The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXVIII, NO. 4 • AUGUST 2023

Reusing Natural Experiments

DAVIDSON HEATH, MATTHEW C. RINGGENBERG, MEHRDAD SAMADI, and INGRID M. WERNER*

ABSTRACT

After a natural experiment is first used, other researchers often reuse the setting, examining different outcome variables. We use simulations based on real data to illustrate the multiple hypothesis testing problem that arises when researchers reuse natural experiments. We then provide guidance for future inference based on popular empirical settings including difference-in-differences, instrumental variables, and regression discontinuity designs. When we apply our guidance to two extensively studied natural experiments, business combination laws and the Regulation SHO pilot, we find that many results that were statistically significant using single hypothesis testing do not survive corrections for multiple hypothesis testing.

OVER THE LAST THREE DECADES, the "credibility revolution" has fundamentally altered empirical research in the field of economics, driven by a newfound emphasis on empirical research design. By exploiting conditions that

*Davidson Heath and Matthew C. Ringgenberg are at the David Eccles School of Business, University of Utah, Mehrdad Samadi is at the Federal Reserve Board of Governors, and Ingrid M. Werner is at The Ohio State University and CEPR. The authors thank the Editor (Stefan Nagel); several anonymous reviewers; Thorsten Beck; Andrew Chen; Yong Chen; David De Angelis; Joey Engelberg; Benjamin Gillen; Todd Gormley; Campbell Harvey; Jonathan Karpoff; Yan Liu; Ye Li; Florian Peters; Peter Reiss; Alessio Saretto; Rik Sen; Sophie Shive; Elvira Sojli; Holger Spamann; Noah Stoffman; Allan Timmermann; Michael Wittry; Michael Wolf; Yuchen Zhang; 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, the 2020 Annual Meeting of the Western Finance Association; and seminar participants at Rutgers University, SMU Cox, The Ohio State University, Texas A&M University, Tulane University, Securities and Exchange Commission, University of Utah, Virtual Finance Seminar (hosted by Michigan State and the University of Illinois at Chicago), Virtual Finance Seminar Series (hosted by the University of Bristol, University of Exeter, University of Lancaster, and University of Manchester), University of Virginia, Federal Reserve Board, University of California at Riverside, University of Maryland, and Norges Handelshøyskole - NHH. The views expressed in this paper are those of the authors and do not represent the views of the Federal Reserve Board, Federal Reserve System, or their staff. No author has received financial support for this research. Heath, Ringgenberg, and Samadi have nothing further to disclose. Werner is an independent director for Dimensional U.S. Mutual Funds and ETF Trust, is a director for the Fourth Swedish Pension Fund (AP4), and serves on the Prize Committee for Riksbanken's Prize in Economic Sciences in Memory of Alfred Nobel.

Correspondence: Ingrid M. Werner, Ohio State University, 2100 Neil Avenue, Columbus, OH 43210; e-mail: werner.47@osu.edu.

DOI: 10.1111/jofi.13250

resemble random assignment, researchers can better estimate the causal effect of one variable on another. Indeed, the use of such "natural experiments" has increased dramatically in recent years. Bowen, Frésard, and Taillard (2017) estimate that 39% of empirical corporate finance articles between 2010 and 2012 use natural experiments, compared to just 8% in the 1970s.[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. Credible 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 the effect of the treatment on other 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 (e.g., the Regulation SHO and U.S. Tick Size pilot programs).[2]

While researchers who reuse a setting may develop testable hypotheses independently of one another, if their research question can be viewed as part of the broader question "What was the treatment effect in this setting?," their tests can be viewed as part of the same "family." This leads to a multiple testing problem.[3] Tests are generally considered part of the same family when they support the same research question and use the same data.[4] In such cases, reusing a given setting without accounting for the other outcomes that have already been examined leads to p-values that cannot be interpreted in the usual manner.

We start by examining the multiple testing problem using simulations. While it is not possible to examine the rate of false positives using real data, simulations based on commonly used data in finance allow us to quantify the potential scope of the problem under conditions that resemble frequently used natural experiments. While the commonly used p-value cutoff of $\alpha = 0.05$ should indicate a 5% probability of observing results at least as extreme as the observed results if the null hypothesis of no effect is true (i.e., a Type I error), we show that the actual probability is often much higher when a natural experiment is reused.[5] In fact, for commonly reused settings and

[1] Bowen, Frésard, and Taillard (2017) classify methods based on the following categories: instrumental variables (IV), difference-in-differences regressions, selection models, regression discontinuity designs (RDD), and randomized experiments.

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

[3] Whether these tests are conducted by different researchers at different times should have no bearing on the multiple testing issue. If it did, then one researcher could address the problem simply by asking different people to push "enter" on their keyboard for them or by waiting a specified length of time before performing the next test (Thompson et al. (2020)).

[4] For more discussion on families of tests, see Thompson et al. (2020) and references therein.

[5] Assuming independence of tests and that all of the null hypotheses are true, the probability of making at least one Type I error is $1 - (1 - \alpha)^S$, where S is the number of hypotheses examined.

estimation techniques, our simulations suggest that when the number of variables examined is large relative to the number of true effects, more than 50% of statistically significant findings may be false positives.

We next examine the properties of several different multiple testing correction methods using simulations. We also use simulations to develop t-statistic cutoffs that researchers can use to improve inference when reusing a setting. Finally, we apply our recommendations to two extensively studied experiments, namely, the staggered enactment of state-specific business combination laws and the Regulation SHO pilot. These two settings have been used to examine several hundred dependent variables. While we find that some of the results in these literatures do survive after applying multiple testing corrections, we find that many more of them do not survive. Moreover, the fraction of results that are significant with single hypothesis testing but not with multiple testing adjustments is consistent with our simulation results.

We use simulation evidence to demonstrate the consequences of reusing a natural experiment without adjusting for multiple testing. As more researchers reuse a setting, our simulation evidence shows it is likely to lead to a large number of Type I errors. Indeed, we show that when the number of variables examined is large relative to the number of true effects, the reuse of natural experiments without correcting for multiple testing may lead to more false positives being discovered than true positives. For example, imagine researchers collectively examine 293 different variables from the Center for Research in Security Prices (CRSP) and Compustat using the same staggered state-level introduction of a law as a source of exogenous variation. Assuming that the law actually causes 10 true effects, our simulation evidence suggests the researchers will document approximately 25 false discoveries in addition to the 10 true discoveries (see Table I).[6]

In light of these findings, we examine the properties of several possible correction methods. To ensure that our findings apply to a wide range of research designs, we examine several different multiple testing correction methods in four popular empirical settings: randomized control trials (RCTs), staggered introductions, IV regressions, and RDD. For each of these settings, we simulate the exogenous independent variables and then sequentially examine the set of 293 outcome variables from Compustat and CRSP. Because multiple testing corrections 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.

For the tests under consideration, commonly studied outcome variables, and corresponding dependence structures, the results are generally similar across multiple testing adjustment methods. Specifically, we examine a number of properties for different correction methods, including the Type I error rate (number of false positives divided by number of null effects), the Type II

[6] While 293 variables may seem like a large number to examine, researchers have examined over 400 different variables using Regulation SHO. See Internet Appendix Table IA.II for details. The Internet Appendix may be found in the online version of this article.

Table I
Simulations

This table presents simulation evidence on the performance of different multiple testing corrections for four types of simulated settings: randomized control trials, the staggered introduction of state-level changes (staggered introductions), instrumental variables, and regression discontinuity designs. We examine 293 outcome variables obtained from Compustat and CRSP in random order. For each new outcome variable, we apply multiple testing corrections to the family of tests that includes that outcome and all previously tested outcomes. The simulated results are then averaged across 10 random orderings for each of 10 independent simulations. In certain simulations, we also incorporate 10 and 20 simulated true effects that are a linear function of the treatment. For each additional outcome added to the family of tests, we apply the Bonferroni (1936) FWER correction, the FDR correction of Benjamini and Yekutieli (2001; BHY), and the FWER correction of Romano and Wolf (2005, 2016; RW) in order to determine statistical significance for that outcome. Average frequencies of true positives and false positives for each empirical setting across simulations are presented in Panel A. Panel B presents the performance of each correction across several criteria. The Type I error rate is (# false positives / # null effects). The Type II error rate is (# false negatives / # true effects). Accuracy is the fraction of all tests with the correct result. Positive predictive value is the probability that a positive finding is a true effect. The 293 dependent variables examined are listed in Internet Appendix Table IA.I.

Panel A: Occurrence of False Positives and True Positives

		Multiple Testing Correction							
		No Correction		Bonferroni		BHY		RW	
Research Design	True Effects	False+	True+	False+	True+	False+	True+	False+	True+
Randomized Control Trials	0	14.7	0.0	0.2	0.0	0.1	0.0	0.2	0.0
	10	15.5	10.0	0.4	7.9	0.1	7.2	0.4	8.1
	20	14.1	20.0	0.1	16.4	0.1	17.1	0.2	17.1
Staggered Introductions	0	23.5	0.0	2.1	0.0	1.3	0.0	2.3	0.0
	10	22.4	10.0	1.8	6.6	1.6	6.3	2.0	6.9
	20	22.2	20.0	2.1	15.1	2.3	16.2	2.2	16.0
Regression Discontinuity Designs	0	15.6	0.0	0.3	0.0	0.1	0.0	0.4	0.0
	10	14.8	10.0	0.3	8.0	0.3	7.9	0.5	8.2
	20	14.8	20.0	0.3	14.5	0.3	14.6	0.5	14.9
Instrumental Variables	0	15.8	0.0	0.4	0.0	0.1	0.0	0.4	0.0
	10	13.8	10.0	0.1	8.3	0.0	7.9	0.1	8.7
	20	14.3	20.0	0.1	15.5	0.1	16.1	0.2	16.3

(*Continued*)

Table I—*Continued*

Panel B: Performance Criteria

Criterion	Research Design	Multiple Testing Correction			
		No Correction	Bonferroni	BHY	RW
Type I Error Rate	Randomized Control Trials	5.2%	0.1%	0.0%	0.1%
	Staggered Introductions	8.0%	0.7%	0.6%	0.8%
	Regression Discontinuity Designs	5.3%	0.1%	0.1%	0.2%
	Instrumental Variables	5.2%	0.1%	0.0%	0.1%
Type II Error Rate	Randomized Control Trials	0.0%	19.5%	21.2%	16.8%
	Staggered Introductions	0.0%	29.2%	28.0%	25.5%
	Regression Discontinuity Designs	0.0%	23.8%	24.0%	21.8%
	Instrumental Variables	0.0%	19.8%	20.2%	15.7%
Accuracy	Randomized Control Trials	95.0%	99.3%	99.3%	99.4%
	Staggered Introductions	92.2%	98.4%	98.6%	98.5%
	Regression Discontinuity Designs	94.8%	99.0%	99.1%	99.1%
	Instrumental Variables	95.0%	99.2%	99.3%	99.4%
Positive Predictive Value	Randomized Control Trials	48.9%	97.3%	99.0%	97.1%
	Staggered Introductions	38.9%	83.2%	83.7%	82.7%
	Regression Discontinuity Designs	48.2%	97.2%	97.2%	95.5%
	Instrumental Variables	50.2%	99.1%	99.7%	98.8%

error rate (number of false negatives divided by number of true effects), and Accuracy (fraction of all tests with the correct result). We find that the Romano and Wolf (2005, 2016) (RW) correction method, which controls the probability of making one or more false rejections across all hypotheses considered (the family-wise error rate [FWER]), performs well across these different dimensions. Other methods, such as the Benjamini and Yekutieli (2001) (BHY) method, which controls the expected value of the ratio of false rejections to total rejections across all hypotheses considered (the false discovery rate [FDR]), perform similarly.

Consequently, we use the RW method to calculate adjusted t-statistic critical values that can be used to make inferences when reusing a setting.[7] We construct adjusted critical values for four commonly used settings: RCTs, staggered introductions, IV regressions, and RDD. We find that the adjusted critical values evolve at a similar rate across these different empirical settings as more dependent variables are examined. To address the multiple testing problem, our results show that a good heuristic is that a new hypothesis should have a t-statistic of at least 2.5 if there are five prior findings and 3.0 if there are 20 prior findings using the same setting (see Table AI for results broken down by empirical setting and the number of prior findings).

Finally, to assess the potential importance of our findings, we apply our adjusted critical values to two commonly studied and distinct real-world settings: business combination laws and the Regulation SHO pilot. The business combination law setting is a natural experiment involving the staggered enactment of state-level laws, while the Regulation SHO pilot is an RCT conducted by the U.S. Securities and Exchange Commission (SEC). We reexamine the empirical evidence on the effects of treatment in these settings after adjusting for multiple testing.

When applying multiple hypothesis corrections to a growing family of tests, the way outcomes are sequenced may affect inference—if a researcher examines outcome C, the multiple testing correction may yield different results depending on whether outcomes A and B were both already examined, versus just outcome A or just outcome B. To account for this possibility, the multiple testing literature proposes different ways to sequence existing outcomes. We examine two different ways to sequence outcomes when applying multiple testing corrections.

Following Harvey, Liu, and Zhu (2016), the first approach orders outcomes by the date they were first reported. Specifically, when we apply multiple testing corrections to a given outcome, we consider the results that had been previously reported sequenced by the order in which the results were first made public. This effectively raises the bar for statistical significance over time, as more outcomes are examined. The second approach that we use is referred to as a "best foot forward policy" in the multiple testing literature (Foster and Stine (2008)). In this approach, outcomes are ordered from high to low based

[7] Alternatively, researchers can directly apply other methods such as BHY, which generate similar results.

on the likelihood that they will be rejected given the experiment. While the ordering of outcomes is ultimately subjective, this approach has been used in clinical trials where the outcomes are ordered based on experimental design (e.g., intended effects of treatment are ordered first). Consequently, the intended treatment effects have a lower statistical hurdle. As new potential treatment effects are proposed, we consider the causal arguments that link them to the intended treatment effect and add related outcomes to the family of tests at the appropriate level.

We find similar results using both sequencing approaches. Specifically, we find that while some results from the literature survive correction for multiple testing, many of them do not. For example, for business combination laws, while the existing literature finds that 73 out of the 114 outcomes examined are statistically significant using single hypothesis testing methods, only 28 (27) of these outcomes remain statistically significant after applying our RW cutoffs when we sequence by the date of reporting (by the second approach). Similar results obtain for outcomes examined using Regulation SHO. Overall, our findings highlight the potential importance of considering multiple testing when making inferences.

Our analyses focus largely on a statistical issue: p-values do not retain their usual interpretation when a large number of outcomes are examined without accounting for multiple testing. However, p-values (and t-statistics) are only one part of inference. We discuss a number of other best practices from the existing literature and how they relate to multiple testing to assist with inference when natural experiments are reused. For example, researchers can provide corroborating evidence, state and provide supporting evidence of causal channels, and reconcile their evidence with existing evidence derived from the same setting.

Overall, our results contribute to a growing literature on multiple testing in economics. Multiple testing corrections have been proposed in several other types of experimental settings where researchers develop prespecified tests in support of the same research question using the same data.[8] 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. In contrast, researchers reusing a given natural experiment develop a variety of different testable hypotheses independently of one another, but these tests effectively investigate the same research question: What was the effect of the treatment? Similarly, researchers in empirical asset pricing independently develop testable

[8] For example, Ludbrook (1998) examines multiple testing in biomedical research and states that "A family of hypotheses is all those actually tested on the results of a single experiment." Clinical trials generally involve multiple comparisons and tests for multiple end points. If multiplicity is taken into account, this situation can lead to the approval of ineffective treatments (a Type I error); see Bretz, Dmitrienko, and Tamhane (2010). Similarly, Thompson et al. (2020) examine multiple testing in psychology and suggest that multiple testing corrections were designed "for situations in which a researcher performs multiple statistical tests within the same experiment."

hypotheses and effectively use the same data to examine whether expected returns are predictable (Harvey and Liu (2013, 2014), Harvey, Liu, and Zhu (2016), Chordia, Goyal, and Saretto (2020), Hou, Xue, and Zhang (2020), Engelberg et al. (2023)). Multiple testing corrections have also been applied to papers estimating asset price variation (Liu, Patton, and Sheppard (2015)), and examining fund performance (Andrikogiannopoulou and Papakonstantinou (2019), Giglio, Liao, and Xiu (2021)). In contrast to these studies, our paper is the first to examine the reuse of natural experiments. The existing literature focuses largely on settings in which the dependent variable is the same while the independent variables vary. Our paper instead focuses on settings in which the independent variable is the same and the dependent variable varies across tests. It is therefore unclear whether the recommendations from the prior literature apply to the reuse of natural experiments. Our paper fills this gap.

The rest of the paper proceeds as follows. Section I provides an overview of multiple testing frameworks. Section II provides simulation evidence. Section III uses multiple testing corrections to reevaluate existing results using business combination laws and Regulation SHO. Section IV discusses other potential solutions to the multiple testing problem, notes caveats, and reviews other issues when reusing natural experiments. Section V concludes.

I. Multiple Testing Corrections

In this section, we provide an overview of several multiple testing corrections and describe our implementation of them. The different methods that we examine are designed to control different metrics: the FWER, the FDR, or the false discovery proportion (FDP). We briefly define these metrics and discuss each correction method below. For more detailed descriptions, we refer the reader to the papers cited below as well as Chordia, Goyal, and Saretto (2020).

A. Family-Wise Error Rate

The FWER is defined as the probability of making one or more false rejections given all hypotheses considered. These corrections allow us to maintain the standard p-value cutoff of $\alpha = 0.05$ for a family of tests. In other words, this implies there is still a 5% probability of observing results at least as extreme as the observed results if the null hypothesis of no effect is true, even when many hypotheses are tested. However, as the number of hypotheses under consideration becomes large, these methods become relatively conservative since they control the probability of even one false positive.

The first correction that we examine is the Bonferroni (1936) method. Under this correction, the critical p-value is equal to $\frac{\alpha}{S}$, where S is the number of outcomes under consideration. While the Bonferroni method is simple to apply, it treats all tests as independent. More powerful FWER procedures account for the dependence structure across hypotheses by resampling (White (2000)) and reject as many null hypotheses as possible by using a step-down approach

(Holm (1979)).[9] The second correction that we apply is the procedure developed in RW and described in further detail by Clarke, Romano, and Wolf (2020). The RW correction combines resampling with a step-down approach. As a consequence, this approach generally has more power than other FWER methods.

B. FDR and FDP

In some applications, researchers may examine tens of thousands of hypotheses and may be willing to tolerate more false positives than the standard $\alpha = 0.05$ allows for. The FDR and FDP were developed to address these situations.[10] Rather than control the probability of *any* false positives, the FDR controls the expected value of the ratio of false rejections to total rejections across tests in the same family. We apply the BHY correction, which builds on earlier work by Benjamini and Hochberg (1995) by controlling the FDR under more arbitrary dependence structures. While the BHY correction is known to be relatively conservative in controlling the FDR, BHY has been applied in several asset pricing settings including Harvey, Liu, and Zhu (2016) and Chordia, Goyal, and Saretto (2020). Following these papers, we control the FDR at the 5% level in all of our applications.

Finally, we examine FDP methods. These methods directly control the ratio of false rejections to rejections for a single application. Romano and Wolf (2007) extend the RW procedure described above for control of the FDP. In our setting, when we apply the Romano and Wolf (2007) FDP correction, we find results that are qualitatively the same as when we apply the RW method.[11] Accordingly, we do not report results for FDP corrections in our main tests.

Overall, both the RW and BHY methods generally have better properties than the Bonferroni (1936) method. While the BHY method generally has more power than the RW method as the number of tests becomes large, it controls a conceptually different error rate. Put differently, the RW and BHY methods have distinct advantages and disadvantages: the RW method has the advantage of leading to fewer false discoveries, but it may miss more true discoveries than the BHY method, especially as the number of tests becomes extremely large. In Section II, we use simulation analyses to examine whether one of these methods performs better than the others within the context of reusing natural experiments.

C. Bootstrap

When applying the RW method, we use bootstrapping to account for the dependence structure across hypotheses. Importantly, the results may depend on

[9] In the settings that we examine, many dependent variables are related due to common firm and/or economic forces, so accounting for their dependence is important.

[10] For example, genome association studies often examine the relation between a disease and tens of thousands of genes that may be related to the disease.

[11] We control the FDP at the 5% proportion and level as in Chordia, Goyal, and Saretto (2020).

the structure of the bootstrap—the bootstrap procedure should preserve the underlying dependence structure in the original data. In our setting, we build bootstrap samples of 1,000 replicants by randomly sampling firms with replacement from each sample. In other words, to preserve the time-series properties of the raw data, we draw all dates for each firm.[12] We then use the same bootstrap sample 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 same set of firms and dates).

II. Simulations

In this section, we use simulation evidence to examine the multiple testing problem associated with reusing natural experiments. We also explore the properties of various multiple testing corrections. We conduct simulations using three commonly used methodologies: difference-in-differences regressions, IV regressions, and RDD. Within the difference-in-differences setting, we examine two research designs: (i) an RCT in which firms are randomly selected for treatment at one point in time and (ii) staggered introductions of state-level changes, which exploit variation across firms and over time.

For each of these four settings, we simulate the independent variable (the source of exogenous variation) and then examine dependent variables based on real data. By using real data for the dependent variables, our simulations account for the actual dependence structures encountered by researchers reusing natural experiments. Our dependent variables are drawn from the set of variables in Compustat and CRSP, which yields 293 variables (discussed in greater detail in Section II.B). These variables are also listed in Internet Appendix Table IA.I.[13] We start by randomly drawing one of the 293 outcome variables. We then continue to randomly add one variable at a time to the family of tests.

A. Empirical Settings

A.1. Randomized Control Trial

To construct the simulated RCT sample, we randomly select a treatment year, then collect five years of annual Compustat firm-level data both before and after the treatment. We randomly assign treated status to one-third of the firms, while the other firms serve as controls in each simulation. We then estimate panel regressions of the form

$$y_{i,t} = \alpha_i + \alpha_t + \beta \cdot Treat_{i,t} + \epsilon_{i,t}, \quad (1)$$

[12] When we apply multiple testing corrections to staggered state-level introductions, we stratify the draws by state of incorporation. After drawing firms, we generate a new firm index to preserve the correct degrees of freedom when absorbing fixed effects.

[13] A database of existing papers that use commonly studied natural experiments can be found on the website https://www.reusingnaturalexperiments.com.

where $y_{i,t}$ is the outcome variable of interest for firm i in year t, and $Treat_{i,t}$ is an indicator variable equal to one if the firm is in the treated firm group and the treatment has taken effect and equal to zero otherwise. We include firm and year fixed effects and cluster standard errors at the firm level.

A.2. Staggered Introductions

The sample consists of Compustat firm-level data with fiscal years ending from 1976 through 1995. To simulate the staggered introductions, we randomly assign the enactment years of a regulatory or legal change, without replacement, to the states of incorporation in the sample, leaving Delaware untreated. We then estimate panel regressions of the form

$$y_{i,s,t} = \alpha_i + \alpha_t + \beta \cdot Treat_{s,t} + \delta' \boldsymbol{L_{s,t}} + \varepsilon_{i,s,t}, \tag{2}$$

where $y_{i,s,t}$ is the outcome variable of interest for firm i in year t incorporated in state $s$and $Treat_{s,t}$ is an indicator variable equal to one if the change has occurred in state s by year t and zero otherwise. Following Spamann (2019), $\boldsymbol{L_{s,t}}$ includes controls for the five antitakeover statutes from Karpoff and Wittry (2018).[14] We include firm and year fixed effects and cluster standard errors at the firm level.

A.3. IV 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 (such that we do not have a weak instrument) plus a noise term. We then estimate two-stage least-squares regressions of the form

$$X_{i,t} = \kappa_i + \kappa_t + \gamma \cdot Z_{i,t} + \eta_{i,t}, \tag{3}$$

$$y_{i,t} = \alpha_i + \alpha_t + \beta \cdot \widehat{X}_{i,t} + \epsilon_{i,t}, \tag{4}$$

where $y_{i,t}$ is the outcome variable of interest for firm i in year t, $X_{i,t}$ is the endogenous independent variable, $Z_{i,t}$ is an IV, and $\widehat{X}_{i,t}$ is the fitted value from the first-stage regression. We include firm and year fixed effects and cluster standard errors at the firm level.

A.4. Regression Discontinuity Design

To construct the simulated RDD sample, we use the 1984 to 2004 period. Each year we randomly simulate a forcing variable and a threshold, and we

[14] This setting mimics the business combination law literature and leads to an unbalanced panel as the majority of firms are incorporated in Delaware. See Bertrand, Duflo, and Mullainathan (2004) and Spamann (2019) for discussions about the resulting issues.

construct a treatment variable ($Treat_{i,t}$) that takes the value of one above the threshold. We then estimate panel regressions of the form

$$y_{i,t} = \alpha_i + \alpha_t + \beta \cdot Treat_{i,t} + \lambda \cdot X_{i,t} \cdot Treat_{i,t} + \epsilon_{i,t}, \quad (5)$$

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, use a bandwidth of 500 firms on either side of each yearly simulated threshold, and cluster standard errors at the firm level.[15]

B. Compustat and CRSP Outcomes

Our dependent variables are drawn from Compustat and CRSP, and include commonly used transformations of each variable. To arrive at a set of Compustat variables, we collect raw variables from financial statements that are nonmissing for at least 70% of observations in a sample between January 1970 and June 2019. For Compustat outcomes, we use the raw variable, raw variable scaled by total assets, and the percentage change in the raw variable scaled by total assets. This approaches results in 96 raw Compustat variables, generating 288 Compustat outcomes in total. We also use monthly CRSP stock data to calculate firm-year average trading volume, average share turnover, cumulative returns, average dollar bid-ask spread, and average percentage bid-ask spread using firms' fiscal years. The resulting sample contains 293 different dependent variables (see Internet Appendix Table IA.I for details).

C. Simulated True Treatment Effects

By construction, the realizations of the treatment indicators are simulated to be independent of the outcomes, so there should be no relation between the independent and dependent variables. To study how different multiple testing corrections perform at detecting true effects, we choose sets of 10 and 20 outcome variables at random, without replacement, and add a linear function of the treatment so that they are related to the independent variables (i.e., we create true effects). These outcomes are constructed to produce a t-statistic from a uniform distribution between 2.8 and 5. The lower cutoff of $t = 2.8$ is chosen to ensure that our simulated natural experiments are adequately powered, that is, a single hypothesis test would reliably detect the effect at $p < 0.05$ at least 80% of the time (Bloom (1995)).

D. Comparing Multiple Testing Frameworks

Table I summarizes the simulation results. As discussed above, we examine the 293 outcome variables in random order. For each new outcome variable,

[15] Results are similar for other choices of the bandwidth and control function.

we apply multiple testing corrections to the family of tests that includes that outcome and all previously tested outcomes. The simulated results are then averaged across 10 random orderings for each of 10 independent simulations, that is, 100 total simulated processes. We repeat this for each research design and each possible number of true effects (which take the value of 0, 10, or 20).

Table I, Panel A, presents the average performance in terms of false and true discoveries (i.e., false positives and true positives). Before applying multiple testing corrections, in each of the four settings with zero true effects there are at least 15 false positive findings on average, with a p-value < 0.05. This is the multiple testing problem. When natural experiments are reused, p-values no longer have their usual interpretation, which may lead to many false positives that are erroneously documented as true positives. Moreover, the occurrence of false positives is higher in the staggered introductions setting, with over 20 false positives on average. This observation is consistent with Bertrand, Duflo, and Mullainathan (2004) and Spamann (2019) who argue that these designs are prone to overstate statistical significance.

When we introduce 10 or 20 true effects, they are successfully identified as expected, but the number of false positives remains at roughly the same level as when we examined zero true effects. Thus, in a real-world setting, a researcher would infer that 25 or 35 outcomes are significant based on a p-value of 0.05, without having any idea of which ones are false positives and which ones are true positives. Put differently, depending on how many true effects there are in a particular simulation (10 or 20), the FDRs for the RCT, IV, or RDD methodologies are similar and range from approximately 43% (15 of 35, if there are 20 true effects) to 60% (15 of 25, if there are 10 true effects), with the results even worse for staggered introductions, which range from approximately 53% (if there are 20 true effects) to 70% (if there are 10 true effects). Overall, the findings suggest that for commonly reused settings and estimation techniques, more than 50% of documented findings may be false positives.[16]

We next examine the performance of several different multiple testing corrections. Specifically, we examine the Bonferroni and RW corrections, which control the FWER, as well as the BHY correction, which controls the FDR. The results are generally similar regardless of research design. As expected, the Bonferroni correction successfully controls the FWER, but it does not detect as many true effects. The RW correction successfully controls the FWER in a manner similar to the Bonferroni correction, but is able to detect more true effects. Finally, the BHY correction generally performs similar to the RW correction. Overall, the results show that the RW and BHY corrections both perform as expected, with the differences between the two minimal and depending on the research design and number of true effects.

To further explore the performance of these methods, Panel B presents four standard criteria used to evaluate methods of statistical inference. The first is

[16] Of course, this depends on the number of true effects, which is unknown. Nonetheless, regardless of the number of true effects, the results indicate that many results are likely to be false positives.

the Type I error (i.e., false positive) rate. We see that the Type I error rate is 5% as expected for the uncorrected hypothesis tests; the exception is the staggered introductions with a rate of 8%. In contrast, all three multiple testing corrections result in a Type I error rate that is close to zero. The second criterion is the Type II error (false negative) rate, which measures statistical power and is equal to one minus the Type I error rate.[17] Across correction methods and research designs, the numbers are generally close to the usual value of 20%. We also compute each method's accuracy, defined as the fraction of all tests with the correct result, and positive predictive value, defined as the fraction of positive results that are a true treatment effect. For all three corrections, the accuracy is at least 99% and the positive predicted value is generally above 95%. The exception again is the staggered introductions, for which the positive predictive value ranges from 83% to 84% depending on the correction method.

Overall, the simulations illustrate the inference problem that arises when researchers reuse natural experiments without accounting for outcomes that have already been examined. They also suggest that based on the number of tests under consideration, commonly used outcome variables, and their corresponding dependence structures, the RW and BHY corrections perform similarly and in a manner that helps make the correct inference when a setting is reused.

D.1. Adjusted t*-Statistic Critical Values*

To provide guidance for future researchers, we calculate adjusted t-statistic critical values that can be used to make inferences when reusing a setting. In our main analyzes, we use the RW correction, but the adjusted t-statistic cutoffs evolve in a similar manner for the BHY correction. For each setting (RCTs, staggered introductions, IVs, and RDDs), the cutoffs indicate the minimum t-statistic necessary for statistical significance that corresponds to the usual interpretation. As the number of outcomes examined increases, the cutoff increases.

Figure 1 presents the adjusted critical values. The evolution of the RW adjusted critical values is similar across all four settings. For further reference, the cutoffs are presented in tabular format in Appendix Table AI.[18] The adjusted critical values for the fifth test in a family are 2.54 in the RCT setting, 2.53 in the staggered introductions setting, 2.53 in the IV setting, and 2.52 in the RDD setting.[19] The critical values for the 10$^{\text{th}}$ outcome increase to 2.76, 2.77, 2.75, and 2.75 for RCT, staggered introductions, IV, and RDD,

[17] Because by construction our simulated natural experiments are adequately powered to detect the true effects, when no correction is applied, the Type II error rate is 0% and power is 100%.

[18] Our critical values assume two-sided tests. See Romano and Wolf (2018) for a discussion of multiple testing with one-sided hypotheses.

[19] These numbers are for the fifth test only (e.g., for the RCT setting, the first test has a critical value of 1.96, and the critical values for the second, third, and fourth tests are 2.18, 2.34, and 2.45, respectively).

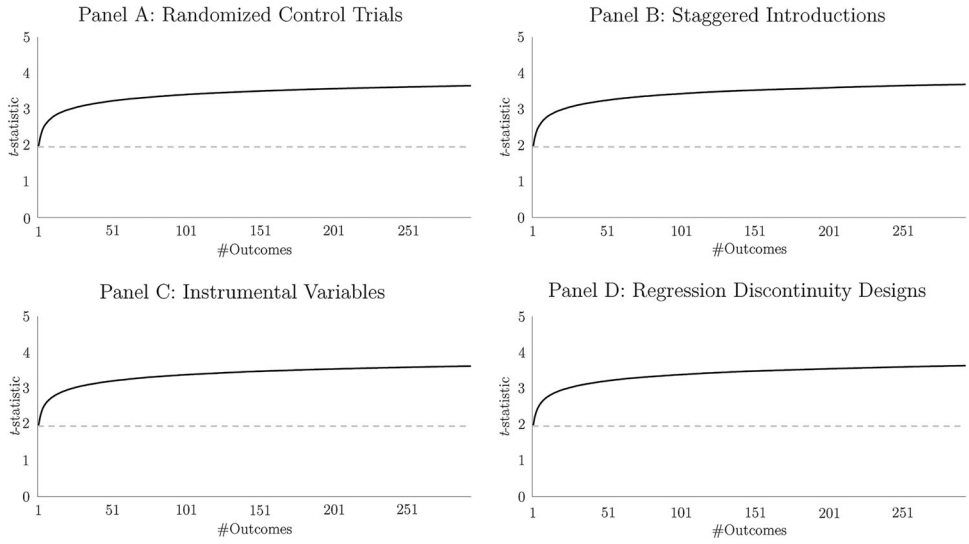

Figure 1. Multiple testing adjusted critical values by research design. This figure presents adjusted t-statistic critical values for four settings: randomized control trials (Panel A), the staggered introduction of state-level changes (Panel B), instrumental variables (Panel C), and regression discontinuity designs (Panel D). We examine 293 outcome variables obtained from Compustat and CRSP in random order. For each new outcome variable, we apply multiple testing corrections to the family of tests that includes that outcome and all previously tested outcomes. The simulated results are then averaged across 10 random orderings for each of 10 independent simulations. For each additional outcome added to the family of tests, we compute adjusted critical values using the FWER correction of Romano and Wolf (2005, 2016; RW). The outcomes variables are listed in Internet Appendix Table IA.I and the adjusted cutoffs are presented in tabular format in Appendix Table AI.

respectively. Finally, the critical values for the 20th outcome are 2.98, 2.99, 2.96, and 2.96 for RCT, staggered introductions, IV, and RDD, respectively.

Ideally, researchers who are using the RW correction should replicate all existing studies that use a setting in order to best reflect the dependence between outcomes. However, if this is not possible, the results in Figure 1 and Appendix AI provide a heuristic for inference when reusing a setting. Our results show that a good heuristic is that a new hypothesis should have a t-statistic of at least 2.5 if there are five prior findings and 3.0 if there are 20 prior findings using the same setting.

III. Evaluating Existing Evidence

To assess the practical importance of our findings, we apply our adjusted critical values to two commonly studied real-world settings. Specifically, we examine the prior empirical evidence on (i) the causal effects of the staggered introduction of state-specific business combination laws and (ii) Regulation SHO. To apply the adjusted cutoffs from our simulations, we collect t-statistics

associated with 114 and 434 unique outcome variables that as of March 31, 2021, have been examined using business combination laws and Regulation SHO, respectively, as a source of exogenous variation.[20]

We start by compiling a list of all papers that use business combination laws or Regulation SHO as a source of exogenous variation. We consider unique outcomes that were examined in these papers using methodologies that could be represented as difference-in-differences regressions of the form

$$y_{i,t} = \alpha + \beta_1 \cdot Treatment_{i,t} + \beta_2 \cdot Post + \beta_3 \cdot Treatment_{i,t} \times Post + \varepsilon_{i,t}, \quad (6)$$

where $y_{i,t}$ is an outcome variable for firm i in year t, $Treatment_{i,t}$ is either the staggered introduction of state-specific business combination laws or a dummy indicating the Regulation SHO pilot stocks, and *Post* is a dummy indicating the period after the beginning of the treatment. We include papers that use various combinations of fixed effects and/or control variables, as long as they examine the β_3 coefficient in a model of the form shown above. We then collect t-statistics from all papers and all models that satisfy the following conditions. If multiple models examine the same dependent variable, we keep the β_3 coefficient from the model with the most controls and/or fixed effects.[21] We exclude models that include additional terms interacted with the main treatment effect and models used in subsample analysis.[22]

Since the BHY and RW methods perform similarly in our simulations, we use both to reevaluate the reported t-statistics from the existing literature. Because the BHY correction does not require bootstrapping, we apply it directly to the reported t-statistics to control the FDR. In contrast, because the RW approach requires bootstrapping, we instead use our reported t-statistic cutoffs (shown in Appendix Table AI) to adjust the reported t-statistics from the literature. Both methods generate similar inferences.

A. Business Combination Laws

U.S. states have enacted business combination laws at different points in time. The enactment of these laws has been used as a source of exogenous variation in the threat of a corporate takeover. Business combination laws

[20] As discussed above, ideally the multiple testing problem should be addressed by applying the RW or BHY method directly to all outcomes examined in the literature. However, for RW this would require replicating all existing results, which is often impractical or, in the case of proprietary data, infeasible. Using adjusted cutoffs together with reported t-statistics from the literature avoids these issues, and this approach has been used in other papers including Harvey, Liu, and Zhu (2016).

[21] Our results are similar if we choose the model with the least controls and/or fixed effects. The fact that many papers examine multiple specifications (with different controls, fixed effects, and/or sample periods) that test the same hypothesis itself generates a multiple testing problem. The issues that result from such specification searches within a paper are beyond the scope of our paper.

[22] For Regulation SHO, we keep treatment coefficients (i.e., β_3) for specifications that examine the start of the Regulation SHO pilot.

impose a moratorium on specified transactions between a target firm and an acquirer firm unless the board of directors votes otherwise before the acquirer become an interested shareholder. Karpoff and Malatesta (1989) and Comment and Schwert (1995) document negative announcement returns and higher takeover premiums for a subset of business combination laws. These state-level changes and those that followed have subsequently been used to examine a wide variety of outcome variables including wages, corporate investment, corporate innovation, board size, dividends, secondary market liquidity, and workplace safety.

Karpoff and Wittry (2018) show that the institutional, political economy, and historical context surrounding the enactment of these laws suggests that they were not exogenous for many firms, which makes results in this setting more difficult to interpret. To mitigate concerns about omitted variable bias, Karpoff and Wittry (2018) introduce a state-of-the-art specification that more accurately captures institutional and legal context. While we stress that the Karpoff and Wittry (2018) specification should be used by researchers examining business combination laws, we apply multiple testing adjustments to outcomes from the existing literature regardless of the specification used by the original authors. Results are also qualitatively similar when we apply adjusted RW cutoffs estimated using the sample and the full set of institutional controls of Karpoff and Wittry (2018) in Internet Appendix Figure IA1.

B. Regulation SHO

We also evaluate the evidence on the causal effects of the Regulation SHO pilot, which was a regulatory experiment enacted by the SEC. The pilot program assigned firms into treated and control groups and suspended Rule 10a-1, "the uptick rule," for firms in the treatment group. The pilot was specifically conducted by the SEC to examine the uptick rule, which restricted short sales so they could only execute when a firm's stock price was above the last traded price (i.e., an uptick). The experiment temporarily suspended Rule 10a-1 as well as any short-sale price test for a stratified sample of 1,000 stocks in the Russell 3000 index. To construct the 1,000 treatment firms, 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 Regulation SHO was set up as an RCT, the study is now effectively being used as a natural experiment: more than 80 papers have reused the setting to examine hypotheses that were not part of the original experiment design. The setting has been reused to examine a wide variety of outcome variables including corporate investment, innovation, payout policies, workplace safety, analysts rounding of forecasts, and banks' loss recognition.

C. Sequencing Tests

As previously discussed, when applying multiple hypothesis corrections, the sequence of outcomes examined may affect inference.[23] The flexibility associated with sequencing the tests conducted by separate research teams and over time raises the question of how such choices should be made in practice (Foster and Stine (2008)). We examine two approaches, discussed below.

We first examine a sequential ordering approach that is based on the date each result was first reported in the public domain. Ideally, we would sequence outcomes based on when each test was undertaken, but since this is not knowable, we use the first reported date as a proxy for when the test was undertaken. Similar to the approach in Harvey, Liu, and Zhu (2016), we manually search SSRN, Google Scholar, and academic journals for the date that each result was first made publicly available. These dates are reported in Internet Appendix Table IA.II. If multiple outcomes share the same reported date, we add them to the family of tests simultaneously. An advantage of this approach is that it is relatively simple and objective. However, it does not consider experimental design.

We next propose an alternative approach referred to as a "best foot forward policy" in the multiple testing literature (Foster and Stine (2008)), which typically focuses on experimental design. In this approach, outcomes are ordered from high to low based on the likelihood that they will be rejected given the experiment. While the ordering of outcomes is ultimately subjective, we base our illustration of the best foot forward policy on experimental design and the causal channels proposed in the business combination laws and Regulation SHO literature. Consequently, the intended treatment effects have a lower statistical hurdle. As new potential treatment effects are proposed, we consider the causal arguments that link them to the intended treatment effect and add related outcomes to the family of tests at the appropriate level. The benefit of this approach compared to hierarchical methods (Dmitrienko and Tamhane (2007), Yekutieli (2008)) is that indirect treatment effects can be examined and properly evaluated.[24] In other words, it recognizes that stakeholders further removed for a given treatment can be affected, and that endogenous adjustments can even contribute to the absence of an observed intended treatment effects.

Figure 2 displays an example of the best foot forward policy for business combination laws and Regulation SHO. In the case of business combination laws, the enactment of state-level laws was designed to restrict hostile takeovers. As a result, researchers have suggested that this could affect takeover-related outcomes, such as takeover premia or realized takeover activity. Accordingly, we define takeover-related outcomes as first-order effects (i.e., these effects are

[23] Sequential multiple testing corrections do not retroactively change the inference of previous tests. For more on sequential multiple testing, see Thompson et al. (2020).

[24] In the hierarchical approach, a rejection of the null at the prior level is required to proceed to a subsequent level of the hierarchy so that the cutoff following the first failure to reject is effectively infinite.

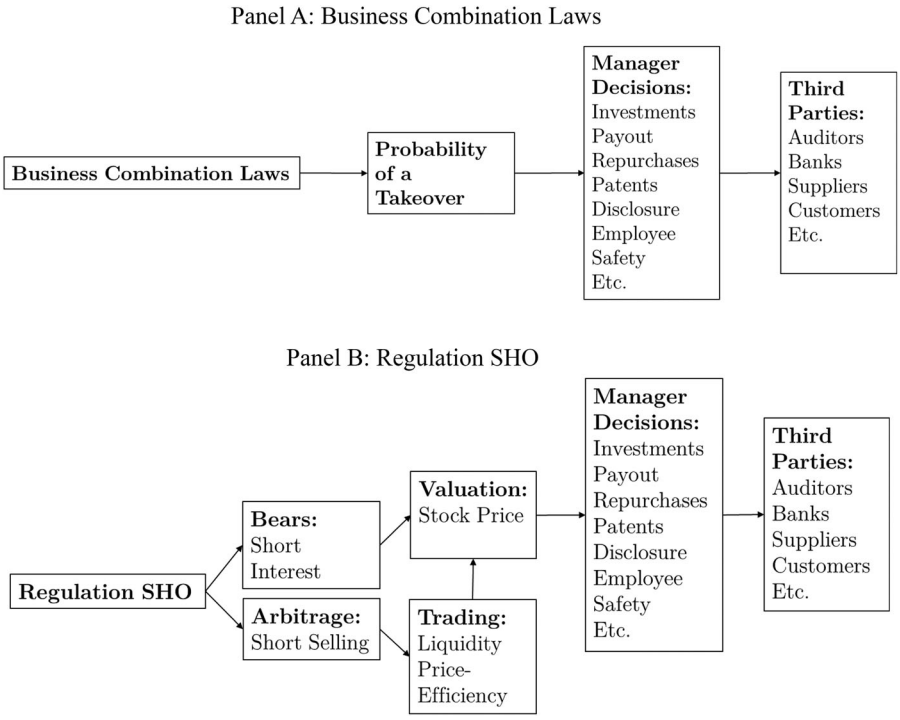

Figure 2. Best foot forward policy. This figure illustrates our implementation of the best foot forward policy for the two natural experiments we evaluate. Panel A displays the best foot forward policy for business combination laws and Panel B displays the best foot forward policy for Regulation SHO.

sequenced first). The causal channels proposed in the literature have further suggested that a change in the threat of takeover could result in a change in corporate governance, which could affect firm-level outcomes in turn. Consequently, we define firm-level outcomes as second-order effects. Finally, external parties may respond to potential changes at the firm. We group outcome variables related to external parties as third-order effects.

The Regulation SHO pilot was designed to loosen restrictions on short selling and we therefore sequence measures of short-selling activity first (i.e., short volume, short interest, etc.). In turn, the causal channels proposed in the literature further suggest that changes in short-selling activity could have implications for liquidity and price formation, and we sequence these variables next, as second-order effects. The literature on the real effects of financial markets suggests that the increased threat of short selling, and the impact it could have on prices, could affect firm-level decisions; we sequence these as third-order effects. Finally, external parties may respond to potential changes at the firm. Consequently, we group outcome variables related to external parties as fourth-order effects.

If multiple outcome variables are examined within the same order (e.g., the literature examined multiple second-order effects), we sequence these variables by the date they were first publicly reported. If variables within the same order are publicly reported on the same date, we add them to the family of tests simultaneously. The best foot forward approach requires greater coordination among researchers as exploration should ideally follow agreed-upon order, and is necessarily more subjective than sequencing the tests by the date they are first reported.

D. Results

Results for the 114 outcomes that have been tested using business combination laws are shown in Figure 3. Panel A plots the reported t-statistic for each outcome as a gray dot, sorted by the date it was first publicly available.[25] The solid gray curve plots the t-statistic cutoffs based on the BHY correction, while the solid black curve plots the cutoffs based on the RW correction. The lines resemble a step function rather than the smooth lines in Figure 1. The reason is that multiple outcomes may be examined in a single paper, in which case they are evaluated against the same threshold.[26] The gray dashed line is the t-statistic cutoff for a single test. While more than half of the outcomes were reported to be significant in the original papers (73 out of 114), Table II, Panel A, shows that fewer than half of these survive the correction for multiple hypothesis testing. Moreover, the BHY correction appears to be slightly more permissive than RW. Based on the RW correction, a t-statistic cutoff of 3.47 should be applied after 114 outcomes have been examined (see Appendix Table AI, Panel B). We zoom in on the first 20 outcomes in Panel C to show that a few wage-related outcomes (*Log Deflated Wage*, *Log of Total CEO Compensation*, and *Log of White Collar Worker Wages*), plant openings and closings, as well as several patent-citation outcomes are at or above the RW t-statistic cutoff (fewer plot above the BHY cutoff).

Panel B of Figure 3 has the same general layout, but here we order the outcomes based on the best foot forward approach. If there are multiple outcomes within a given grouping, we sort the outcomes by the date they were first publicly reported. The figure shows that none of outcomes that we define as takeover related (first-order effects) plot above the t-statistic cutoffs for the BHY or RW corrections, respectively. This is easier to see in Panel D, where we zoom in on the first 20 outcomes. In other words, this approach shows that none of the first-order effects is significant based on the single-test cutoff, and thus they certainly do not survive corrections for multiple hypothesis testing. This result echoes findings by Cain, McKeon, and Solomon (2017) and Karpoff and Wittry (2018).

[25] The vertical axis in the figure is truncated at a t-statistic cutoff of 5.0, and dots representing outcomes with higher t-statistics in the original paper(s) are plotted at 5.0.

[26] Note: the BHY correction may produce cutoffs that increase nonmonotonically in the number of examined outcomes, particularly for low numbers of evaluated outcomes. Similar patterns are found in Harvey, Liu, and Zhu (2016).

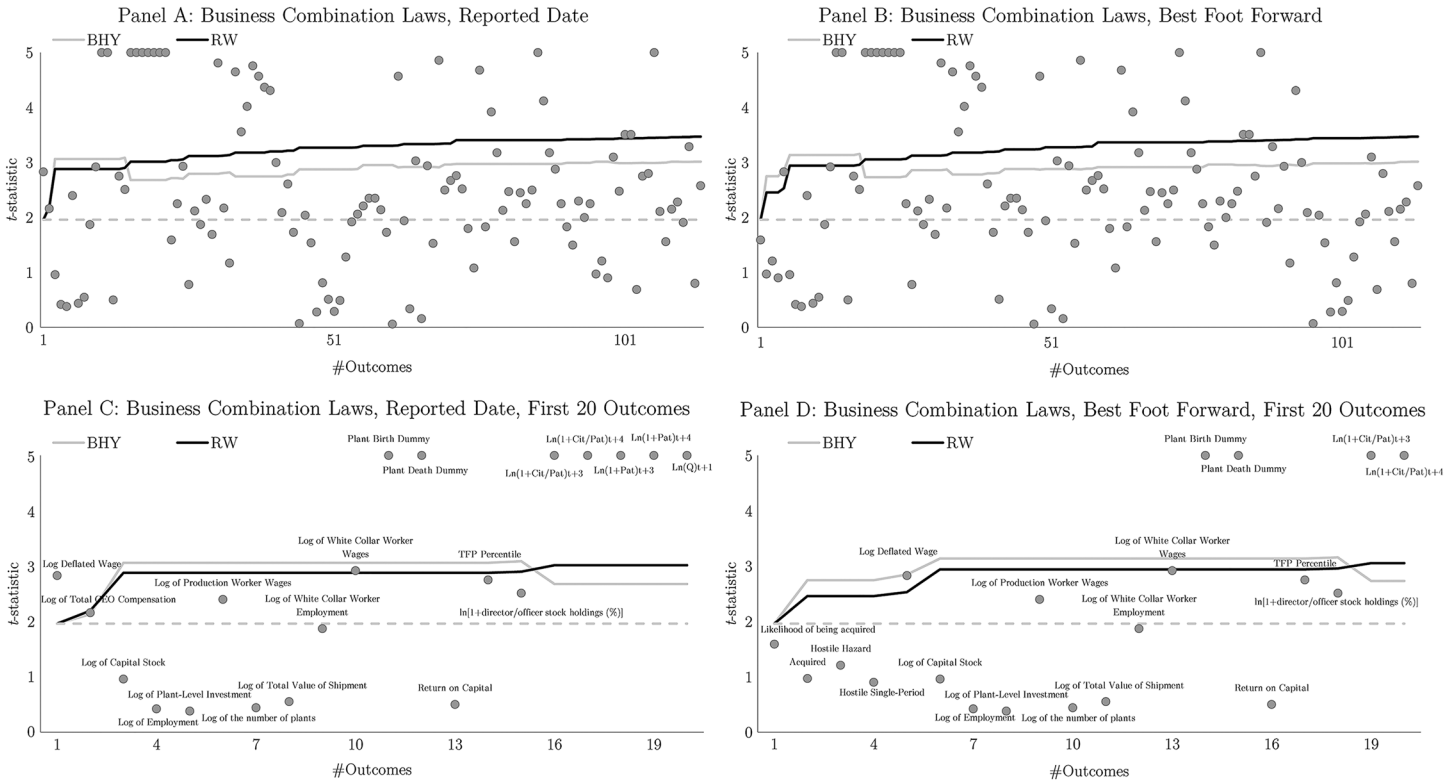

Figure 3. Evaluating existing evidence, business combination laws. This figure presents results from applying multiple testing corrections to reported results examining the treatment effects of business combination laws. Reported t-statistics are obtained for 114 unique outcomes examined using business combination laws (gray dots). The outcomes are listed in Internet Appendix Table IA.II. We directly apply the FDR (gray curve) correction of Benjamini and Yekutieli (2001; BHY) to the reported t-statistics. We separately apply the Romano and Wolf (2005, 2016; RW) adjusted critical values obtained from simulations (black curve) and presented in Figure 1 (Panel B) to the reported t-statistics. Panel A presents results when outcomes are sequenced by the date they were first reported. Panel B presents results when outcomes are sequenced using the best foot forward policy. Panel C presents results when outcomes are sequenced by the date they were first reported for the first 20 outcomes. Panel D presents results when outcomes are sequenced using the best foot forward policy for the first 20 outcomes.

Table II

Evaluating Existing Evidence

This table presents results from applying multiple testing corrections to reported results examining the treatment effects of business combination laws and Regulation SHO. Reported t-statistics are obtained for 114 and 434 unique outcomes examined using business combination laws and Regulation SHO, respectively. The outcomes are listed in Internet Appendix Table IA.II. We directly apply the Bonferroni (1936) FWER correction and the FDR correction of Benjamini and Yekutieli (2001; BHY) to the reported t-statistics. We separately apply the Romano and Wolf (2005, 2016; RW) adjusted critical values obtained from simulations and presented in Figure 1 and Appendix Table AI to the reported t-statistics. Panel A presents results for the reported treatment effects of business combination laws, where outcomes are sequenced by the date they were first reported. Panel B presents results for the reported treatment effects of business combination laws, where outcomes are sequenced using the best foot forward policy. Panel C presents results for the reported treatment effects of Regulation SHO, where outcomes are sequenced by the date they were first reported. Panel B presents results for the reported treatment effects of Regulation SHO, where outcomes are sequenced using the best foot forward policy.

Panel A: Business Combination Laws, Reported Date

Multiple Testing Correction	#Statistically Significant	%Outcomes
No correction	73	64.04%
BHY	36	31.58%
RW	28	24.56%

Panel B: Business Combination Laws, Best Foot Forward

Multiple Testing Correction	#Statistically Significant	%Outcomes
No correction	73	64.04%
BHY	33	28.95%
RW	27	23.68%

Panel C: Regulation SHO, Reported Date

Multiple Testing Correction	#Statistically Significant	%Outcomes
No correction	219	50.46%
BHY	26	5.99%
RW	21	4.84%

Panel D: Regulation SHO, Best Foot Forward

Multiple Testing Correction	#Statistically Significant	%Outcomes
No correction	219	50.46%
BHY	31	7.14%
RW	33	7.60%

We proceed similarly for the 434 outcomes that have been examined using Regulation SHO. The results are shown in Figure 4. Panel A plots the reported t-statistics for each outcome sorted by the date it was first publicly available. There are several unique features of this figure. First, the dots for early papers are plotted as either 2.58 or 1.96. The reason is that early papers in this literature only report asterisks indicating significance at the 1% and 5% levels. Accordingly, we assume that the t-statistics for these variables correspond to the single hypothesis cutoffs for significance at the 1% and 5% levels.[27] Second, the t-statistic cutoff for both BHY and RW start at 3.07 (see Appendix Table AI, Panel A) because the first paper (Alexander and Peterson (2008)) tested 28 unique outcomes. Third, the Regulation SHO literature examines more outcomes than we cover in our simulations, so we provide Bonferroni cutoffs (indicated by a dashed black curve) after more than 293 outcomes have been examined.

The Regulation SHO results mirror the results from business combination laws. While more than half of outcomes are statistically significant in the original papers (219 out of 434), Table II, Panel C, shows that only approximately 10% survive the correction for multiple hypothesis testing (the BHY correction is slightly more permissive than RW). After 434 outcomes have been examined, the t-statistic cutoff that should be applied based on the Bonferroni correction is 3.86. For transparency, we again zoom in on the first 20 outcomes in Panel C and none of the outcomes plot above either the BHY or the RW cutoff.

Finally, in Panel B we sequence outcomes from the Regulation SHO literature based on the best foot forward approach. The clustering of dots at 2.58 makes it difficult to discern if the first-order effects plot above the t-statistic cutoffs, so we again zoom in on the first 20 outcomes in Panel D. The first four outcomes examined in Panel D relate to short-selling activity (*No. Trades*, *Short/Long*, *Trade Size (Short Sale)*, and *Volume*), and these all plot above the RW t-statistic cutoff. However, none of the outcomes measuring short positions (*Short Interest (NYSE)*, *Abnormal Short Interest*, and *SHORT RATIO*) plot above the RW (or the BHY) t-statistic cutoff. Hence, while arbitrage activity appears to have increased, we find no evidence that traders increased their short positions as a result of Regulation SHO. This finding agrees with the evidence in Diether, Lee, and Werner (2009).

IV. Discussion

In this section, we note caveats to our findings, discuss other potential solutions to the multiple testing problem, and provide an overview of best practices when reusing natural experiments.

[27] If these papers report an insignificant outcome, we assume that it has a t-statistic of one.

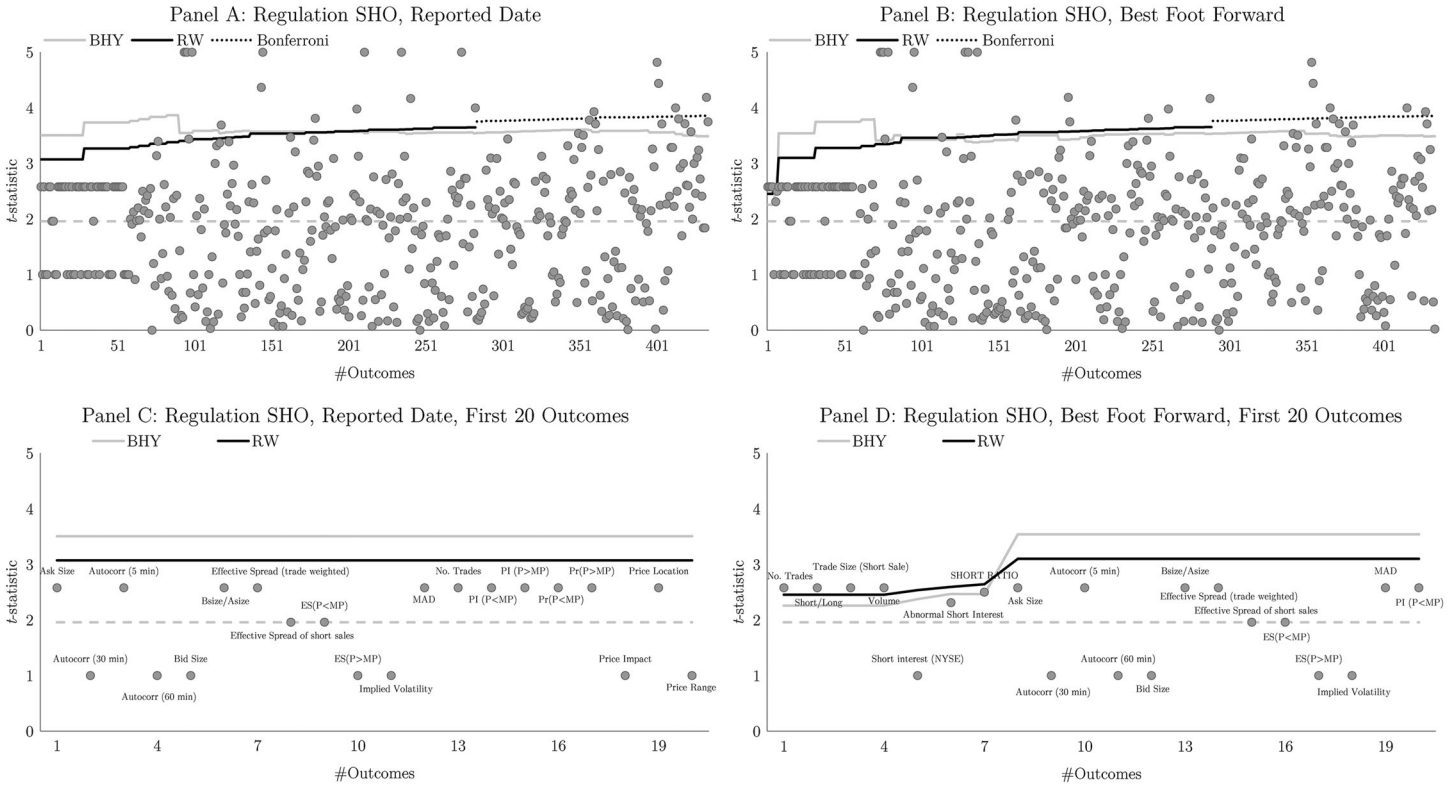

Figure 4. Evaluating existing evidence, Regulation SHO. This figure presents results from applying multiple testing corrections to reported results examining the treatment effects of Regulation SHO. Reported t-statistics are obtained for 434 unique outcomes examined using Regulation SHO (gray dots). The outcomes are listed in Internet Appendix Table IA.II. We directly apply the FDR (gray curve) correction of Benjamini and Yekutieli (2001; BHY) to the reported t-statistics. We separately apply the Romano and Wolf (2005, 2016; RW) adjusted critical values obtained from simulations (black curve) and presented in Figure 1 (Panel A) to the reported t-statistics. When the number of outcomes exceeds 293, we replace the RW cutoffs with the FWER correction of Bonferroni (1936) (black dashed curve). Panel A presents results when outcomes are sequenced by the date they were first reported. Panel B presents results when outcomes are sequenced using the best foot forward policy. Panel C presents results when outcomes are sequenced by the date they were first reported for the first 20 outcomes. Panel D presents results when outcomes are sequenced using the best foot forward policy for the first 20 outcomes.

A. Caveats

A.1. p-Values are Only One Input for Inference

Fisher (1925) presents the *p*-value as only one of many inputs that should be used in evaluating research and making decisions. Similarly, we caution that multiple testing correction methods are not a panacea; simply clearing the hurdle of adjusted critical values does not mean that a research design is valid. Moreover, researchers should take care in interpreting *p*-values: a *p*-value of 0.09 should not be viewed as proving that a result exists, nor should a *p*-value of 0.11 be viewed as proving there is no result. Rather, *p*-values are just one of many inputs that assist with inference, along with information about the proposed economic mechanism and the validity of the research design.

A.2. What is the Right Burden of Proof?

Motivated by our simulation evidence, we suggest that researchers use the adjusted critical values in Appendix Table AI when reusing a setting. However, some may wonder whether the medicine is worse than the disease: in other words, do the additional complications associated with our recommendations, as well as the possible increase in Type II error rates, result in improved inference?

While the trade-off between Type I and Type II error rates is ultimately a philosophical question that is beyond the scope of this paper, we do note that the results in Panel B of Table I show strong evidence that the RW and BHY methods that we recommend have better accuracy (and better positive predicted values) than uncorrected results. In other words, the simulation results suggest that our recommendations will lead to test statistics that more frequently yield the correct result.

Some may argue that the multiple testing problem should not apply since researchers develop a variety of different testable hypotheses when reusing a natural experiment and are not interested in the experiment per se. However, policymakers evaluate the evidence surrounding a given natural experiment when making policy that affects many stakeholders. For example, the SEC (2007) considered papers that examine the treatment effects of removing the uptick rule during the Regulation SHO pilot, and in large part decided to repeal the uptick rule because of this evidence. As Leamer (1983) put it, economics research has "customers in government, business, and on the boardwalk at Atlantic City." Policy implementations that are based on false positive results could potentially be very costly for society.

B. Improving Inference when Reusing a Natural Experiment

Our analyses focus largely on a statistical issue: *p*-values do not have their usual interpretation when a large number of outcomes is examined in a family of tests. In this section we provide other guidelines for improving inference when reusing a setting.

B.1. Corroborating Evidence

It is important to note that our results do not consider the various types of additional tests that researchers can provide when conducting inference. One way researchers address the multiple hypothesis testing problem is by gathering new data, that is, if researchers can find a new experiment that can be used to study the same question, then the resulting new test will not be in the same family of tests as the existing literature.[28] Researchers also base their inference on multiple outcomes in a given setting. Put differently, if existing theory predicts effects on more than one variable, then testing more than one variable may create additional information to improve inference. Basing inference on multiple outcomes can be accommodated by bootstrap-based multiple testing correction procedures, as they maintain control of false positives while not unnecessarily penalizing correlated outcomes. Researchers also develop and test additional hypotheses of heterogeneous treatment effects. Yekutieli (2008) and Dmitrienko and Tamhane (2007) discuss hierarchical corrections for heterogeneous treatment effects.

B.2. State and Test Causal Channels

Experimental research in the social sciences is often complicated by the fact that humans change their behavior in complex ways. For example, even if the Regulation SHO pilot did not change the cost of short selling, it may have changed firm outcomes if firm managers believed that the experiment would change short selling in their stock. Consequently, the Regulation SHO pilot could result in a change in manager behavior without an increase in actual short-selling activity. In such a case, the authors must establish how, and why, such an effect is possible. Researchers should establish and attempt to provide supporting evidence of causal channels. For example, in their analysis of Regulation SHO, De Angelis, Grullon, and Michenaud (2017) cite a letter to the SEC that argues that many firms were worried that the removal of the uptick rule could affect their stock prices.

B.3. Compound Exclusion Restrictions

Finally, Morck and Yeung (2011) note that "each successful use of an instrument creates an additional latent variable problem for all other uses of that instrument." This concern applies more generally within the context of all natural experiments, not just IV settings. Researchers reusing an experimental setting should reconcile their exclusion restrictions with existing empirical evidence available when their study is written. As a hypothetical example, suppose that a research team discovers a natural experiment that changes variable Y_1 because it changes variable X. Suppose another research team later

[28] Note that simply adding more observations surrounding the same experiment does not solve the problem, as the source of the exogenous variation is still the same.

examines the same setting and finds a statistically significant result for variable Y_2. The typical exclusion restriction states that the experiment affects Y_2 only through X, but there is already evidence that Y_1 changes too. Accordingly, the researchers should reconcile their exclusion restriction with this existing evidence.[29] While some recent work attempts to obtain statistical inference on the validity of exclusion restrictions (Kiviet (2020)), these issues are typically addressed through rhetorical reasoning. In practice, few of the business combination laws and Regulation SHO papers reconcile their exclusion restriction with the voluminous 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)).

V. 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 along many dimensions (Angrist and Pischke (2010)), we show that the repeated reuse of these settings may lead to p-values that cannot be interpreted in the usual manner. 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 that, "[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.

We document that two extensively studied natural experiments, namely, business combination laws and the Regulation SHO pilot, have been used to examine more than 500 different dependent variables. We also note that business combination laws and Regulation SHO are not alone—our arguments apply to many other frequently reused natural experiments in social sciences. For example, Mellon (2021) documents 176 different outcomes examined using rainfall as an IV, the Russell stock index reconstitution has been reused in more than 80 different studies, the U.S. Tick Size Pilot has been reused in more than 60 different studies, and Universal Demand Laws have been reused in more than 30 studies.[30]

To aid future research, we provide guidelines for inference when an experiment is reused. We use simulations to estimate adjusted critical values as a function of the number of times a setting is examined. We also show that multiple testing adjusted t-statistics are significantly more accurate than unadjusted t-statistics. Finally, we apply our recommendations to existing findings from research on business combination laws and the Regulation SHO pilot. We find that many results in the literature that were statistically significant using single hypothesis testing do not survive corrections for multiple hypothesis

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

[30] These numbers are based on Google Scholar and Appel (2019) for Universal Demand Laws.

testing. Overall, 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.

Initial submission: December 8, 2021; Accepted: February 15, 2022
Editors: Stefan Nagel, Philip Bond, Amit Seru, and Wei Xiong

Appendix

Table AI
Adjusted Critical Values

This table presents adjusted *t*-statistic critical values for four types of simulated settings: randomized control trials, the staggered introduction of state-level changes (staggered introductions), instrumental variables, and regression discontinuity designs. We examine 293 outcome variables obtained from Compustat and CRSP in random order. For each new outcome variable, we apply multiple testing corrections to the family of tests that includes that outcome and all previously tested outcomes. The simulated results are then averaged across 10 random orderings for each of 10 independent simulations. For each additional outcome added to the family of tests, we compute adjusted critical values using the FWER correction of Romano and Wolf (2005, 2016; RW). Panel A presents results for randomized control trials, Panel B presents results for staggered introductions, Panel C presents results for instrumental variables, and Panel D presents results for regression discontinuity designs. The outcomes variables are listed in Internet Appendix Table IA.I.

Panel A: Randomized Control Trials

#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*
1	1.96	60	3.28	119	3.45	178	3.55	237	3.61
2	2.18	61	3.28	120	3.46	179	3.55	238	3.61
3	2.34	62	3.29	121	3.46	180	3.55	239	3.61
4	2.45	63	3.29	122	3.46	181	3.55	240	3.61
5	2.54	64	3.30	123	3.46	182	3.55	241	3.61
6	2.60	65	3.30	124	3.46	183	3.55	242	3.61
7	2.65	66	3.30	125	3.46	184	3.55	243	3.61
8	2.69	67	3.31	126	3.47	185	3.56	244	3.61
9	2.73	68	3.31	127	3.47	186	3.56	245	3.62
10	2.76	69	3.31	128	3.47	187	3.56	246	3.62
11	2.80	70	3.32	129	3.47	188	3.56	247	3.62
12	2.82	71	3.32	130	3.47	189	3.56	248	3.62
13	2.85	72	3.32	131	3.47	190	3.56	249	3.62
14	2.87	73	3.33	132	3.48	191	3.56	250	3.62
15	2.89	74	3.33	133	3.48	192	3.56	251	3.62
16	2.91	75	3.34	134	3.48	193	3.57	252	3.62
17	2.93	76	3.34	135	3.48	194	3.57	253	3.62
18	2.95	77	3.34	136	3.48	195	3.57	254	3.62
19	2.96	78	3.35	137	3.49	196	3.57	255	3.62
20	2.98	79	3.35	138	3.49	197	3.57	256	3.63

(*Continued*)

Table AI—*Continued*

Panel A: Randomized Control Trials

#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*
21	2.99	80	3.35	139	3.49	198	3.57	257	3.63
22	3.00	81	3.36	140	3.49	199	3.57	258	3.63
23	3.01	82	3.36	141	3.49	200	3.57	259	3.63
24	3.03	83	3.36	142	3.49	201	3.58	260	3.63
25	3.04	84	3.36	143	3.50	202	3.58	261	3.63
26	3.05	85	3.37	144	3.50	203	3.58	262	3.63
27	3.06	86	3.37	145	3.50	204	3.58	263	3.63
28	3.07	87	3.37	146	3.50	205	3.58	264	3.63
29	3.08	88	3.38	147	3.50	206	3.58	265	3.63
30	3.09	89	3.38	148	3.50	207	3.58	266	3.63
31	3.10	90	3.38	149	3.51	208	3.58	267	3.63
32	3.11	91	3.39	150	3.51	209	3.58	268	3.64
33	3.12	92	3.39	151	3.51	210	3.58	269	3.64
34	3.12	93	3.39	152	3.51	211	3.59	270	3.64
35	3.13	94	3.40	153	3.51	212	3.59	271	3.64
36	3.14	95	3.40	154	3.51	213	3.59	272	3.64
37	3.15	96	3.40	155	3.52	214	3.59	273	3.64
38	3.16	97	3.40	156	3.52	215	3.59	274	3.64
39	3.16	98	3.41	157	3.52	216	3.59	275	3.64
40	3.17	99	3.41	158	3.52	217	3.59	276	3.64
41	3.18	100	3.41	159	3.52	218	3.59	277	3.64
42	3.18	101	3.41	160	3.52	219	3.59	278	3.64
43	3.19	102	3.42	161	3.52	220	3.59	279	3.65
44	3.20	103	3.42	162	3.53	221	3.60	280	3.65
45	3.20	104	3.42	163	3.53	222	3.60	281	3.65
46	3.21	105	3.42	164	3.53	223	3.60	282	3.65
47	3.21	106	3.43	165	3.53	224	3.60	283	3.65
48	3.22	107	3.43	166	3.53	225	3.60	284	3.65
49	3.23	108	3.43	167	3.53	226	3.60	285	3.65
50	3.23	109	3.43	168	3.53	227	3.60	286	3.65
51	3.24	110	3.43	169	3.54	228	3.60	287	3.65
52	3.24	111	3.44	170	3.54	229	3.60	288	3.65
53	3.25	112	3.44	171	3.54	230	3.60	289	3.65
54	3.25	113	3.44	172	3.54	231	3.60	290	3.65
55	3.26	114	3.44	173	3.54	232	3.60	291	3.65
56	3.26	115	3.45	174	3.54	233	3.61	292	3.66
57	3.27	116	3.45	175	3.54	234	3.61	293	3.66
58	3.27	117	3.45	176	3.54	235	3.61		
59	3.28	118	3.45	177	3.55	236	3.61		

Panel B: Staggered Introductions

#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*
1	1.96	60	3.30	119	3.48	178	3.58	237	3.65
2	2.19	61	3.30	120	3.48	179	3.57	238	3.64
3	2.35	62	3.31	121	3.48	180	3.57	239	3.65

(*Continued*)

Table AI—*Continued*

Panel B: Staggered Introductions

#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*
4	2.46	63	3.31	122	3.49	181	3.57	240	3.65
5	2.53	64	3.32	123	3.49	182	3.57	241	3.65
6	2.59	65	3.32	124	3.49	183	3.58	242	3.65
7	2.65	66	3.32	125	3.49	184	3.58	243	3.65
8	2.70	67	3.33	126	3.49	185	3.58	244	3.65
9	2.74	68	3.33	127	3.50	186	3.58	245	3.65
10	2.78	69	3.34	128	3.50	187	3.58	246	3.65
11	2.81	70	3.34	129	3.50	188	3.58	247	3.65
12	2.84	71	3.34	130	3.50	189	3.58	248	3.66
13	2.86	72	3.35	131	3.50	190	3.58	249	3.66
14	2.88	73	3.35	132	3.50	191	3.59	250	3.66
15	2.90	74	3.36	133	3.51	192	3.59	251	3.66
16	2.92	75	3.36	134	3.51	193	3.59	252	3.66
17	2.94	76	3.36	135	3.51	194	3.59	253	3.66
18	2.96	77	3.37	136	3.51	195	3.59	254	3.66
19	2.98	78	3.37	137	3.51	196	3.59	255	3.66
20	2.99	79	3.37	138	3.51	197	3.59	256	3.66
21	3.00	80	3.38	139	3.52	198	3.60	257	3.66
22	3.02	81	3.38	140	3.52	199	3.60	258	3.67
23	3.03	82	3.38	141	3.52	200	3.60	259	3.67
24	3.04	83	3.39	142	3.52	201	3.60	260	3.67
25	3.05	84	3.39	143	3.52	202	3.60	261	3.67
26	3.07	85	3.39	144	3.52	203	3.60	262	3.67
27	3.08	86	3.40	145	3.53	204	3.60	263	3.67
28	3.09	87	3.40	146	3.53	205	3.61	264	3.67
29	3.10	88	3.40	147	3.53	206	3.61	265	3.67
30	3.11	89	3.40	148	3.53	207	3.61	266	3.67
31	3.12	90	3.41	149	3.53	208	3.61	267	3.67
32	3.12	91	3.41	150	3.53	209	3.61	268	3.67
33	3.13	92	3.41	151	3.53	210	3.61	269	3.67
34	3.14	93	3.42	152	3.53	211	3.61	270	3.68
35	3.15	94	3.42	153	3.54	212	3.62	271	3.68
36	3.16	95	3.42	154	3.54	213	3.62	272	3.68
37	3.17	96	3.42	155	3.54	214	3.62	273	3.68
38	3.17	97	3.43	156	3.54	215	3.62	274	3.68
39	3.18	98	3.43	157	3.54	216	3.62	275	3.68
40	3.19	99	3.43	158	3.54	217	3.62	276	3.68
41	3.19	100	3.43	159	3.55	218	3.62	277	3.68
42	3.20	101	3.44	160	3.55	219	3.62	278	3.68
43	3.21	102	3.44	161	3.55	220	3.63	279	3.68
44	3.21	103	3.44	162	3.55	221	3.63	280	3.68
45	3.22	104	3.44	163	3.55	222	3.63	281	3.68
46	3.23	105	3.45	164	3.55	223	3.63	282	3.68
47	3.23	106	3.45	165	3.55	224	3.63	283	3.69
48	3.24	107	3.45	166	3.55	225	3.63	284	3.69
49	3.24	108	3.45	167	3.56	226	3.63	285	3.69
50	3.25	109	3.46	168	3.56	227	3.63	286	3.69

(*Continued*)

Table AI—*Continued*

Panel B: Staggered Introductions

#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*
51	3.26	110	3.46	169	3.56	228	3.63	287	3.69
52	3.26	111	3.46	170	3.56	229	3.64	288	3.69
53	3.27	112	3.46	171	3.56	230	3.64	289	3.69
54	3.27	113	3.47	172	3.56	231	3.64	290	3.69
55	3.28	114	3.47	173	3.56	232	3.64	291	3.69
56	3.28	115	3.47	174	3.57	233	3.64	292	3.69
57	3.29	116	3.47	175	3.57	234	3.64	293	3.69
58	3.29	117	3.48	176	3.57	235	3.64		
59	3.30	118	3.48	177	3.57	236	3.64		

Panel C: Instrumental Variables

#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*
1	1.96	60	3.25	119	3.42	178	3.52	237	3.58
2	2.19	61	3.26	120	3.42	179	3.52	238	3.58
3	2.34	62	3.26	121	3.43	180	3.52	239	3.58
4	2.45	63	3.27	122	3.43	181	3.52	240	3.58
5	2.53	64	3.27	123	3.43	182	3.52	241	3.58
6	2.59	65	3.27	124	3.43	183	3.52	242	3.58
7	2.64	66	3.28	125	3.44	184	3.52	243	3.59
8	2.68	67	3.28	126	3.44	185	3.53	244	3.59
9	2.72	68	3.29	127	3.44	186	3.53	245	3.59
10	2.75	69	3.29	128	3.44	187	3.53	246	3.59
11	2.78	70	3.29	129	3.44	188	3.53	247	3.59
12	2.80	71	3.30	130	3.44	189	3.53	248	3.59
13	2.83	72	3.30	131	3.45	190	3.53	249	3.59
14	2.85	73	3.30	132	3.45	191	3.53	250	3.59
15	2.87	74	3.31	133	3.45	192	3.53	251	3.59
16	2.89	75	3.31	134	3.45	193	3.54	252	3.59
17	2.91	76	3.32	135	3.45	194	3.54	253	3.59
18	2.92	77	3.32	136	3.46	195	3.54	254	3.59
19	2.94	78	3.32	137	3.46	196	3.54	255	3.60
20	2.96	79	3.33	138	3.46	197	3.54	256	3.60
21	2.97	80	3.33	139	3.46	198	3.54	257	3.60
22	2.98	81	3.33	140	3.46	199	3.54	258	3.60
23	3.00	82	3.33	141	3.47	200	3.54	259	3.60
24	3.01	83	3.34	142	3.47	201	3.54	260	3.60
25	3.02	84	3.34	143	3.47	202	3.55	261	3.60
26	3.03	85	3.34	144	3.47	203	3.55	262	3.60
27	3.04	86	3.34	145	3.47	204	3.55	263	3.60
28	3.05	87	3.35	146	3.47	205	3.55	264	3.60
29	3.06	88	3.35	147	3.48	206	3.55	265	3.60
30	3.07	89	3.35	148	3.48	207	3.55	266	3.60
31	3.08	90	3.36	149	3.48	208	3.55	267	3.60
32	3.09	91	3.36	150	3.48	209	3.55	268	3.61
33	3.10	92	3.36	151	3.48	210	3.55	269	3.61

(*Continued*)

Table AI—*Continued*

Panel C: Instrumental Variables

#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*
34	3.10	93	3.36	152	3.48	211	3.55	270	3.61
35	3.11	94	3.37	153	3.48	212	3.56	271	3.61
36	3.12	95	3.37	154	3.48	213	3.56	272	3.61
37	3.13	96	3.37	155	3.49	214	3.56	273	3.61
38	3.13	97	3.38	156	3.49	215	3.56	274	3.61
39	3.14	98	3.38	157	3.49	216	3.56	275	3.61
40	3.15	99	3.38	158	3.49	217	3.56	276	3.61
41	3.15	100	3.38	159	3.49	218	3.56	277	3.61
42	3.16	101	3.39	160	3.49	219	3.56	278	3.61
43	3.17	102	3.39	161	3.50	220	3.56	279	3.61
44	3.17	103	3.39	162	3.50	221	3.57	280	3.62
45	3.18	104	3.39	163	3.50	222	3.57	281	3.62
46	3.18	105	3.39	164	3.50	223	3.57	282	3.62
47	3.19	106	3.40	165	3.50	224	3.57	283	3.62
48	3.20	107	3.40	166	3.50	225	3.57	284	3.62
49	3.20	108	3.40	167	3.50	226	3.57	285	3.62
50	3.21	109	3.40	168	3.50	227	3.57	286	3.62
51	3.21	110	3.41	169	3.50	228	3.57	287	3.62
52	3.22	111	3.41	170	3.51	229	3.57	288	3.62
53	3.22	112	3.41	171	3.51	230	3.57	289	3.62
54	3.23	113	3.41	172	3.51	231	3.57	290	3.62
55	3.23	114	3.41	173	3.51	232	3.58	291	3.62
56	3.24	115	3.41	174	3.51	233	3.58	292	3.62
57	3.24	116	3.42	175	3.51	234	3.58	293	3.62
58	3.24	117	3.42	176	3.52	235	3.58		
59	3.25	118	3.42	177	3.52	236	3.58		

Panel D: Regression Discontinuity Designs

#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*
1	1.96	60	3.26	119	3.43	178	3.52	237	3.59
2	2.19	61	3.26	120	3.43	179	3.52	238	3.59
3	2.34	62	3.27	121	3.43	180	3.52	239	3.59
4	2.44	63	3.27	122	3.44	181	3.52	240	3.59
5	2.52	64	3.27	123	3.44	182	3.53	241	3.59
6	2.58	65	3.28	124	3.44	183	3.53	242	3.59
7	2.63	66	3.28	125	3.44	184	3.53	243	3.59
8	2.68	67	3.29	126	3.44	185	3.53	244	3.59
9	2.71	68	3.29	127	3.45	186	3.53	245	3.60
10	2.75	69	3.29	128	3.45	187	3.53	246	3.60
11	2.78	70	3.30	129	3.45	188	3.53	247	3.60
12	2.81	71	3.30	130	3.45	189	3.53	248	3.60
13	2.83	72	3.30	131	3.45	190	3.54	249	3.60
14	2.85	73	3.31	132	3.46	191	3.54	250	3.60
15	2.87	74	3.31	133	3.46	192	3.54	251	3.60
16	2.89	75	3.31	134	3.46	193	3.54	252	3.60

(*Continued*)

Table AI—*Continued*

Panel D: Regression Discontinuity Designs

#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*	#Outcomes	*t*
17	2.91	76	3.32	135	3.46	194	3.54	253	3.60
18	2.92	77	3.32	136	3.46	195	3.54	254	3.60
19	2.94	78	3.32	137	3.47	196	3.54	255	3.60
20	2.96	79	3.33	138	3.47	197	3.54	256	3.61
21	2.97	80	3.33	139	3.47	198	3.55	257	3.61
22	2.98	81	3.33	140	3.47	199	3.55	258	3.61
23	2.99	82	3.34	141	3.47	200	3.55	259	3.61
24	3.01	83	3.34	142	3.47	201	3.55	260	3.61
25	3.02	84	3.34	143	3.47	202	3.55	261	3.61
26	3.03	85	3.34	144	3.48	203	3.55	262	3.61
27	3.04	86	3.35	145	3.48	204	3.55	263	3.61
28	3.05	87	3.35	146	3.48	205	3.55	264	3.61
29	3.06	88	3.35	147	3.48	206	3.56	265	3.61
30	3.07	89	3.36	148	3.48	207	3.56	266	3.62
31	3.08	90	3.36	149	3.48	208	3.56	267	3.62
32	3.09	91	3.36	150	3.49	209	3.56	268	3.62
33	3.10	92	3.36	151	3.49	210	3.56	269	3.62
34	3.10	93	3.37	152	3.49	211	3.56	270	3.62
35	3.11	94	3.37	153	3.49	212	3.56	271	3.62
36	3.12	95	3.37	154	3.49	213	3.56	272	3.62
37	3.13	96	3.37	155	3.49	214	3.56	273	3.62
38	3.13	97	3.38	156	3.49	215	3.56	274	3.62
39	3.14	98	3.38	157	3.49	216	3.57	275	3.62
40	3.15	99	3.38	158	3.50	217	3.57	276	3.62
41	3.15	100	3.39	159	3.50	218	3.57	277	3.62
42	3.16	101	3.39	160	3.50	219	3.57	278	3.62
43	3.17	102	3.39	161	3.50	220	3.57	279	3.63
44	3.17	103	3.39	162	3.50	221	3.57	280	3.63
45	3.18	104	3.39	163	3.50	222	3.57	281	3.63
46	3.19	105	3.40	164	3.50	223	3.57	282	3.63
47	3.19	106	3.40	165	3.51	224	3.57	283	3.63
48	3.20	107	3.40	166	3.51	225	3.58	284	3.63
49	3.20	108	3.40	167	3.51	226	3.58	285	3.63
50	3.21	109	3.41	168	3.51	227	3.58	286	3.63
51	3.21	110	3.41	169	3.51	228	3.58	287	3.63
52	3.22	111	3.41	170	3.51	229	3.58	288	3.63
53	3.22	112	3.41	171	3.51	230	3.58	289	3.63
54	3.23	113	3.42	172	3.51	231	3.58	290	3.64
55	3.24	114	3.42	173	3.52	232	3.58	291	3.64
56	3.24	115	3.42	174	3.52	233	3.58	292	3.64
57	3.24	116	3.42	175	3.52	234	3.58	293	3.64
58	3.25	117	3.43	176	3.52	235	3.59		
59	3.25	118	3.43	177	3.52	236	3.59		

REFERENCES

Alexander, Gordon J., and Mark A. Peterson, 2008, The effect of price tests on trader behavior and market quality: An analysis of Reg SHO, *Journal of Financial Markets* 11, 84–111.

Andrikogiannopoulou, Angie, and Filippos Papakonstantinou, 2019, Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? *Journal of Finance* 74, 2667–2688.

Angrist, Joshua D., and Alan B. Kreuger, 2001, Instrumental variables and the search for identification: From supply and demand to natural experiments, *Journal of Economic Perspectives* 15, 69–85.

Angrist, Joshua D., and Jörn-Steffen Pischke, 2010, The credibility revolution in empirical economics: How better research design is taking the con out of econometrics, *Journal of Economic Perspectives* 24, 3–30.

Appel, Ian, 2019, Governance by litigation. Available at SSRN 253227.

Benjamini, Yoav, and Yosef Hochberg, 1995, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 289–300.

Benjamini, Yoav, and Daniel Yekutieli, 2001, The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics* 29, 1165–1188.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, 2004, How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* 119, 249–275.

Bloom, Howard S., 1995, Minimum detectable effects: A simple way to report the statistical power of experimental designs, *Evaluation Review* 19, 547–556.

Bonferroni, Carlo E., 1936, *Teoria statistica delle classi e calcolo delle probabilità* (Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, Florence, Italy).

Bowen, Donald E., III, Laurent Frésard, and Jérôme P. Taillard, 2017, What's your identification strategy? Innovation in corporate finance research, *Management Science* 63, 2529–2548.

Bretz, Frank, Alex Dmitrienko, and Ajit C. Tamhane, 2010, *Multiple Testing Problems in Pharmaceutical Statistics* (Chapman & Hall/CRC Biostatistics Series, Boca Raton, FL).

Cain, Matthew D., Stephen B. McKeon, and Steven Davidoff Solomon, 2017, Do takeover laws matter? Evidence from five decades of hostile takeovers, *Journal of Financial Economics* 124, 464–485.

Chordia, Tarun, Amit Goyal, and Alessio Saretto, 2020, Anomalies and false rejections, *Review of Financial Studies* 33, 2134–2179.

Clarke, Damian, Joseph P. Romano, and Michael Wolf, 2020, The Romano–Wolf multiple-hypothesis correction in Stata, *Stata Journal* 20, 812–843.

Comment, Robert, and G. William Schwert, 1995, Poison or placebo? Evidence on the deterrence and wealth effects of modern antitakeover measures, *Journal of Financial Economics* 39, 3–43.

De Angelis, David, Gustavo Grullon, and Sébastien Michenaud, 2017, The effects of short-selling threats on incentive contracts: Evidence from an experiment, *Review of Financial Studies* 30, 1627–1659.

Diether, Karl B., Kuan-Hui Lee, and Ingrid M. Werner, 2009, It's SHO time! Short-sale price tests and market quality, *Journal of Finance* 64, 37–73.

Dmitrienko, Alex, and Ajit C. Tamhane, 2007, Gatekeeping procedures with clinical trial applications, *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 6, 171–180.

Engelberg, Joseph, R. David McLean, Jeffrey Pontiff, and Matthew C. Ringgenberg, 2023, Do cross-sectional predictors contain systematic information? *Journal of Financial and Quantitative Analysis* 58, 1172–1201.

Fisher, Ronald A, 1925, *Statistical Methods for Research Workers* (Oliver & Boyd, Edinburgh, Scotland).

Foster, Dean P., and Robert A. Stine, 2008, α-Investing: A procedure for sequential control of expected false discoveries, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 429–444.

Fuchs-Schündeln, Nicola, and Tarek A. Hassan, 2017, Natural experiments in macroeconomics, *Handbook of Macroeconomics* 2, 923–2012.

Giglio, Stefano, Yuan Liao, and Dacheng Xiu, 2021, Thousands of alpha tests, *Review of Financial Studies* 34, 3456–3496.

Harvey, Campbell R., and Yan Liu, 2013, Multiple testing in economics. Available at SSRN 2358214.

Harvey, Campbell R., and Yan Liu, 2014, Evaluating trading strategies, *Journal of Portfolio Management* 40, 108–118.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, … and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

Holm, Sture, 1979, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* 6, 65–70.

Hou, Kewei, Chen Xue, and Lu Zhang, 2020, Replicating anomalies, *Review of Financial Studies* 33, 2019–2133.

Karpoff, Jonathan M., and Paul H. Malatesta, 1989, The wealth effects of second-generation state takeover legislation, *Journal of Financial Economics* 25, 291–322.

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

Kiviet, Jan, 2020, Testing the impossible: Identifying exclusion restrictions, *Journal of Econometrics* 218, 294–316.

Leamer, Edward E., 1983, Let's take the con out of econometrics, *American Economic Review* 73, 31–43.

Leamer, Edward E., 2010, Tantalus on the road to asymptopia, *Journal of Economic Perspectives* 24, 31–46.

List, John A., Azeem M. Shaikh, and Yang Xu, 2019, Multiple hypothesis testing in experimental economics, *Experimental Economics* 22, 773–793.

Liu, Lily Y., Andrew J. Patton, and Kevin Sheppard, 2015, Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes, *Journal of Econometrics* 187, 293–311.

Ludbrook, John, 1998, Multiple comparison procedures updated, *Clinical and Experimental Pharmacology and Physiology* 25, 1032–1037.

Mellon, Jonathan, 2021, Rain, rain, go away: 176 potential exclusion-restriction violations for studies using weather as an instrumental variable, Working paper.

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

Morck, Randall, and Bernard Yeung, 2011, Economics, history, and causation, *Business History Review* 85, 39–63.

Pearl, Judea, 1995, Causal diagrams for empirical research, *Biometrika* 82, 669–688.

Pearl, Judea, 2009, Causal inference in statistics: An overview, *Statistics Surveys* 3, 96–146.

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

Romano, Joseph P., and Michael Wolf, 2007, Control of generalized error rates in multiple testing, *Annals of Statistics* 35, 1378–1408.

Romano, Joseph P., and Michael Wolf, 2010, Balanced control of generalized error rates, *Annals of Statistics* 38, 598–633.

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

Romano, Joseph P., and Michael Wolf, 2018, Multiple testing of one-sided hypotheses: Combining Bonferroni and the bootstrap, International Conference of the Thailand Econometrics Society, pp. 78–94.

Rozenzwieg, Mark R., and Kenneth I. Wolpin, 2000, Natural "natural experiments" in economics, *Journal of Economic Literature* 38, 827–274.

SEC, 2007, Economic analysis of the short sale price restrictions under the Regulation SHO pilot, Securities and Exchange Commission Special Study.

Spamann, Holger, 2019, On inference when using state corporate laws for identification, Harvard Law School John M. Olin Center Discussion Paper.

Thompson, William Hedley, Jessey Wright, Patrick G Bissett, and Russell A. Poldrack, 2020, Meta-research: Dataset decay and the problem of sequential analyses on open datasets, *ELife* 9, e53498.

White, Halbert, 2000, A reality check for data snooping, *Econometrica* 68, 1097–1126.

Yekutieli, Daniel, 2008, Hierarchical false discovery rate-controlling methodology, *Journal of the American Statistical Association* 103, 309–316.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Internet Appendix.
Replication Code.