to supplement the analyses provided in the main text. Below, we briefly discuss each of the included figures and tables.

1. Table A1 provides detailed data definitions for each of the variables we use in the regulation SHO and business combinations analyses.

2. Table A2 provides detailed data definitions for the 293 variables we examine in Table 3.

3. Table A3 presents adjusted critical values (similar to Table 3 of the main text) using alternate methods to adjust for multiple hypothesis testing; our conclusions are unchanged.

4. Figures A1 displays causal chains for business combination laws (Panel A) and Regulation SHO (Panel B).

5. Figures A2 through A6 present graphical evidence on the robustness of the Romano-Wolf adjusted critical values we construct in the simulations; we examine a variety of alternate specifications and the conclusions are unchanged.

Table A1: Definition of Dependent Variables used in Re-evaluation of Business Combination Laws and Regulation SHO

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: Definition of Dependent Variables used in Calculation of Critical Values

This table presents the list of CRSP/Compustat outcomes examined in Table 3 of the main text. We examine the raw version of each variable, as well as two possible transformations of each variable, resulting in 293 different outcome variables. *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

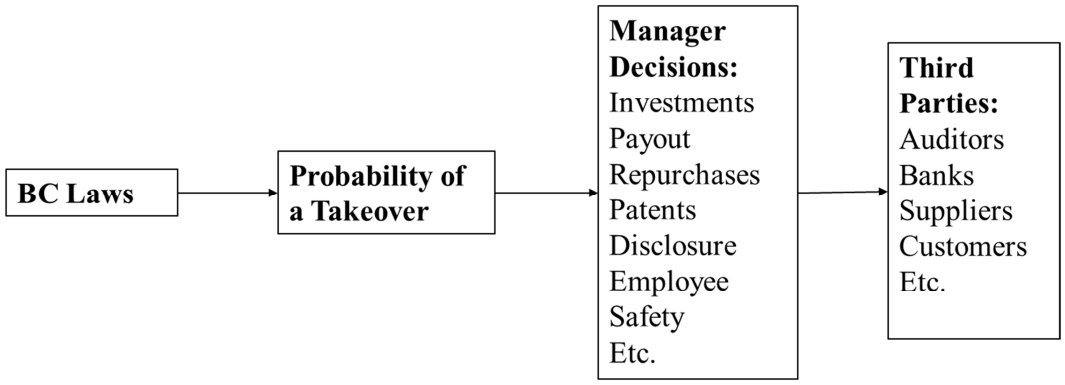
Table A3: Critical Values for Alternate Multiple Testing Correction Methods

This table presents adjusted t-statistic critical values from alternate multiple testing correction procedures. Panels A and B present results for the 23 re-evaluated business combination laws and Regulation SHO outcomes, respectively. Panels C and D present results for the 293 data mined business combination laws and Regulation SHO outcomes, respectively. Adjusted t-statistic critical values are computed using the Bonferroni FwER procedure, the Benjamini and Hochberg (1995) FDR procedure, the Benjamini and Yekutieli (2001) FDR procedure, and the Romano and Wolf (2007) FDP procedure. The FwER and FDR are controlled at the 5% level. The FDP is controlled at the 5% proportion and level.

Panel A: Business Combination Laws Re-evaluations		Panel B: Regulation SHO Re-evaluations	
Procedure	Critical Value	Procedure	Critical Value
FwER: Bonferroni	3.07	FwER: Bonferroni	3.07
FwER: Romano-Wolf	2.99	FwER: Romano-Wolf	2.95
FDR: Benjamini-Hochberg	3.07	FDR: Benjamini-Hochberg	2.62
FDR: Benjamini-Yekutieli	3.43	FDR: Benjamini-Yekutieli	3.24
FDP: Romano-Wolf	2.99	FDP: Romano-Wolf	2.95
Panel C: Business Combination Laws Data Mining		Panel D: Regulation SHO Data Mining	
Procedure	Critical Value	Procedure	Critical Value
FwER: Bonferroni	3.76	FwER: Bonferroni	3.76
FwER: Romano-Wolf	3.67	FwER: Romano-Wolf	3.68
FDR: Benjamini-Hochberg	2.96	FDR: Benjamini-Hochberg	3.76
FDR: Benjamini-Yekutieli	3.69	FDR: Benjamini-Yekutieli	4.19
FDP: Romano-Wolf	3.67	FDP: Romano-Wolf	3.68

Figure A1: Causal Chains. This figure illustrates causal chain diagrams for the two natural experiments we reevaluate. Panel A displays the causal chain for Business Combination laws and Panel B displays the causal chain for Regulation SHO.

Panel A: Business Combination Laws



Panel B: Regulation SHO

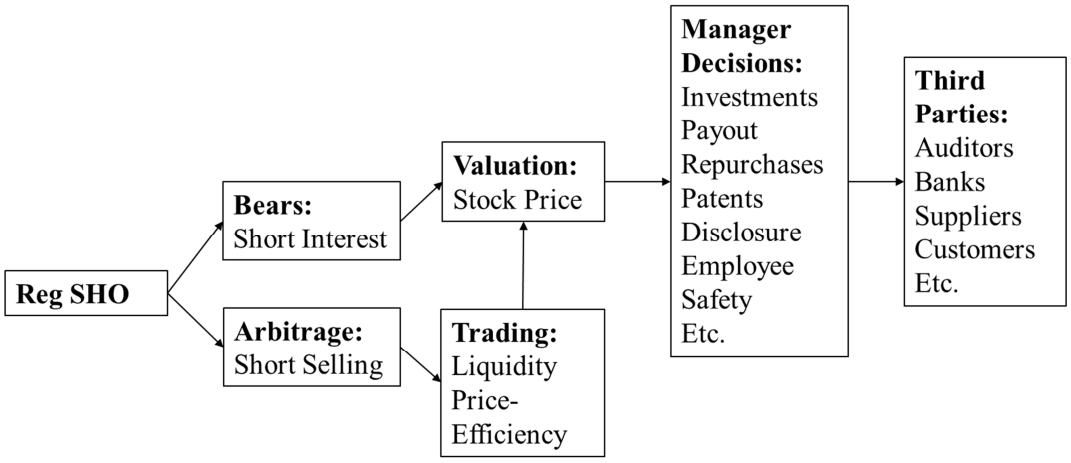


Figure A2: Simulation Evidence, 10 True Treatment Effects. This figure presents adjusted t-statistic critical values for three simulated settings: the staggered introduction of a state-level shock, an RCT, and an RDD setting. We examine a set of 293 outcomes drawn from Compustat and CRSP with pre-specified coverage at the annual frequency, conducting 20 simulations in each setting and report averages. We add 10 outcomes which are a linear function of the treatment with noise. The treatment effects are generated so that they have an individual uncorrected t-statistic of approximately 3.0 on average. That is, the simulated natural experiment is adequately powered to detect the treatment effect in a single hypothesis test (Bloom, 1995). In each simulation, we repeatedly sample the 303 outcome variables without replacement, retaining outcomes that are significant at the 5% level before correcting for multiple testing, as the set of "prior findings." For each new finding, we apply an iteration of the Romano-Wolf correction including the new finding and prior findings, computing adjusted t-statistic critical values. Panel A presents results for the staggered shock, Panel B presents results for RCTs, and Panel C presents results for RDDs. The data mined outcomes are listed in Appendix Table A2.

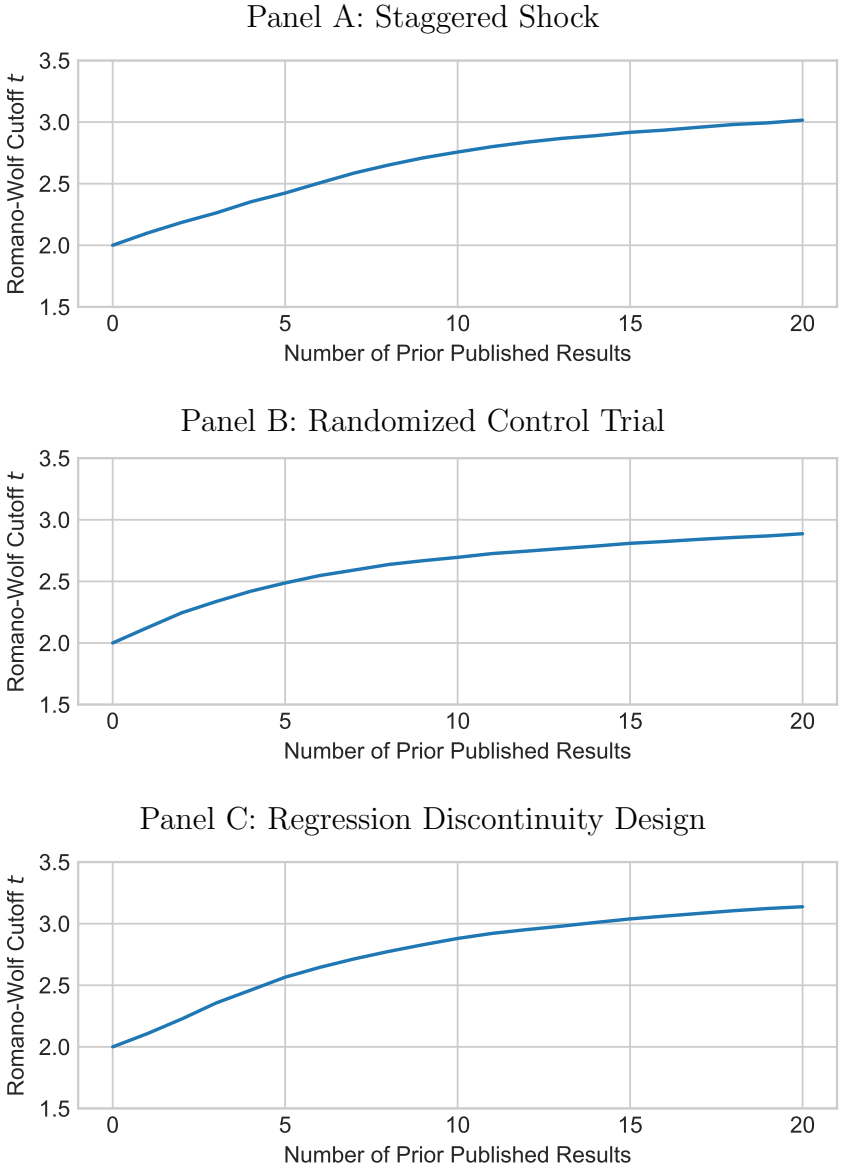


Figure A3: Simulation Evidence, 20 True Treatment Effects. This figure presents adjusted t-statistic critical values for three simulated settings: the staggered introduction of a state-level shock, an RCT, and an RDD setting. We examine a set of 293 outcomes drawn from Compustat and CRSP with pre-specified coverage at the annual frequency, conducting 20 simulations in each setting and report averages. We add 20 outcomes which are a linear function of the treatment with noise. The treatment effects are generated so that they have an individual uncorrected t-statistic of approximately 3.0 on average. That is, the simulated natural experiment is adequately powered to detect the treatment effect in a single hypothesis test (Bloom, 1995). In each simulation, we repeatedly sample the 313 outcome variables without replacement, retaining outcomes that are significant at the 5% level before correcting for multiple testing, as the set of "prior findings." For each new finding, we apply an iteration of the Romano-Wolf correction including the new finding and prior findings, computing adjusted t-statistic critical values. Panel A presents results for the staggered shock, Panel B presents results for RCTs, and Panel C presents results for RDDs. The data mined outcomes are listed in Appendix Table A2.

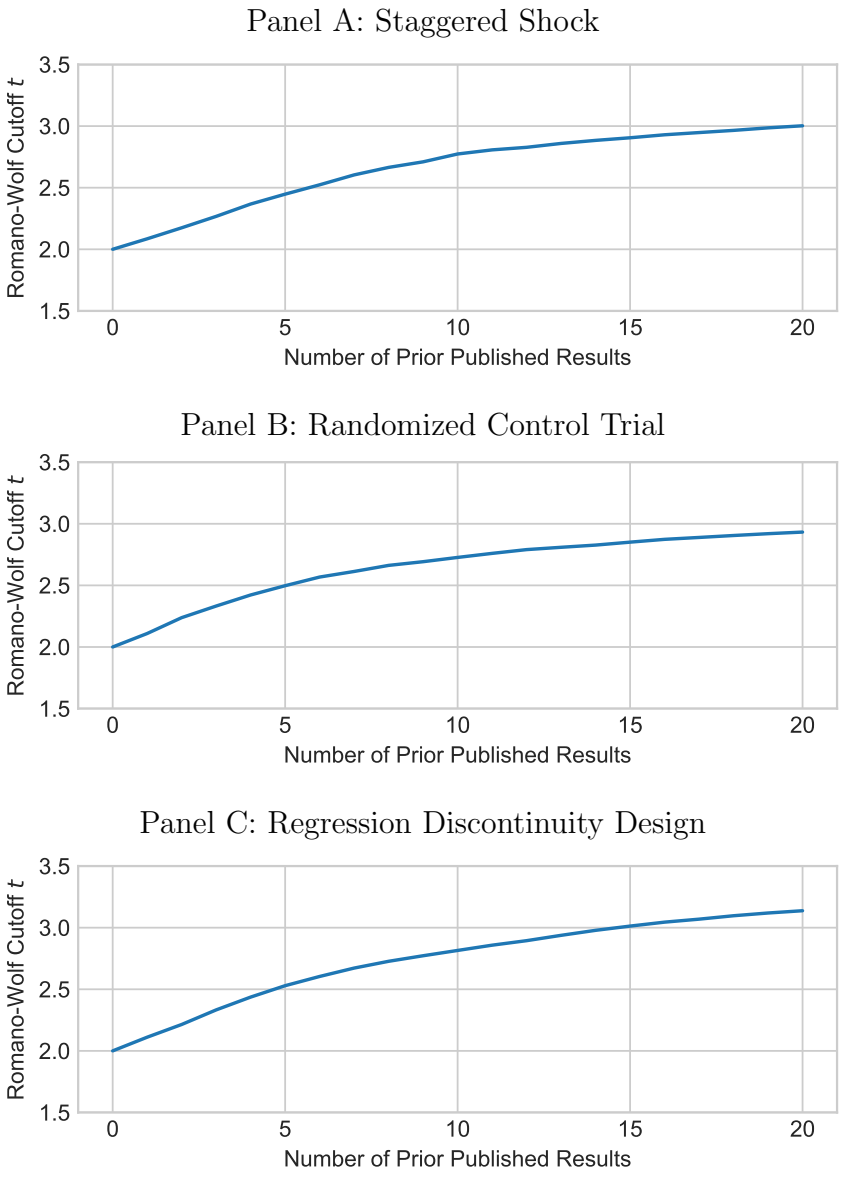
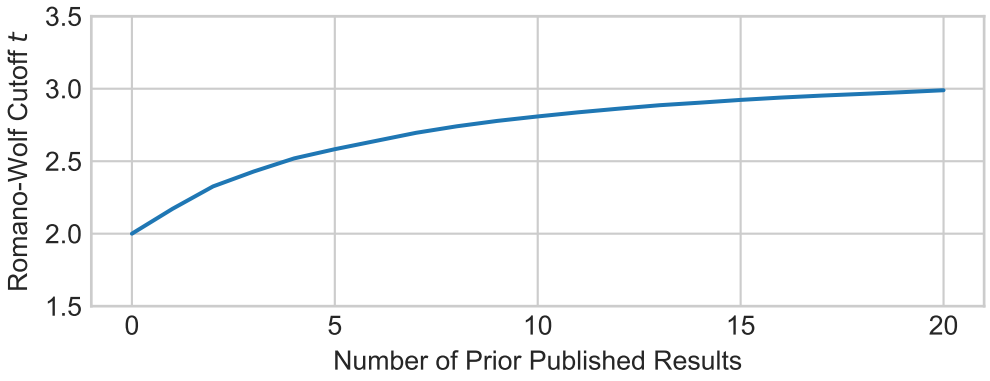


Figure A4: Simulation Evidence, RCT Treated Fraction. This figure presents adjusted t-statistic critical values for simulated RCTs. We examine a set of 293 outcomes drawn from Compustat and CRSP with pre-specified coverage at the annual frequency, conducting 20 simulations in each setting and report averages. We vary the proportion of treated firms to $\frac{1}{2}$ and $\frac{2}{3}$. In each simulation, we repeatedly sample the 293 outcome variables without replacement, retaining outcomes that are significant at the 5% level before correcting for multiple testing, as the set of "prior findings." For each new finding, we apply an iteration of the Romano-Wolf correction including the new finding and prior findings, computing adjusted t-statistic critical values. Panel A presents results for simulations in which $\frac{1}{2}$ of the firms are treated. Panel B presents results for simulations in which $\frac{2}{3}$ of the firms are treated. The data mined outcomes are listed in Appendix Table A2.

Panel A: RCT, Treated fraction = $\frac{1}{2}$



Panel B: RCT, Treated fraction = $\frac{2}{3}$

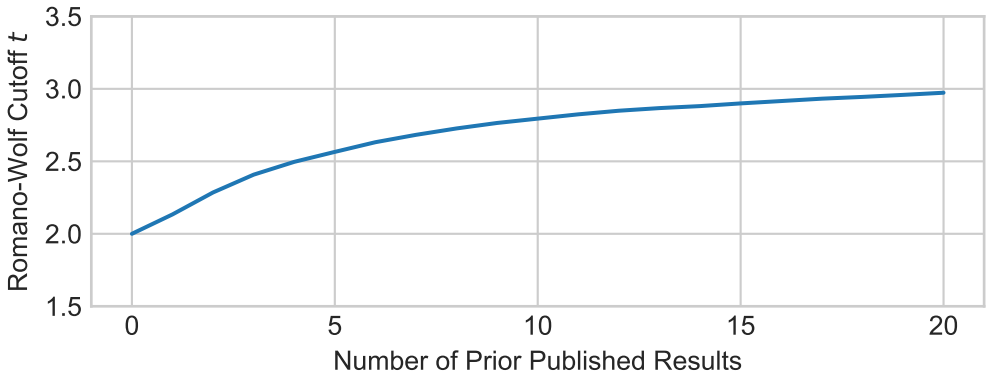
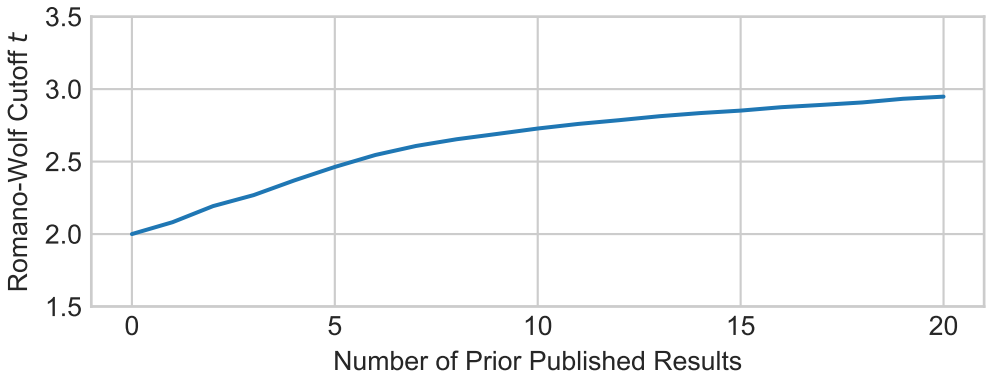


Figure A5: Simulation Evidence, RDD Bandwidth. This figure presents adjusted t-statistic critical values for simulated RDD settings. We examine a set of 293 outcomes drawn from Compustat and CRSP with pre-specified coverage at the annual frequency, conducting 20 simulations in each setting and report averages. We vary the bandwidth of the sample around the treatment threshold to $+/-250$ firms and $+/-500$ firms. In each simulation, we repeatedly sample the 293 outcome variables without replacement, retaining outcomes that are significant at the 5% level before correcting for multiple testing, as the set of "prior findings." For each new finding, we apply an iteration of the Romano-Wolf correction including the new finding and prior findings, computing adjusted t-statistic critical values. Panel A presents results for simulations with a bandwidth of $+/-250$ firms. Panel B presents results for simulations with a bandwidth of $+/-500$ firms The data mined outcomes are listed in Appendix Table A2.

Panel A: RDD, Bandwidth +/-250 firms



Panel B: RDD, Bandwidth +/-500 firms

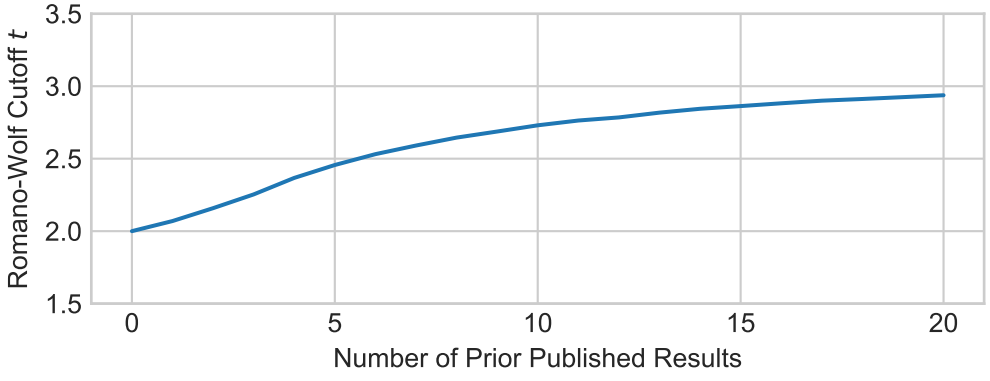
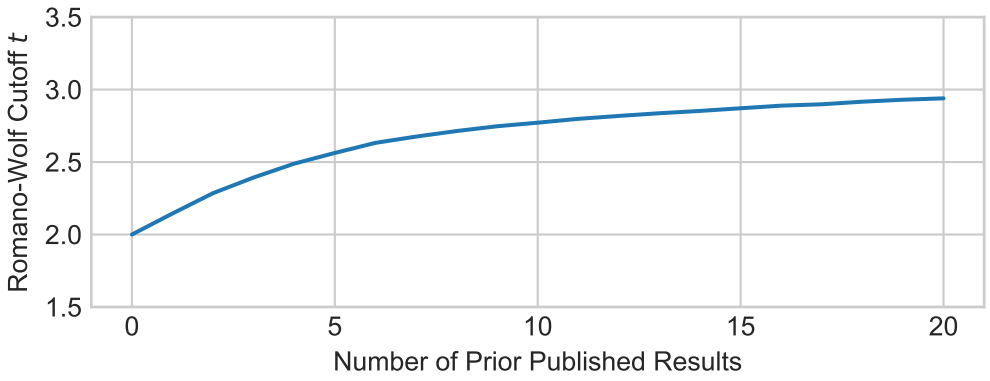


Figure A6: Simulation Evidence, RDD Control Function. This figure presents adjusted t-statistic critical values for simulated RDD settings. We examine a set of 293 outcomes drawn from Compustat and CRSP with pre-specified coverage at the annual frequency, conducting 20 simulations in each setting and report averages. We vary the control function to quadratic and cubic control functions. In each simulation, we repeatedly sample the 293 outcome variables without replacement, retaining outcomes that are significant at the 5% level before correcting for multiple testing, as the set of "prior findings." For each new finding, we apply an iteration of the Romano-Wolf correction including the new finding and prior findings, computing adjusted t-statistic critical values. Panel A presents results for simulations with a quadratic control function. Panel B presents results for simulations with a cubic control function. The data mined outcomes are listed in Appendix Table A2.

Panel A: RDD, quadratic control function



Panel B: RDD, cubic control function