Contents lists available at ScienceDirect

Journal of Accounting and Economics

journal homepage: www.journals.elsevier.com/journal-of-accounting-and-economics

Causality redux: The evolution of empirical methods in accounting research and the growth of quasi-experiments☆

Christopher Armstrong [b], John D. Kepler [b], Delphine Samuels [c], Daniel Taylor [a, *]

[a] The Wharton School, University of Pennsylvania, USA
[b] Graduate School of Business, Stanford University, USA
[c] Booth School of Business, University of Chicago, USA

ARTICLE INFO

Article history:
Received 6 July 2022
Accepted 7 July 2022
Available online xxx

Keywords:
Causal inferences
Methodology
Quasi-experiments

ABSTRACT

This paper reviews the empirical methods used in the accounting literature to draw causal inferences. Recent years have seen a burgeoning growth in the use of methods that seek to exploit as-if random variation in observational settings—i.e., "quasi-experiments." We provide a synthesis of the major assumptions of these methods, discuss several practical considerations relevant to the application of these methods in the accounting literature, and provide a framework for thinking about whether and when quasi-experimental and non-experimental methods are well-suited for addressing causal questions of interest to accounting researchers. While there is growing interest in addressing causal questions within the literature, we caution against the idea that one should restrict attention to only those causal questions for which there are quasi-experiments. We offer a complementary approach for addressing causal questions that does not rely on the availability of a quasi-experiment, but rather relies on a combination of economic theory, developing and falsifying alternative explanations, triangulating results across multiple settings, measures, and research designs, and caveating results where appropriate.

1. Introduction

1.1. Objective

We review the empirical archival literature in accounting. Unlike traditional literature reviews that focus on topical areas, our review focuses on empirical methodology. Specifically, we review the evolution of empirical methods accounting researchers use to draw causal inferences from archival data. From research seminars, to publications, to conference discussions

☆ We appreciate helpful comments and suggestions from seminar participants at University of Minnesota, Baruch, Georgia State, Kellogg, Tulane, Tilburg University, UT Dallas, Florida State, and the 2021 JAE Conference, as well as from Mary Barth, Mary Billings, Jonathan Bonham, Donal Byrd, Ed deHaan, Stephen Glaeser, David Godsell, Ian Gow, Wayne Guay, Michelle Hanlon, Jeff Hoopes, Sterling Huang, Mark Lang (editor), Dave Larcker, Christian Leuz (*JAE* discussant), Bradford Lynch, Iván Marinovic, Miguel Minutti-Meza, Shawn Shi, Kevin Smith, and Toni Whited (*JAE* discussant). We thank our respective schools for financial support.

* Corresponding author.
E-mail addresses: carmstro@stanford.edu (C. Armstrong), jdkepler@stanford.edu (J.D. Kepler), delphine.samuels@chicagobooth.edu (D. Samuels), dtayl@wharton.upenn.edu (D. Taylor).

https://doi.org/10.1016/j.jacceco.2022.101521
0165-4101/© 2022 Elsevier B.V. All rights reserved.

Please cite this article as: C. Armstrong, J.D. Kepler, D. Samuels et al., Causality redux: The evolution of empirical methods in accounting research and the growth of quasi-experiments, Journal of Accounting and Economics, https://doi.org/10.1016/j.jacceco.2022.101521

and interactions, anecdotal evidence points to an increasing focus within the profession on improving and sharpening causal inferences. Our aim with this review is not to make normative statements about how accounting researchers *should* draw causal inferences, but rather to make positive statements about how accounting researchers *are* drawing causal inferences while highlighting the implementation challenges and tradeoffs associated with the current state-of-the-art.

The genesis of this review stems from philosophical and methodological questions we commonly receive from authors, reviewers, colleagues, and PhD students. Several of these questions, for instance, are perennial favorites at the Deloitte Doctoral Consortium. While there is a great deal of heterogeneity in the purpose of papers—and not all papers seek to address causal questions—PhD students are increasingly concerned about the importance of drawing causal inferences. For example: Does my paper need to explicitly draw causal inferences? What if the evidence doesn't support such statements, should I find another idea? What method do I need to use to draw causal inferences? Should I motivate my paper as being about this interesting setting, or as testing the underlying broader theory? Does that choice matter for my research design? Because these questions are often researcher- and setting-specific, there is no “one-size-fits-all” answer to any of these questions. As a result, this review does not seek to answer these questions *per se*, but rather seeks to provide a framework and discussion for how to think about these questions and weigh the tradeoffs from potential answers.

Notably, such a framework is not present in prior review papers on causal inferences in accounting (e.g., Gow et al., 2016), and is rarely present in common econometrics textbooks (Angrist and Pischke, 2008). As anyone who has tried to work with data knows, rarely do the data comport with the assumptions and convenient linear representations commonly found in econometrics textbooks. Indeed, the approach taken in many textbooks is that there is an unobserved linear data generating process, and the chief role of the researcher is to estimate a specification that recovers the parameters in the unobserved model. Rarely do these textbooks discuss model building and fitting when the data generating process is unknown, and rarely do they discuss the philosophy of science surrounding causal inferences. If anything, a common criticism of econometrics textbooks is that they give the impression that drawing causal inferences is straightforward—that one just needs to find a setting, argue the setting provides as-if random variation, and then estimate a canned statistical routine for a difference-in-differences estimator (see e.g., Hennessy and Strebulaev, 2020 for this criticism). As a consequence, there is a gap between the theory of causal inference (i.e., how researchers *should* draw causal inferences) and the practice of causal inference in the accounting literature (i.e., how researchers *are* drawing causal inferences). This literature review aims to highlight and span this gap.

Our review and framework emphasizes (1) the need for theory and institutional knowledge in drawing causal inferences, (2) the need for researchers interested in causal inferences to triangulate inferences across multiple research designs, test specifications, and empirical measures, and (3) the complementarity of various approaches and methodologies to drawing causal inferences. If nothing else, a reader should come away appreciating the fact that as empirical researchers, we rarely know the underlying data generating process. As a consequence, it is virtually impossible to be certain that one particular econometric specification is necessarily “more robust” or strictly dominates another. Indeed, without knowledge of the underlying data generating process, it is hard to say which econometric specification is the “better identified specification.” This is where domain expertise and theory fit in—two crucial elements for causal inference that allow researchers to rule out alternative explanations and develop well-specified tests.

1.2. *Summary*

The following provides a brief summary of our review and key takeaways from each section.

Section 2 surveys a comprehensive set of all empirical archival papers published in the *Journal of Accounting and Economics*, *Journal of Accounting Research*, and *The Accounting Review* since 2005.[1] Our survey focuses on the methods these papers employ and whether the papers seek to explicitly draw causal inferences. Similar to Angrist and Pischke (2008), we define as “quasi-experimental methods” methods that seek to use exogenous shocks and/or other stylized settings that are intended to provide as-if random variation in the explanatory variable of interest (e.g., difference-in-differences, instrumental variables, and regression discontinuity) and define all other methods as “non-experimental.” Although non-experimental methods can sometimes be used to approximate experiments in the absence of random assignment, the distinguishing feature of quasi-experimental methods is that they explicitly seek to emulate random assignment to facilitate causal inference (e.g., Angrist and Pischke, 2008).[2]

We report our survey results in Section 2.1 and provide commentary in Section 2.2. Over the past 15 years, our survey reveals a 4—5 fold increase in papers using quasi-experimental methods to draw causal inferences, that more than 75% of such papers use variations of the classic difference-in-differences design, and that 65% of papers using quasi-experimental methods to draw causal inferences study the effects of regulation. We make no judgments about whether the trends in the literature are “good” or “bad.” Instead, we discuss examples of how these trends manifest in recent research, and discuss these trends in the context of a Bayesian learning framework in which the necessary evidentiary standard to revise beliefs varies with the novelty of the theory being tested.

[1] Our survey explicitly excludes laboratory and field experiments.

[2] “In most cases [...] regression is used with observational data. Without the benefit of random assignment, regression estimates may or may not have a causal interpretation.” (Angrist and Pischke 2008, p. 22).

The survey makes clear that causal inferences are undoubtedly “en vogue,” and do not appear to be a passing fad. Given the clear and rising interest in causal inference and difference-and-differences designs, Section 3 discusses two key ingredients to causal inference—empirical methods and theoretical assumptions; Section 4 discusses practical implementation issues; and Section 5 offers a conceptual framework for evaluating the role of non-experimental methods (i.e., methods that do not seek to approximate an experiment) in facilitating causal inferences.

Section 3.1 begins by discussing how a simple pooled ordinary least squares (OLS) regression can be used to draw causal inferences. As common econometrics textbooks discuss, if the OLS assumptions hold—e.g., no correlated omitted variables exist—then pooled OLS can be used to draw causal inferences (e.g., Stock and Watson, 2003; Angrist and Pischke, 2008). This is an excellent example of how the choice of method itself does not imbue the researcher with the ability to draw causal inferences—but rather how that method's assumptions comport with the theory and data. In practice, researchers are often reluctant to assume away correlated omitted variables in simple pooled OLS models and thus rarely feel comfortable drawing strong causal inferences using pooled regressions. We discuss three popular methods that accounting researchers tend to employ to deal with correlated omitted variables and discuss the strengths and weaknesses of each method: identification of specific omitted variables, fixed effects, and cross-sectional interactions.

After introducing the omitted variable threat, Section 3.2 covers the single most common method used for drawing causal inferences in accounting research—difference-in-differences (DiD) designs.[3] Our discussion of DiD designs links back to the econometrics of cross-sectional interaction designs and the omitted variable threat. Specifically, we discuss how a DiD design represents a specific case of the more general cross-sectional interaction design—whereby a variable partitions the data into two distinct subsamples, and pooled regressions are estimated on each subsample, comparing coefficients across the two samples. We highlight that the parallel trends assumption of DiD designs is effectively equivalent to the “no correlated omitted variable assumption” in the cross-sectional interaction design. Our discussion emphasizes that the ability to draw causal inferences stems not from the application of a particular econometric method, but rather from the validity of the theoretical assumptions of that method, which often hinge critically on the researcher's institutional knowledge about the particular setting being studied.

Having discussed the *methods* commonly used to draw causal inferences, Section 3.3 discusses two important roles of *theory* in causal inference. First, theory—whether it be informal intuition or a formal economic model—is required to interpret correlations. All econometric methods estimate correlations. Some correlations have more meaning than others (see Vigen, 2015 for a compilation of purely spurious correlations). What grants correlations meaning is how we interpret them in light of theoretical assumptions. Without theory, the evidentiary value of bivariate correlations is no greater or less than the evidentiary value of correlations estimated from the most rigorous econometric method. Theory is what allows researchers to separate spurious inferences—i.e., those based on random correlations with no particular meaning—from causal inferences. The more precise the theory, the more meaningful the estimated correlations, and the more credible the resulting inferences.

To illustrate this point, we extend the analysis in Vigen (2015) to the accounting literature and use a simple exercise to illustrate the danger of an atheoretical approach to causal inference. We provide the reader with a dataset (and code) for over a dozen commonly used measures of earnings management (e.g., discretionary accruals and restatements) and voluntary disclosure (e.g., management forecasts and voluntary Forms 8-K).[4] We then use a staggered adoption DiD design with *Firm*, *Industry* x *State*, and *Industry* x *Year* fixed effects to document the “causal effect” of state laws restricting workplace smoking, as well as state laws restricting access to firearms, on management forecasts. If one believes that the ability to draw causal inferences is solely a matter of applying a particular econometric method to a particular setting, then we are the first to show that workplace smoking laws and gun laws *cause* voluntary disclosure. Alternatively, if causal inference requires theory, then having offered none, these results should not be interpreted in a causal manner.

Here it is useful to point out that a staggered DiD design still estimates correlations. It is up to the reader whether these correlations should be interpreted in a causal manner. We emphasize, correlations—even when estimated using state-of-the-art methods—are merely descriptive; they have no interpretable content without an underlying theory. Indeed, in the absence of theory, a researcher can arbitrarily select the regression specification and measure of voluntary disclosure (as we have done). And, in the absence of triangulation across multiple specifications and multiple measures, the reader should be hesitant to draw causal inferences (a point we return to in Section 4).[5]

Second, theory is required to generalize inferences. The process of generalizability refers to extrapolating inferences learned from a single empirical test (or set of tests) to circumstances occurring out of sample—and ultimately to the underlying theory being tested. The ability to generalize one's inferences beyond a stylized setting hinges critically on the strength of the paper's theoretical foundation. If the theory is not compelling, then inferences are necessarily limited to the setting being studied. We discuss settings and circumstances where generalizability is (and is not) a concern. We discuss how the importance of generalizability depends not on the method, or the setting, but on the specific research question being studied. For example, if one is interested in the causal effect of International Financial Reporting Standards (IFRS) on a

[3] See Larcker and Rusticus (2010) for a discussion of instrumental variables and Lee and Lemieux (2010) for a discussion of regression discontinuity designs.

[4] See Internet Appendix.

[5] In subsequent sections we will show that the Table 1 results are an artifact of selective reporting. The results in Table 1 are ‘knife edge,’ they do not hold with alternative fixed effect structures or alternative measures of voluntary disclosure.

particular outcome, then one need not worry about generalizing inferences beyond IFRS (Barth et al., 2012). However, if one is interested in testing a general theory of how information quality relates to corporate payout policy using IFRS as an exogenous shock, then generalizability beyond IFRS is a concern (Hail et al., 2014). In our discussion of these points, we draw on several examples in the literature that highlight how concerns about generalizability are a function of the research question and rely on the theoretical links between the empirical measures and the underlying theoretical constructs.

Having covered the conceptual underpinnings for causal inference, Section 4 turns to more practical issues and associated implementation challenges. Our literature survey reveals three core implementation challenges concerning causal inferences that researchers often face. We cover each of these in turn.

Section 4.1 highlights the distinction between an event that is exogenous and an event that provides as-if random assignment. Even though this distinction is critical for causal inference (see, for example, Atanasov and Black (2016) and Hennessey and Strebulaev (2020) for recent discussions in finance), many papers in our survey still do not make this distinction. Consider the definition of exogenous in the Oxford dictionary: “Having an external cause or origin. Often contrasted with endogenous: e.g., technological changes exogenous to the oil industry.” This definition makes clear that an exogenous event (or variable) refers to something that originates outside the system being studied. This speaks to the origin of the event, but not whether it provides as-if random variation, i.e., whether the event randomly distributes the quantity of interest among affected firms. In this regard, an event can be plausibly exogenous to the firm being studied (e.g., the introduction of an accounting standard) but fail to provide as-if random assignment to treatment and control groups.

We take two approaches to illustrating the importance of distinguishing between the concepts of exogeneity and as-if random variation. First, we discuss these concepts in the context of several regulatory settings studied in accounting research. In particular, the NYSE/NASDAQ board independence requirements and the California and Norwegian board gender diversity requirements. Although arguably exogenous, the effects of such regulations on the firm are often—by design—a mechanical function of the firm's previously endogenous choices, and therefore not as-if random.[6] Second, we discuss the concepts in the context of a theoretical economy in which a previously voluntary behavior is made mandatory. In this theoretical economy, the effect of the mandatory action, although plausibly exogenous, is explicitly conditional on choices that were previously voluntary, which we show can lead to biased estimates of the causal effect.

Section 4.2 discusses common tests that are useful for assessing whether the parallel trends assumption of the DiD design holds. Although the parallel trends assumption is inherently untestable, diagnostic tests can provide useful insights on potential violations of the assumption. Indeed, it is becoming increasingly common for authors to present estimates of the “treatment effect” (i.e., the difference between treatment and control groups) graphically over time: several periods before and after the treatment was administered. Although these diagnostic tests are useful, given the reliance on graphical representations and the fact that different readers can interpret the same graph differently, there is often considerable subjectivity in whether a given graph does or does not support the parallel trends assumption. Consequently, these diagnostic tests are not a panacea—they are neither necessary nor sufficient for causal inferences.

Section 4.3 discusses two tradeoffs associated with high-dimensional fixed effect designs common in the literature. One of the hallmarks of many difference-in-differences designs used in the literature is the inclusion of high-dimensional fixed effects (i.e., a large number of fixed effects). As discussed in Section 3.1, these methods are useful for helping rule out correlated omitted variables. However, these methods are not without tradeoffs. First, we show that when the source of the variation in a correlated omitted variable is within-group (rather than across groups), including group fixed effects can *exacerbate* omitted variable bias. Second, we show that including high-dimensional fixed effects can induce significant multicollinearity and increase the sensitivity of regression results to a handful of observations. Given the potential for false positives and fragility that we document, we encourage researchers to: (1) explicitly motivate their choice of fixed effects, (2) present regression diagnostics for multicollinearity, (3) report the amount of variation in the independent variable absorbed by the fixed effects, and (4) triangulate inferences across alternative fixed effect structures. If results are sensitive to a particular specification, we encourage transparent reporting and discussion of that sensitivity (see Bianchi et al., 2021 for an excellent example). We caution against the temptation to infer (or ex post justify) that the specification that yields statistical significance in the predicted direction is the correct specification.

Having discussed the implementation challenges associated with quasi-experimental methods, Section 5 takes a step back and offers a conceptual framework for evaluating the role of non-experimental methods (i.e., methods that do not seek to approximate an experiment) in facilitating causal inferences. In particular, we ask three questions: (i) Does one need an experimental or quasi-experimental setting to address a causal question? (ii) How can non-experimental and quasi-experimental evidence be combined in the context of a single study to identify causal mechanisms? (iii) When quasi-experimental evidence seemingly conflicts with non-experimental evidence, should we prioritize the former over the latter?

In contrast to the conventional wisdom in the accounting literature, we offer the view that there are some causal questions for which quasi-experiments are ill-suited. We adopt several stylized examples to illustrate how evidence from settings *without* as-if random variation can nonetheless be helpful for addressing causal questions. We discuss the advantages of

[6] For example, a regulation that the board must include at least two female directors only “treats” those boards who endogenously chose not to have two (or more) female directors prior to the regulation. In such a circumstance, whether a firm or board is “treated” is a mechanical function of an endogenous choice.

combining quasi-experimental and non-experimental settings in the context of a single study, and we discuss the dangers of dismissing evidence from non-experimental settings—even those with pervasive endogeneity issues.

We offer concluding thoughts in Section 6. Our review of the accounting literature leads us to conclude that drawing reliable causal inferences is very challenging—and much more challenging than simply the choice of method. Reliable causal inferences require compelling economic theory, methods that make assumptions that comport with the institutional setting being studied, and a plethora of robustness tests to triangulate inferences across (often implicit) theoretical assumptions. Despite their best efforts, sometimes researchers cannot find a setting that approximates the experimental ideal and, in these cases, it is acceptable—even desirable—to provide evidence using non-experimental methods with appropriate caveats. We caution against the idea that one should restrict attention to only those causal questions for which there are settings conducive to quasi-experimental methods.

2. Evolution of empirical methods

2.1. Survey of empirical papers

2.1.1. Survey method

In order to identify trends in the evolution of empirical archival methods in the accounting literature, we build an inventory of all empirical studies published in the JAE, JAR, and TAR between 2005 and 2019. To detect empirical archival studies, we identify all papers that use any of the following keywords: {standard errors, t-statistic, p-value}. Next, we read the title and abstract of each of these studies and eliminate (i) field or laboratory experiments, (ii) review papers and discussions, (iii) methods papers (e.g., Larcker and Rusticus, 2010), and (iv) papers that exclusively use theoretical models. This leaves us with a sample of 1417 empirical studies.

We next seek to identify all papers that use quasi-experimental methods and explicitly draw causal inferences from these methods. We define as "quasi-experimental methods" methods that seek to use exogenous shocks and/or other stylized settings that are intended to provide as-if random variation in the explanatory variable of interest (e.g., difference-in-differences, instrumental variables, and regression discontinuity). Although non-experimental methods (e.g., a panel regression used with observational data) can sometimes be used to approximate experiments in the absence of random assignment, the distinguishing feature of quasi-experimental methods is that they explicitly seek to emulate random assignment to facilitate causal inference (e.g., Angrist and Pischke, 2008).

To identify papers using quasi-experimental methods, we begin by flagging all studies that use any of the following keywords in their title, abstract, or body, but excluding footnotes, table captions, and references: {causal, exogenous, natural experiment, quasi, shock}. Within this set of studies, we use two approaches to identify papers using quasi-experimental methods. First, to minimize subjectivity and our own biases, we use an "automated approach" that classifies papers as quasi-experimental if they use at least two of the keywords {exogenous, natural experiment, quasi, shock}.[7] Second, we use a "manual approach" to classify the study both using such methods and explicitly drawing causal inferences by reading the title, abstract, and the two sentences surrounding each keyword. Three authors read all the material independently, and in the event of disagreement, discussed the appropriate classification.

2.1.2. Survey results

Fig. 1 plots trends in the evolution of empirical methods in accounting research and shows a significant upward trend in the number of studies using quasi-experimental methods. This trend is common to all three accounting journals.[8] Panel B reveals a noticeable rise in the percentage of papers published in JAR and JAE in 2013, with a peak at JAR (JAE) in 2015 (2018). The trend is also steadily increasing at TAR, although it appears somewhat muted relative to the other two journals.

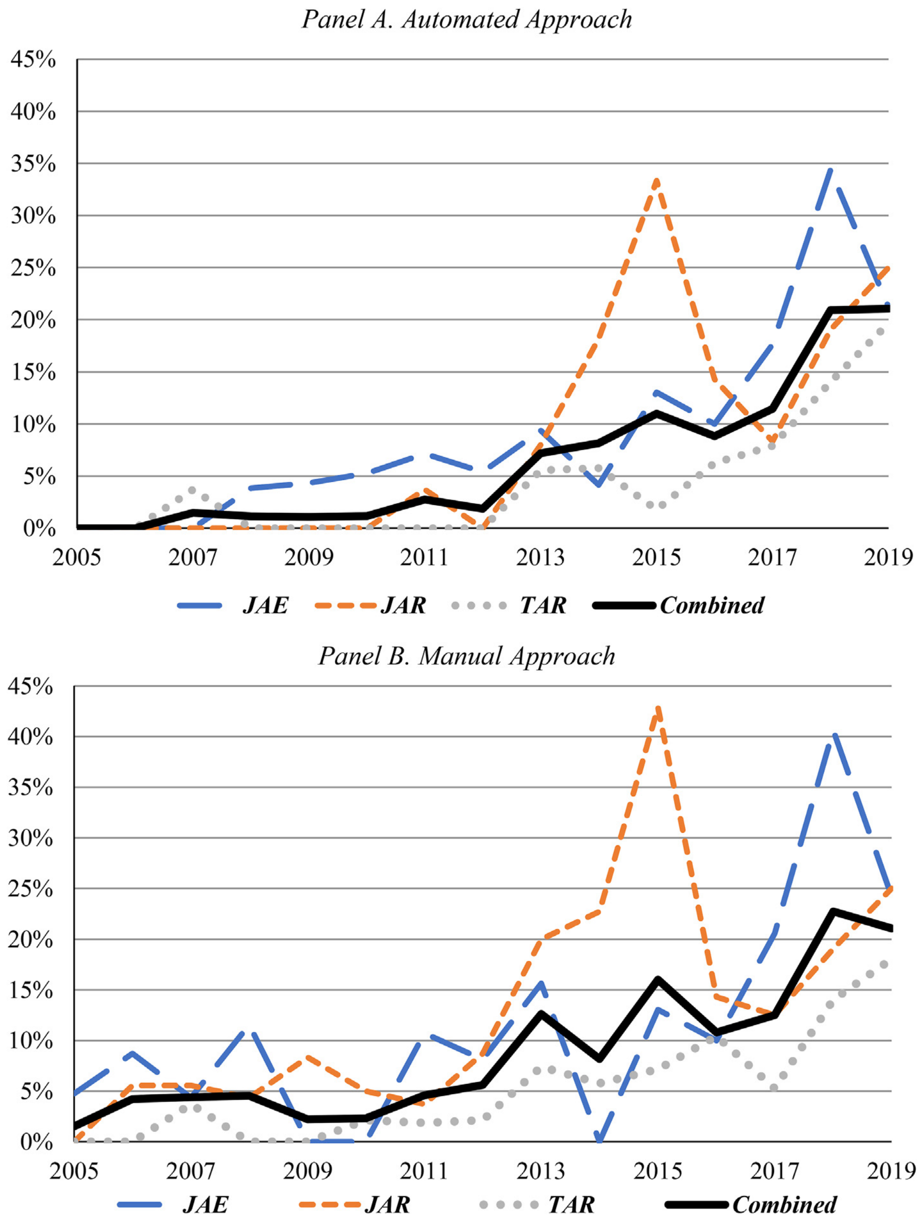
Next, we identify the primary research design that was used in these papers. Fig. 2 illustrates that the overwhelming majority of papers using quasi-experimental methods to draw causal inferences employ a differences-in-differences design, followed by instrumental variable designs, and regression discontinuity designs.

Finally, we also identify the primary research setting. Panel A of Fig. 3 illustrates that the overwhelming majority of papers study regulatory settings, followed by economic shocks at the industry or country level (e.g., the financial crisis of 2008), court cases (e.g., Supreme Court decisions), index composition (e.g., Russell 2000), analyst brokerage closures and natural phenomena (e.g., changes in weather and death).

Given the predominant focus of the literature on regulatory settings, Panel B of Fig. 3 presents statistics on the type of regulatory setting. 26% of papers studying regulatory settings examine accounting and/or auditing standards (e.g., IFRS, FASB, and PCAOB rules). 21% study disclosure standards promulgated by securities regulators (e.g., Regulation Fair Disclosure and

[7] We drop the keyword "causal" from this subset because of a significant proportion (66%) of papers use the keyword in the context of a disclaimer about the inability to draw causal inferences (e.g., "we cannot draw causal inferences").

[8] We are agnostic about the extent to which these trends reflect authors' preferences, reviewers' preferences, and/or editors' preferences.



Panel A. Automated Approach

Panel B. Manual Approach

Fig. 1. Trends in Percentage of Papers Using Quasi-Experimental Methods. This figure plots trends in the percentage of papers using quasi-experimental methods using our automated approach in Panel A and our manual approach in Panel B. We plot separate trends for papers published in the JAE, JAR and TAR, as well as all three journals combined.

Regulation SHO), 13% study state laws (e.g., universal demand laws and the Inevitable Disclosure Doctrine), 11% study tax laws, and 5% study the Sarbanes-Oxley Act of 2002.[9]

2.1.3. Citation analysis

We next attempt to quantify the impact of papers in our survey. On the one hand, if papers employing quasi-experimental methods provide novel insights and ideas to a greater extent than papers using non-experimental (i.e., observational) methods, we would expect such papers to be more highly cited. On the other hand, if papers employing quasi-experimental methods are testing well-established ideas (e.g., proprietary costs reduce voluntary disclosure), rather than generating new ideas, we would expect such papers to be less highly cited.

[9] We break SOX out into a separate category because of sheer volume of papers (5%) and the multi-faceted aspect of the Act.

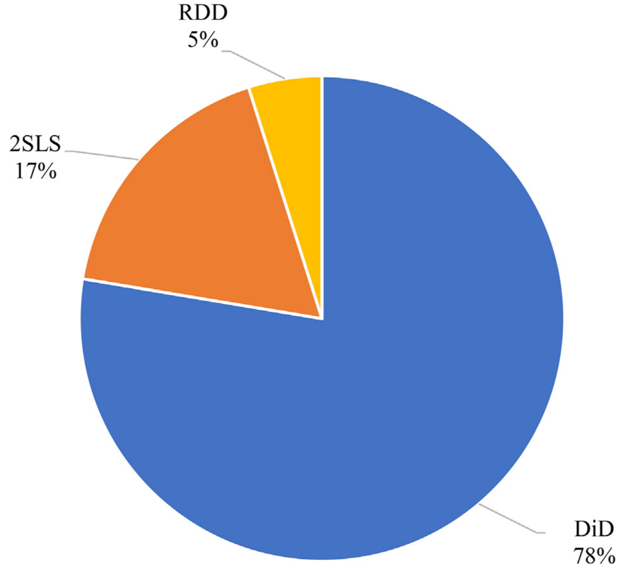


Fig. 2. Summary of Methods. This figure illustrates the empirical design in our sample of papers using quasi-experimental methods classified using the manual approach.

To identify highly cited papers, we collect Clarivate Analytics' Journal Citation Reports for each of the journals in our survey every year. These reports list the top 5 most highly cited papers published in that journal-year, hereafter referred to as "highly cited papers." Fig. 4 reports the percentage of highly cited papers using quasi-experimental methods. We also report the expected percentage of highly cited papers using quasi-experimental methods based on the earlier trends identified in the literature. For example, if 20% of empirical papers in 2014 use quasi-experimental methods, then we expect 20% of highly cited papers in 2014 to employ quasi-experimental methods (20% x 3 journals x 5 papers = 3).

On average, we find that papers using quasi-experimental methods represent 8% (13%) of the top five most highly cited papers using the automated (manual) approach to classifying papers, and that expected percentages are 19% (19%). Interestingly, we find that the proportion of highly cited papers that use quasi-experimental methods is less than the proportion of all empirical papers using quasi-experimental methods.

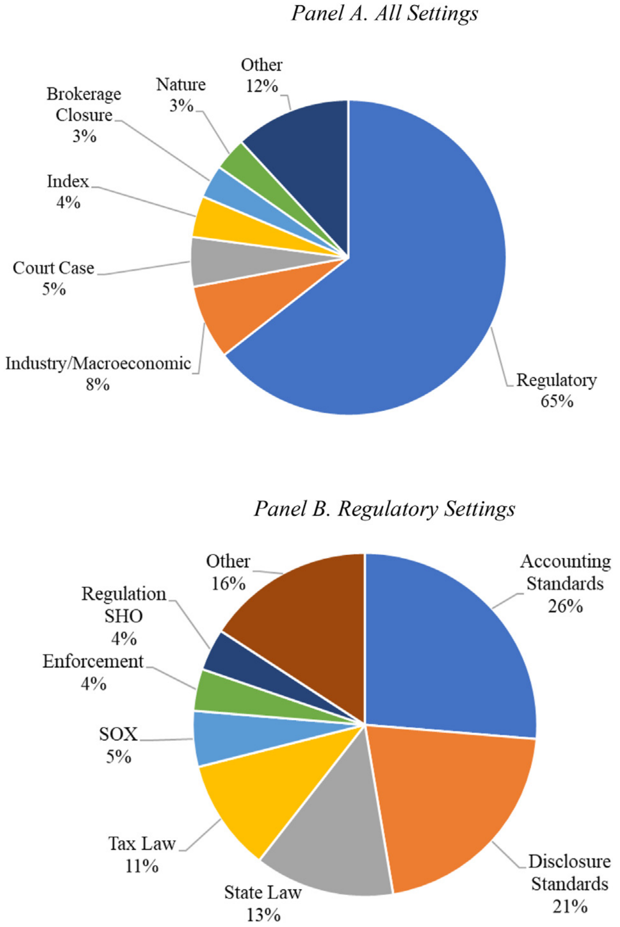
2.2. *Commentary*

In this section, we provide our thoughts on the meaning of these trends for the accounting literature. Fig. 1 suggests an increased emphasis on causal inference in accounting using quasi-experimental methods. This emphasis is a natural and welcomed progression in the empirical accounting literature—to the extent that it helps researchers draw more credible inferences. As inferences become more credible, our priors solidify, which in turn raises the evidentiary standard necessary to realize an incremental contribution.

Consider a Bayesian learning framework in which readers' priors are shaped by evidence (e.g., Glaeser and Guay, 2017; Christensen, 2019). If no evidence exists on a particular theory, then the reader has diffuse priors, and will heavily update prior beliefs when presented with new evidence. In contrast, if five decades of observational evidence on a particular theory exist (e.g., proprietary costs reduce voluntary disclosure), then the reader likely has well-defined priors, and will require very compelling evidence to update their beliefs—a much higher evidentiary standard. This framework—i.e., viewing a paper's contribution as the extent to which the reader revises their prior beliefs—makes clear that the evidentiary standard for a particular paper differs depending on the novelty of the phenomenon being studied.

Fig. 5 illustrates this tradeoff between the novelty of the theory being tested and the evidentiary standard needed to make an incremental contribution.[10] Papers in the top-left quadrant tend to provide initial evidence on new theories, for which the evidentiary standard may not be particularly high. By their very nature, however, it is difficult to design powerful tests of novel, immature theories. As theory becomes better developed and fleshed out over time, more powerful tests can be developed. As the set of collective theories studied in the literature matures, we observe that the literature as a whole tends to shift from the top-left to the bottom-right quadrant of Fig. 5, where papers tend to provide increasingly stronger evidence (e.g., causal inference).

[10] We thank Ed deHaan for suggesting this figure.



Panel A. All Settings

Panel B. Regulatory Settings

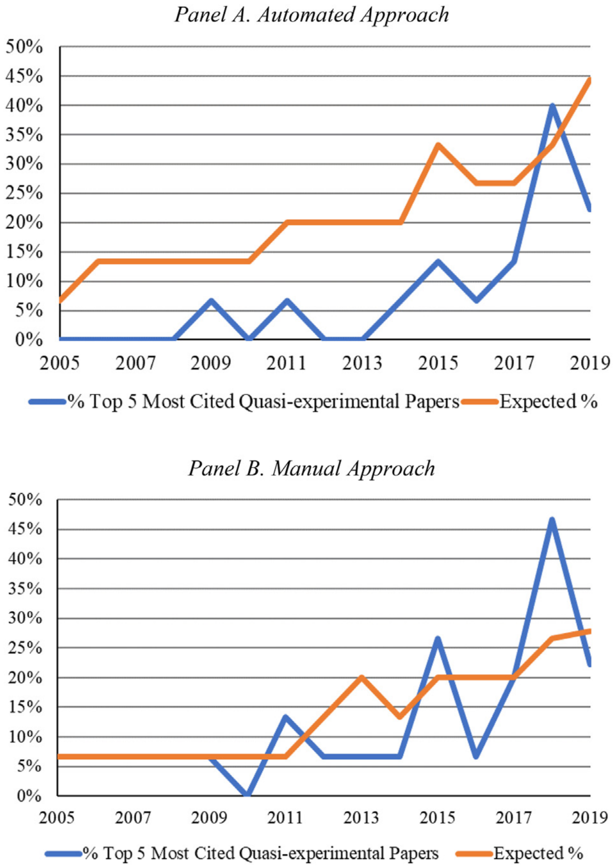
Fig. 3. Summary of Settings. Panel A of this figure illustrates the settings used in the sample of papers in Fig. 2. Among the sample of papers using "Regulatory" settings in Panel A, Panel B illustrates the types of regulations that were used.

Our assessment is that the majority of papers using quasi-experimental methods fall into the bottom-right quadrant, and that much of their contribution lies in the novelty or cleverness of the setting, rather than the novelty of the underlying conceptual research question. Indeed, several studies in our survey are explicit that their contribution is to use quasi-experimental methods to upgrade inferences from "association" to "causal."

For example, consider Boland and Godsell (2020), which tests the political cost hypothesis of Watts and Zimmerman (1978) by relating discretionary accruals to political costs using local soldier fatalities as a source of as-if-random variation in the threat of political costs, and Huang et al. (2017), which tests the proprietary cost hypothesis of Verrecchia (1983) using U.S. import tariffs as an exogenous shock to proprietary costs. Because the political cost and proprietary cost hypotheses are mature theories that each span four decades, recent papers that examine these topics tend to couch their contribution primarily in the novelty of their settings—rather than the novelty of the underlying theory being tested.

We make no judgments about whether these trends in the literature are "good" or "bad," or about the contributions of any specific paper, but rather seek to point out methodological trends in the literature and examples of how these trends manifest in recent research. To conclude this section, our survey suggests that the literature is gravitating from the upper-left to the bottom-right quadrant of Fig. 5.[11] It is useful to reflect on why this might be the case. Does this indicate that the field is "maturing"? Does this reflect a stagnation in the development of new theories? Does this reflect a lack of important, unaddressed research questions in accounting? Does it reflect an increased interest in drawing causal inferences?

[11] Other observers have noted this trend in the economics and finance literatures (Deaton, 2009; Heckman and Urzua, 2010; Bowen et al., 2017; Panhans and Singleton, 2017).



Panel A. Automated Approach

Panel B. Manual Approach

Fig. 4. Citations Analysis. This figure plots the percentage of the top five most-cited empirical papers according to Clarivate Analytics Journal Citation Reports among our sample of papers. The blue line in Panel A (B) presents results using the automated (manual) approach to classifying papers. The orange line plots an expected percentage of highly cited papers using quasi-experimental methods, based on the proportion of all papers using quasi-experimental methods every year (e.g., if 25% of empirical papers in 2016 are classified as quasi-experimental, then we expect 4 highly cited papers in 2016 to be classified as such (25% x 3 journals x 5 papers). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3. Quasi-experiments and the path toward causal inference

Our survey of the accounting literature reveals an increasing interest in drawing causal inferences using quasi-experimental methods. In this section, we discuss two key ingredients to causal inference: empirical methods and theoretical assumptions.

In Section 3.1, we begin by discussing the most common estimation technique in accounting: linear regression with panel data. We discuss the assumptions under which simple linear regression does and does not allow for causal inferences and the common methods researchers have used to alleviate concerns about violations of these assumptions, e.g., correlated omitted variables.

In Section 3.2, we discuss the most popular quasi-experimental method used to draw causal inferences: difference-in-differences.[12] Our discussion makes transparent the similarity in assumptions between difference-in-differences designs and standard linear regression using panel data. Our discussion of these methods and their assumptions is intended as an overview, and to supplement—but not substitute for—material in standard econometrics texts.

In Section 3.3, we discuss the critical importance of theory (whether it be informal intuition or a formal economic model) for drawing causal inferences. Specifically, we focus on two subtle, but important, roles of theory. First, all empirical methods estimate correlations, and it is the researcher's assumptions that allow us to interpret these correlations in a causal manner. In the absence of theory, even the correlations estimated from quasi-experimental methods do not have any specific meaning. Second, theory is required to generalize from the researcher's sample to a broader population of interest. In the absence of a compelling theory, there is no basis on which to generalize inferences beyond the specific setting being studied.

[12] Fig. 2 indicates over 75% of accounting papers drawing causal inferences employ this design.

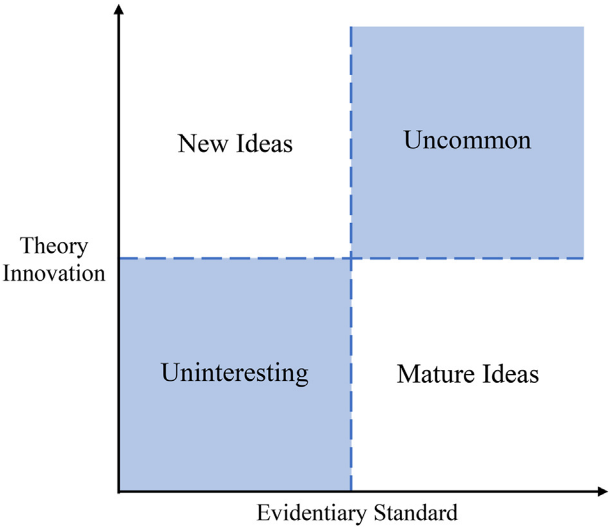


Fig. 5. Tradeoff Between Novel Ideas and the Evidentiary Standard. This figure illustrates the tradeoff between new ideas (i.e., theory innovation) and the evidentiary standard. Novel ideas for which the theory is not yet well developed typically only allow for tests where the evidentiary standard is relatively low (i.e., upper-left quadrant). In contrast, ideas that have well-developed theories allow for much stronger tests where the evidentiary standard is relatively high (i.e., lower-right quadrant). We thank Ed deHaan for suggesting this figure.

3.1. Linear regression and the omitted variable threat

3.1.1. Brief review of OLS assumptions

A common feature of the vast majority of the papers from our survey is the use of linear regression (i.e., OLS) on panel data to estimate correlations. These regressions take the general form:

$$y_{i,t} = \alpha + \beta x_{i,t} + \varepsilon_{i,t} \tag{1}$$

where *i* indexes cross-sectional units (typically a firm) and *t* indexes time-series data (typically a year or quarter). Henceforth, for expositional purposes, we assume that the cross-sectional unit is a firm and the time-series unit is a year. Many studies interpret estimates of β as the "correlation," "association," or "relation" between *x* and *y*. As all common econometrics textbooks point out (e.g., Stock and Watson, 2003; Angrist and Pischke, 2008), if the standard OLS assumptions hold, then OLS can be used to draw causal inferences, and β can be interpreted as a causal effect. The OLS assumptions are:

1. The conditional distribution of $\varepsilon_{i,t}$, given $x_{i,t}$, has a mean of zero.
2. The pair $(x_{i,t}, y_{i,t})$ are independently and identically distributed across observations.
3. $x_{i,t}$ and $\varepsilon_{i,t}$ have non-zero and finite fourth moments.

For researchers interested in drawing causal inferences, identifying settings and circumstances in which these assumptions hold is of paramount importance. In the interest of parsimony, we limit the scope of our discussion to assumption #1, which is violated when the regression omits a variable that is correlated with both x_i and y_i. As Stock and Watson (2003) discuss, common concerns about "endogeneity" (e.g., a specific omitted variable, reverse causality, and simultaneity) can all be framed as various forms of correlated omitted variables.[13] Thus, convincingly dealing with omitted variable bias is of critical importance to empirical researchers interested in causal inference. As Angrist and Pischke (2008, p.39) articulate, in the absence of omitted variable bias, regression coefficients can be interpreted in a causal manner, but "[t]he big question [...] is what these variables are, or should be."

3.1.2. Brief review of omitted variable bias

Before turning to the methods researchers use to rule out this bias, and the potential role of quasi-experimental methods in mitigating this bias, it is important to first understand the source and nature of the bias. Suppose that the true data generating process is given by the following equation:

[13] See Barth and Clinch (2009) for a discussion of this point in the context of valuation models.

$$y_{i,t} = \alpha + \beta x_{i,t} + \varphi z_{i,t} + \varepsilon_{i,t} \tag{2}$$

but that we estimate equation (1), which omits $z_{i,t}$ from the regression. As a result, the expected coefficient on $x_{i,t}$ is given by:

$$E[\widehat{\beta}] = \beta + \frac{\varphi cov(z_{i,t}, x_{i,t})}{var(x_{i,t})} \tag{3}$$

where the latter term represents the omitted variable bias. This formula makes clear that, to bias the OLS estimator, the omitted variable $z_{i,t}$ must vary with *both* the outcome variable $y_{i,t}$ (i.e., $\varphi \neq 0$) and the independent variable of interest $x_{i,t}$ (i.e., $cov(z_{i,t}, x_{i,t}) \neq 0$). If either condition does not hold, there is no omitted variable bias.

3.1.3. *Common approaches to ruling out omitted variable bias*

Our survey reveals that studies in the accounting literature tend to take three common approaches to ruling out omitted variable bias: specific identification, fixed effects, and cross-sectional interactions. We discuss these approaches in turn.

3.1.3.1. Specific identification. The first approach, which we refer to as "specific identification," uses theory (whether it be informal intuition or a formal economic model) to develop plausible economic alternative explanations for the phenomenon at hand and uses these alternative explanations to guide the search for and measurement of potential omitted correlated variables to include in the regression. For example, one might intuit that "corporate governance" is a potential omitted variable and include a noisy measure thereof in the regression (e.g., staggered boards). The benefit of this approach is that it explicitly identifies *falsifiable* alternative economic explanations that can enable the researcher to design more powerful tests to rule out these alternatives. The drawback of this approach is that the researcher may not consider an exhaustive set of alternatives (e.g., out of ten omitted correlated variables, the researcher only identifies two). For example, Christensen et al. (2013) identify changes in enforcement laws as a confound to estimating the causal effect of IFRS. Having identified a specific confound allows Christensen et al. (2013) to design more powerful and better specified tests of the effect of IFRS and its interaction with changes in securities laws.

3.1.3.2. Fixed effects. The second approach seeks to rule out omitted variables without identifying a specific correlated omitted variable. This approach recognizes the existence of unknown omitted variables that the researcher may not be able to specifically identify and seeks to rule out these omitted variables—collectively—without needing to individually identify them, or for that matter identify any economic alternative explanation for the phenomenon being studied. A common way of implementing this approach is to include a vector of fixed effects along a particular dimension that would absorb *any* known or unknown omitted variables that vary exclusively along that dimension. For example, suppose that the data generating process is given by equation (2), and the researcher estimates equation (1) after including a vector of firm fixed effects. The goal of this estimation is to use the fixed effects to absorb as much problematic variation in the unknown omitted variable $z_{i,t}$ as possible.

The key benefit of this approach is that it can remove significant problematic variation from omitted variables without the researcher needing to specify what those omitted variables are. When firm fixed effects are included, only omitted variables containing within-firm variation are a threat. However, this approach is not a panacea, and there are many settings where including firm fixed effects would exacerbate the omitted variable bias (see Section 4.3 for details).

Fixed effects can also allow researchers to examine the sources of variation driving the underlying relations. In many cases, understanding the sources of the variation can help discriminate among alternative explanations for a given phenomenon. In this regard, it is instructive to consider Bianchi et al. (2021, BMMP). BMMP examine the relation between a firm's connection to the Italian mafia and financial statement characteristics that might be indicative of money laundering. Such relations can manifest from either (i) mafia connections causing these financial statement outcomes, or (ii) the mafia merely selecting firms with certain characteristics (making them "an offer they cannot refuse"). BMMP use firm fixed effects to distinguish between these two explanations—i.e., "treatment" versus "selection." BMMP find no evidence of a relation between mafia connections and their financial outcomes in the presence of firm effects. This null result is informative: despite not being able to establish causal inferences, it suggests a selection explanation for the phenomenon, and provides authorities with sufficient information to develop a set of crude metrics potentially indicative of firms targeted by the mafia.

3.1.3.3. Cross-sectional interactions. A third approach uses theory (whether it be informal intuition or a formal economic model) to identify a setting or subsample within the data in which the marginal effect of $x_{i,t}$ is conjectured to be particularly pronounced, but in which the effect of the omitted correlated variable is conjectured to be similar. In this case, the researcher can estimate a "cross-sectional interaction" to effectively difference out the effect of the omitted variable. To develop this formally, consider the following two data generating processes for a sample partitioned based on the indicator variable $D_{i,t}$:

$$y_{i,t} = \alpha_A + \beta_A x_{i,t} + \varphi_A z_{i,t} + \varepsilon_{i,t} \text{ where } D_{i,t} = 1$$

$$y_{i,t} = \alpha_B + \beta_B x_{i,t} + \varphi_B z_{i,t} + \varepsilon_{i,t} \text{ where } D_{i,t} = 0 \tag{4}$$

For ease of exposition, suppose than when $D_{i,t} = 0$, $x_{i,t}$ is unrelated to $y_{i,t}$ (i.e., $\beta_B = 0$). That is, the partitioning variable indicates a setting in which $x_{i,t}$ has no causal effect on $y_{i,t}$.

If one were to estimate a regression of $y_{i,t}$ on $x_{i,t}$ within each sample, the expected coefficients on $x_{i,t}$ would be given by:

$$E[\widehat{\beta_A}] = \beta_A + \frac{\varphi_A cov(z_{i,t}, x_{i,t}|D=1)}{var(x_{i,t}|D=1)} \tag{5}$$

and

$$E[\widehat{\beta_B}] = \beta_B + \frac{\varphi_B\, cov(z_{i,t}, x_{i,t}|D=0)}{var(x_{i,t}|D=0)}\ . \tag{6}$$

By virtue of $\beta_B = 0$, the latter estimation simply reduces to the omitted variable bias:

$$E[\widehat{\beta_B}] = \frac{\varphi_B\, cov(z_{i,t}, x_{i,t}|D=0)}{var(x_{i,t}|D=0)}\ . \tag{7}$$

In this regard, the subsample in which $D_{i,t} = 0$ can be viewed as a placebo. In truth, there is no effect of $x_{i,t}$ on $y_{i,t}$ when $D_{i,t} = 0$, so any estimated effect must be the result of the omitted variable bias. If we assume that the effect of bias does not vary with the sample partition, $D_{i,t}$, then:

$$\frac{\varphi_A\, cov(z,x|D=1)}{var(x|D=1)} = \frac{\varphi_B\, cov(z,x|D=0)}{var(x|D=0)} \tag{8}$$

and we can difference out the effect of the omitted correlated variable by taking the difference in coefficients across the two sample partitions:

$$E[\widehat{\beta}_A] - E[\widehat{\beta}_B] = \beta_A. \tag{9}$$

The difference in the coefficients is an unbiased estimate of the causal effect when $D_{i,t} = 1$.

In practice, this methodology for mitigating omitted variable bias is typically implemented in one of two ways, either (i) estimating separate regressions in each sample partition and testing for differences in the estimated coefficients across the partitions, or (ii) stacking the sample partitions and estimating a fully interacted model:

$$\widehat{y}_{i,t} = \widehat{\alpha}_1 + \widehat{\alpha}_2 D_{i,t} + \widehat{\beta}_1 x_{i,t} + \widehat{\beta}_2 (D_{i,t} * x_{i,t}). \tag{10}$$

Here, the expected coefficient on the interaction term, $E[\widehat{\beta}_2]$, is equivalent to $E[\widehat{\beta}_A] - E[\widehat{\beta}_B]$. Thus, under the maintained assumption that the marginal effect of the omitted variable does not vary with $D_{i,t}$, the expected coefficient on the interaction term is free of any omitted variable bias. This illustrates why we often see authors write language similar to “to explain these results, an omitted correlated variable would not only need to be related to both $x_{i,t}$ and $y_{i,t}$, but this relationship would also need to vary with $D_{i,t}$.” For example, Guay et al. (2016) study the relation between financial statement complexity and voluntary disclosure. To help rule out the possibility that the results are attributable to an omitted variable that affects financial statement complexity and voluntary disclosure (e.g., time trends), the authors conduct several cross-sectional analyses. Guay et al. (2016) predict and find that the relation between financial statement complexity and voluntary disclosure is stronger in firms where managers have incentives to be transparent (i.e., when their firms have lower liquidity and greater external monitoring), and is weaker in firms when managers have incentives to obfuscate performance (i.e., in the presence of earnings losses and greater earnings management). Generating and testing multiple predictions makes it less likely that the collective results are attributable to alternative explanations (relative to a study that offers and tests a single prediction).

There are two key takeaways from this section. First, each method makes different theoretical assumptions about the data generating process—a process that is, in truth, unknown. Because the data generating process is unknown, no one method of dealing with omitted variables is a panacea. Ex ante, without knowing the specific institutional setting or sample, it is impossible to say with any degree of certainty that one approach unambiguously dominates the others.

Second, researchers’ ability to draw causal inferences depends on the methodological assumptions they are willing to make, and often these assumptions are implicit. Fortunately, empirical methods are not mutually exclusive, and papers often employ all the methods described above to triangulate inferences across a variety of different tests. Finding consistent results across multiple approaches that each make different assumptions about the data generating process strengthens the credibility of the inferences obtained from any single approach.

3.2. Quasi-experiments as a potential solution

The preceding section makes clear that, if the OLS assumptions hold, one can use OLS on panel data to estimate a causal effect. In practice, however, researchers are often unwilling to assume the absence of correlated omitted variables. Section 3.1.3 discusses three designs researchers commonly use to rule out such variables. Notably, the third approach—cross-sectional tests—seeks to identify a sample or setting in which the omitted variable bias is present but the underlying causal effect is less pronounced, and use this setting as a placebo. A similar approach increasingly used in contemporary accounting research is to seek a setting in which the underlying causal effect is present, but the omitted variable bias is absent.

The overarching objective of this approach—which we refer to as "quasi-experimental"—is to find a setting or sample that replicates the experimental ideal of "random assignment" in the variable of interest. Suppose we could conduct a laboratory experiment to randomly assign the value of our variable of interest $x_{i,t}$ to each unit of observation, i.e., each $\{i,t\}$ pair. By virtue of random assignment of $x_{i,t}$ values to each observation, by construction, there would be no correlated omitted variables. We could then regress our desired outcome variable of interest on the randomly assigned $x_{i,t}$ value to recover the causal effect of $x_{i,t}$ on $y_{i,t}$ (e.g., Angrist and Pischke, 2008).

For example, suppose we randomly assigned patients in our laboratory to treatment and control groups, and administer the treatment only to the patients in the treatment group. By virtue of random assignment (i.e., assuming there is no systematic difference between the treatment and control groups prior to administering the treatment), we would recover the causal effect of the treatment by simply measuring the difference in mean outcome between the treatment and control groups. This simple difference in means can be operationalized using an OLS regression of the outcome, $y_{i,t}$, on an indicator variable for whether the observation belongs to the treatment group.

In practice, however, the experimental ideal of random assignment is rarely observed outside of a laboratory. The *quasi*-experimental approach is to find a setting—in nature—that approximates this experimental ideal, and generates "as-if" random assignment of the $x_{i,t}$ to our units of observation. If we find such a setting, we can directly estimate the causal effect of $x_{i,t}$ on our outcome of interest. The concern with this approach is that the observations may not quite be randomly assigned to treatment and control groups (e.g., the adoption of IFRS was not random). This is where the "quasi" portion of the term "quasi-experiment" comes in. In particular, we might want to check whether, in fact, in our chosen setting, observations are randomly assigned to treatment and control groups, and if not, control for the non-random aspect of the assignment.

To do so, the literature has embraced a method known as difference-in-differences (DiD). Technically, the DiD method is just an OLS regression estimated on panel data. The reason this technique gets its own name is because it implies a particular regression specification, rather than a general form. However, it is important to point out that DiD estimators inherit *all* of the OLS assumptions. The reason DiD estimators do not make fewer assumptions than OLS is because DiD is implemented using OLS. There are many different flavors of DiD specifications. In each case, the idea is to use panel data to approximate an experiment in which observations are assigned to treatment and control groups, but also recognizing that the assignment may not be random in the strict sense of the term. We briefly cover three types of DiDs.

3.2.1. The classic DiD

The first specification is the classic DiD:

$$y_{i,t} = \alpha_1 + \alpha_2 D_t + \beta_1 x_i + \beta_2 (D_t * x_i) + \varepsilon_{i,t} \tag{11}$$

In this specification, x_i is an indicator variable equal to one if firm *i* was treated, and D_t is an indicator equal to one if year *t* occurred after the treatment. The key to the classic DiD design is that all observations receive treatment at the same point in time, and so the period after treatment, or "post-period," is the same for all observations. This design can best be represented in the context of a two-by-two grid, where each cell represents a conditional expectation (or conditional mean) of $y_{i,t}$:

		$x_i = 0$	$x_i = 1$	*Difference*
		Control	*Treatment*	
$D_t = 0$	*Pre*	α_1	$\alpha_1 + \beta_1$	β_1
$D_t = 1$	*Post*	$\alpha_1 + \alpha_2$	$\alpha_1 + \alpha_2 + \beta_1 + \beta_2$	$\beta_1 + \beta_2$

This diagram makes clear that β_1 captures the difference in the outcome between the two groups prior to the treatment. If the observations were truly randomly assigned to treatment and control groups (i.e., their assignment is not conditional on underlying characteristics of the observations), then, in expectation, this difference should be zero. In practice, however, assignment is rarely random and it is not uncommon to observe a difference between the two groups in the pre-period.

Recognizing that differences between the two groups (may) exist in the pre-period—similar to the cross-sectional interactions described earlier—the pre-period sample is used as a placebo to de-bias and remove the differences in outcome that are attributable to non-random assignment. That is, the focus of the DiD design is not the difference between treatment and control groups (or the marginal effect of x_i) but rather how that difference *changes* after the treatment is administrated (i.e., the *difference* in the marginal effect of x_i). This is where the term "difference-in-differences" comes from: the focus is on the β_2 term.

At this point, the parallels between the DiD design and the cross-sectional interaction design should be apparent. In fact, the cross-sectional regression design discussed earlier is actually a generalization of the difference-in-differences design. To see this, note that the cross-sectional design did not specify the form of the variable of interest $x_{i,t}$ or the partitioning variable $D_{i,t}$. In contrast, the DiD design specifies: (1) that the partitioning variable $D_{i,t}$ is based on whether year t occurs after a specific time threshold (e.g., post-2003), which is why one drops the i subscript from $D_{i,t}$, and (2) replaces the general $x_{i,t}$ variable with an indicator for whether the firm was treated, i.e., it drops the t subscript from $x_{i,t}$. In this regard, the DiD design *is a specific type of cross-sectional interaction design*: the coefficient of interest represents the difference in the marginal effect of $x_{i,t}$ between the two sample partitions.[14]

Similar to the cross-sectional regression, the primary threat to causal inference in the DiD design is not an omitted variable that varies with $x_{i,t}$. Under the assumption that the effect of the omitted variable does not vary with the sample partition (i.e., does not vary with D_t), the difference between the treatment and control sample when $D_t = 0$ will capture the related omitted variable bias (i.e., β_1). Instead, as in the cross-sectional approach, the primary threat to inferences in the DiD design is a correlated omitted variable whose effect varies with D_t. Because in the DiD specification, the partitioning variable D_t is based on units of time, this is known as the "parallel trends assumption." In other words, to bias the DiD estimator, the correlated omitted variable must not only vary with treatment and control groups (i.e., with $x_{i,t}$), but its effect must also vary over time (i.e., with D_t). If its effect does not vary over time, then it will simply create a constant difference between treatment and control groups in each period, captured in β_1, and the time-series trends in the outcome variable between treatment and control groups will be parallel. See Roberts and Whited (2013) for more exposition on this point.

To illustrate studies that use a DiD to draw causal inferences, Fang et al. (2016), Hope et al. (2017), Li and Zhang (2015), and Kecskés et al. (2013) use Regulation SHO as a setting that aims to provide random assignment of short selling constraints between treatment and control groups. Under the assumption that the short selling constraints were randomly assigned by the regulators—the operating assumption of these papers—these papers use a classic DiD to estimate the causal effect of short selling constraints on earnings management, audit fees, management forecasts, and credit ratings. Under the assumption that short selling constraints were randomly assigned by regulators, controls for correlated omitted variables are not necessary.

3.2.2. Extensions

We next briefly discuss the two most popular extensions of the classic DiD design in the accounting literature. The first extension is the "generalized" DiD design:

$$y_{i,t} = \theta(D_t * x_i) + Firm_i + Year_t \tag{12}$$

The distinguishing feature of this design, relative to the classic design, is that it includes both firm fixed effects ($Firm_i$) and year fixed effects ($Year_t$), which control for any fixed differences between the treatment and control groups (firm fixed effects absorb the *Treated* main effect) and control for any common time trends (year fixed effects absorb the D_t main effect). See Hansen (2007) and Angrist and Pischke (2008) for details on this design. A recent example is Christensen et al. (2017), who examine the effect of the disclosure of mine safety records in financial statements on mine safety practices. Christensen et al. (2017) use a mine-year panel and include mine—rather than firm as in our Equation (12)—and year fixed effects in a generalized DiD design. This controls for any time-invariant fixed differences between treatment and control groups—i.e., mines the were and were not affected by the disclosure regulation, respectively—as well as for any common time trends that have a similar effect on treatment and control groups.

At this point it is useful to highlight that the dependent variable in a DiD design ($y_{i,t}$) need not be continuous. Even in this case, there are compelling reasons to estimate (12) using a linear model (i.e., a linear probability model in the context of a binary dependent variable, see also Angrist and Pischke, 2008). First, Greene (2004) and Arellano and Hahn (2007) raise concerns about bias and consistency of probit and logit models with high-dimensional fixed effects. Given such effects are a key feature of these more advanced DiD designs, linear design models may be more appropriate. Second, Ai and Norton (2003) show that in probit and logit models (and non-linear models more generally), the coefficient on the interaction term does *not* represent the marginal effect ($\frac{\partial^2 y}{\partial x \partial D}$). Consequently, θ will not recover the DiD estimator when the function is a probit or logit specification.

The second extension is the "staggered adoption" DiD design:

$$y_{i,t} = \theta(D_{i,t} * x_i) + Firm_i + Year_t \tag{13}$$

The distinguishing feature of this design, relative to the generalized design, is that each firm (or *i*th unit) has its own value of $D_{i,t}$. That is, whereas previously D_t took a common value for all observations in a given year, now $D_{i,t}$ varies across firms. For

[14] Although the classic DiD is easiest to understand when all the independent variables are binary, there is also a generalization in which the treatment variable is continuous. Atanasov and Black (2016) refer to this method as "DiD-Continuous." This makes the specification even closer to the general cross-sectional regression discussed above.

example, in the classic and generalized designs, if the treatment was administered in 2003, $D_t = 1$ for all observations after 2003. In contrast, the staggered adoption design allows each unit to receive treatment at different points in time. For example, one observation might be treated in 2003. For this observation, $D_t = 1$ for all years after 2003. Another observation might be treated in 2010. For this observation, $D_t = 1$ for all years after 2010, hence the inclusion of the *i*-subscript, $D_{i,t}$. Conversely, if the firm is never treated, the value of $D_{i,t}$ is irrelevant, because $x = 0$. For this reason, we often see the design expressed parsimoniously without the interaction, as follows:

$$y_{i,t} = \theta(Adopt_{i,t}) + Firm_i + Year_t \tag{14}$$

where $Adopt_{i,t} = 0$ for both untreated observations and treated observations prior to the treatment. A recent example of a staggered DiD is Barrios (2022), who examines the effect of occupational licensing on the quality of Certified Public Accountants (CPAs). Barrios (2022) exploits the staggered adoption across states of the 150-hour rule, which increases the educational requirements for a CPA license. In this case, the staggered adoption of the rule allows for comparisons between treated and untreated observations using variation in the time series (i.e., before and after the rule) and the cross section of states (i.e., states that have not yet adopted the rule in a given year).

It is important to note that commingling these two groups in the $Adopt_{i,t} = 0$ group in Equation (14) has recently been the topic of controversy. Goodman-Bacon (2021) points out circumstances under which this commingling can lead to bias. Baker et al. (2021, p.2) describe the intuition for this bias as follows: "when treatment effects can evolve over time—staggered DiD estimates can obtain the opposite sign [as the true causal effect] … The intuition is that in the standard staggered DiD approach, already-treated units can act as effective controls, and changes in their outcomes over time are subtracted from the changes of later-treated units (the treated)." This takes us to the cutting edge of methodological research in accounting: the Goodman-Bacon critique is relatively recent and, consequently, no paper in our survey applied these insights. We refer interested readers to Barrios (2021) for a discussion of Goodman-Bacon in the context of accounting research and Baker et al. (2021) for applications in finance.

There are two key takeaways from this section. First, the classic DiD estimator is a particular implementation of the more general cross-sectional interaction approach discussed in Section 3.1.3. Both the cross-sectional interaction design and the DiD design share the same common estimation techniques and estimation assumptions. In both cases, the threat to causal inference is an omitted variable that is correlated with both the outcome, the independent variable of interest, and the partitioning variable, *D*. Thus, ex ante, without specifying a setting or sample, DiD designs are no more or less robust than cross-sectional interaction designs.

Second, the ability to draw meaningful causal inferences stems from the assumptions that the researcher is willing to make. For example, one might have a more compelling reason to believe that the assumptions underlying the cross-sectional approach hold if the partitioning variable is a unit of time (e.g., post 2007), rather than a firm characteristic (e.g., high analyst coverage). In this case, a DiD would be the appropriate method. Alternatively, one might have reason to believe that the assumptions hold if the partitioning variable is a firm characteristic (e.g., market competition) rather than a unit of time, in which case a cross-sectional approach is appropriate. So, rather than framing the debate about which methods do and do not allow for causal inference, the debate should be framed around whether the implicit assumptions embedded in a particular method are valid in a particular setting. Here, we emphasize that causal inferences are not driven by the method *per se*—or even the assumptions those methods make—but whether the assumptions comport with the particular setting the researcher is analyzing.

3.3. *The importance of theory for causal inference*

In this section, we discuss the importance of theory for drawing causal inferences. Throughout our discussion, we define theory according to The Oxford Dictionary: "a supposition or a *system of ideas* intended to explain something, especially one based on *general principles independent of the thing to be explained*" [emphasis added]. This definition makes clear that theory is more general than the inference one draws from any single empirical test (which is conditional on the choice of setting, sample, and methods), and makes clear that theory is much broader than formal analytical models.

3.3.1. *The importance of theory for interpreting empirical facts*

A sound, well-defined theory is necessary to draw causal inferences from observational data. Theory provides a framework for making predictions and interpreting estimated correlations. In the absence of a theory, correlations do not carry any economic meaning. Correlations may have statistical meaning, in the sense that two variables might co-move, but interpreting that co-movement requires a theory. Indeed, Heckman (2005) suggests that the first two tasks that confront empirical researchers seeking to draw causal inferences are: (i) use theory to describe a hypothetical world, and (ii) identify the causal channel in that hypothetical world (see also Heckman and Vytlacil, 2007).

Our prior discussion illustrates that the methods used in our survey are all estimating correlations and it is our willingness to make particular underlying assumptions that allow us to interpret these correlations in a causal manner. For example, each of the methods discussed earlier assume there is no omitted variable that is correlated with both the dependent variable and the independent variable of interest. Theory tells us the extent to which we might be concerned about such a variable in a specific setting or a specific research design. For example, in a panel regression of firms' illiquidity on analyst coverage, theory might tell us that "corporate governance" is an omitted variable. As a result, we might be skeptical of empirical relations that do not explicitly control for this variable. Theory might tell us that if we were to look at changes in analyst coverage around the September 11th terrorist attacks, then omitted variables related to corporate governance are less of a concern because such variables do not vary around the terrorist attacks (Kelly and Ljungvist, 2012).

Theory can also inform us about the endogenous nature of the theoretical constructs being studied. For example, theory might tell us that mandatory and voluntary disclosure decisions are interdependent (e.g., Beyer et al., 2010; Heinle et al., 2020). As a consequence, a regression of voluntary disclosure characteristics on mandatory disclosure characteristics should not be interpreted in a causal manner, but rather as consistent with a theory in which managers who choose a particular characteristic of mandatory disclosure also choose a particular type of voluntary disclosure. Whether researchers realize it or not, whenever they justify a research design choice—or suggest a specific setting provides as-if random variation in a particular outcome—they are implicitly invoking theory.

For example, as discussed in Bertomeu et al. (2016), when researchers measure systematic risk, they often employ estimates of β from the Capital Asset Pricing Model—an empirical measure of systematic risk derived from formal theory; when researchers measure information asymmetry, they often employ estimates of Kyle's λ or the probability of informed trade (PIN)—empirical measures of information asymmetry between market participants derived from formal theory (Kyle, 1985). Theory provides the assumptions that guide how we interpret the relations in the data—for example, how we interpret the covariance between a firm's returns and the market return.

Indeed, the scientific process is premised on the idea that the researcher starts with a theory, tests the theory, refines the theory in light of the test results, and that the refined theory leads to new and more precise tests. This process highlights the integral part of theory to causal inferences—and scientific inquiry more generally.

Ultimately, as the scientific process in Fig. 6 makes clear, the process of drawing causal inferences is not about random, or as-if-random variation *per se*, but about ruling out alternative explanations (e.g., Kahn and Whited, 2018). Consequently, not only is theory valuable for causal inference, but more precise theory is even more valuable. As the precision of the underlying theory increases, researchers can better articulate (and test) multiple predictions and rule out alternative explanations, which increases the credibility of the resulting inferences.

3.3.2. An alternative view

An alternative view is that theory is not necessary to estimate causal effects. Proponents of this view point to laboratory experiments, field experiments, and standard "A/B" testing procedures with random assignment as settings where theory is not necessary for the estimation of causal effects. In other words, theory is not necessary in the presence of true random assignment. In these settings, if the experiment is well-specified, the researcher is inducing random variation, and observations are randomly assigned to treatment and control groups by construction. Consequently, there is no correlated omitted

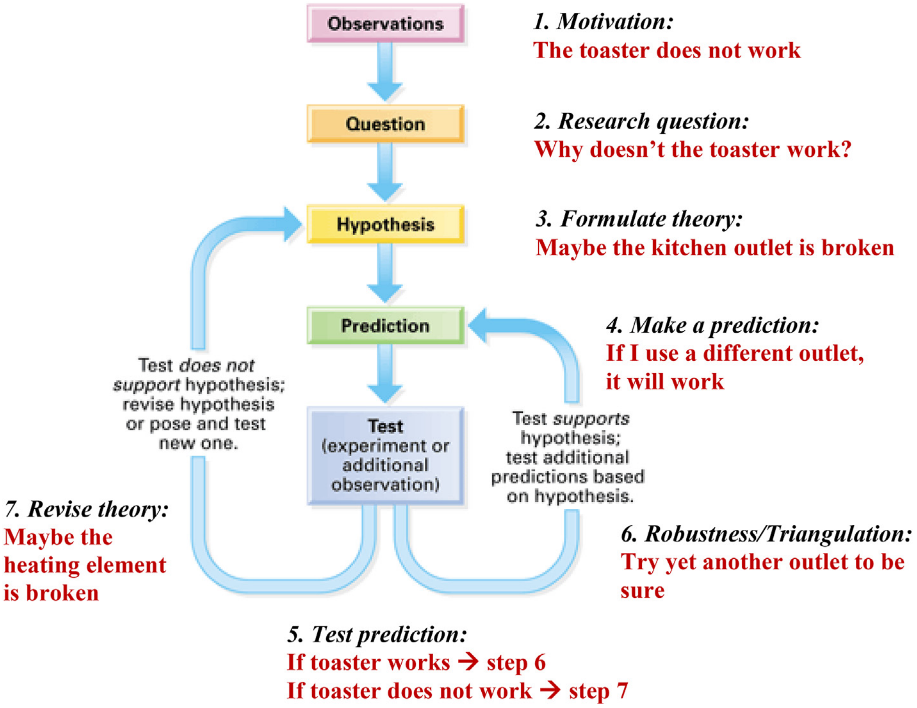

Observations
1. Motivation:
The toaster does not work
Question
2. Research question:
Why doesn't the toaster work?
Hypothesis
3. Formulate theory:
Maybe the kitchen outlet is broken
Prediction
4. Make a prediction:
If I use a different outlet, it will work
Test *does not support* hypothesis; revise hypothesis or pose and test new one.
Test (experiment or additional observation)
Test *supports* hypothesis; test additional predictions based on hypothesis.
7. Revise theory:
Maybe the heating element is broken
6. Robustness/Triangulation:
Try yet another outlet to be sure
5. Test prediction:
If toaster works → step 6
If toaster does not work → step 7

Fig. 6. Scientific Process. This figure adapts the scientific process as illustrated in Figs. 2–10 from Campbell et al. (2009).

variable and the experiments are presumably replicable under identical conditions.[15] We agree that settings where the researcher generates random variation are particularly promising in estimating causal effects and drawing causal inferences. To the extent that the literature is increasingly emphasizing the importance of random assignment, we would expect to see—and encourage—a renaissance in laboratory and field experiments in accounting research.[16]

When considering this alternative view however, it is important to keep in mind two points. First, any test of a causal effect—either using experimental data or archival data—implicitly invokes theory when choosing measures and research design. Second, the nature of the data in common archival accounting settings is unlike those discussed above. Rarely do researchers have data with truly random assignment to treatment and control groups, and instead must rely on settings that provide as-if random assignment (see Fig. 3) and econometric methods that rely on certain untestable theoretical assumptions. Consequently, as a practical matter, when using archival data, theory is required to estimate causal effects, and the more precise the theory the more credible the resulting causal inferences.

3.3.3. *Using theory to separate spurious relations from causal relations*

The danger in ignoring theory when drawing inferences is that the researcher will be unable to separate spurious inferences—i.e., those based on random correlations with no particular meaning—from causal inferences. As a starting point to understanding the dangers of ignoring theory, it is important to understand that correlations by themselves do not reveal truth. Spurious correlations exist in the data. For example, Vigen (2015) shows a high correlation between (i) suicides by hanging, strangulation, and suffocation and U.S. spending on science; (ii) the annual number of people who drowned and number of films in which Nicolas Cage appeared; (iii) the age of the Miss America beauty pageant winner and the number of people murdered by hot liquid, and (iv) the number of doctorates in civil engineering and the consumption of mozzarella cheese.

Theory (i.e., economic intuition) tells us that these correlations are patently absurd, and thus we should exercise caution in interpreting them in a causal manner. We might be reluctant to believe the existence of a causal relation because we have no compelling theory that links, for example, Miss America's age with murders by hot liquids. Alternatively, we might ignore the lack of underlying theoretical foundations and be skeptical because the data are bivariate correlations, and we might be willing to interpret these empirical facts in a causal manner only if they were the result of more sophisticated econometric techniques.

However, spurious correlations are possible even with more sophisticated econometric techniques (e.g., Brodeur et al., 2020)—suggesting common robustness tests often reported in the literature are often not sufficient to rule out spurious relations. For example, a methodologically rigorous study by Brown et al. (2015) purports to show that the staggered introduction of state laws that banned texting while driving reduced individuals' information search activity and reduced market liquidity. Despite using state-of-the-art econometric methods (e.g., a staggered DiD design with high-dimensional fixed effects) and triangulating results across a number of different specifications, a follow-up paper by one of the authors (White and Webb, 2021), reports that similar results are obtained on a random sample of dates 25% of the time—calling into question the inferences of Brown et al. (2015).

To illustrate spurious correlations are possible even with more sophisticated econometric techniques, we extend the analysis in Vigen (2015) to accounting research and compile a dataset of over two dozen common measures of voluntary disclosure and earnings management (e.g., discretionary accrual models, restatements, Accounting and Auditing Enforcement Releases, management forecasts, voluntary Form 8-K filings, and press releases). We provide the corresponding code and data in the Internet Appendix and encourage readers to explore this dataset. Having compiled a set of outcome variables commonly used in accounting research, we then use a staggered adoption regression design that includes *Firm*, *Industry* x *State*, and *Industry* x *Year* fixed effects to correlate these data with various state-year level variables. Does the use of a state-of-the-art method with high-dimensional fixed effects ensure there are no spurious relations?

Using this design, Table 1 suggests that the adoption of state laws that restrict workplace smoking (Panel A), and the adoption of state laws that restrict access to firearms (Panel B), are both highly negatively correlated with voluntary disclosure (measured using the number of management earnings forecasts during the year). Notice that we used the term "negatively correlated" to describe the empirical relations we estimated. Can these negative correlations always be interpreted in a causal manner? Do we feel comfortable concluding that workplace smoking laws and gun waiting laws both cause reductions in voluntary disclosure?

Given that the staggered adoption method employed in Table 1 is used by prior research to draw causal inferences (including our own papers), it is useful to consider why we may or may not feel comfortable drawing causal inferences from these particular results. First, we have not advanced a theory about why these laws would be related to voluntary disclosure, nor have we advanced a theory about why management earnings forecasts are an appropriate measure of voluntary

[15] To see why replicability is important to causal inference, consider Bem (2011). Published in a prominent psychology journal, Bem (2011) purports to show evidence of extra sensory perception using randomized control trials. Gelman and Loken (2014) point out that subsequent studies have failed to replicate the results under ostensibly identical conditions, and argue that the original results stem from ex post hypothesizing and "researcher degrees of freedom" in designing and interpreting the results (see also Simmons et al., 2011). See Hail et al. (2020) for a discussion of the "reproducibility problem" in accounting research.

[16] Recent examples include Lawrence et al. (2018), Belnap (2020), Belnap et al. (2020), and Umar (2020).

Table 1
Ad hoc quasi-experiments.

Panel A. Staggered Adoption of Workplace Smoking Laws	
Dependent Variable:	*VolDisc*
Variable	(1)
Smoking Law$_t$	−0.48**
	(−2.52)
Firm Fixed Effects	yes
Industry x State Fixed Effects	yes
Industry x Year Fixed Effects	yes
Adj R^2	51.8
N	56,516
Panel B. Staggered Adoption of State Gun Laws	
Dependent Variable:	*VolDisc*
Variable	(1)
Gun Law$_t$	−0.06***
	(−2.78)
Firm Fixed Effects	yes
Industry x State Fixed Effects	yes
Industry x Year Fixed Effects	yes
Adj R^2	60.2
N	108,381

This table presents results from a generalized difference-in-differences estimation using staggered state adoption of workplace smoking laws and gun waiting laws respectively (see Gao et al., 2020 and Edwards et al., 2018, for data on these laws). *VolDisc* is the number of management forecasts issued during the respective year. Each column includes untabulated *Firm*, *Industry* x *State*, and *Industry* x *Year* fixed effects. Sample of 56,516 firm-years from 1996 to 2008 for workplace smoking laws; and 108,381 firm-years from 1995 to 2017 for gun laws. Industries are defined using two-digit SIC codes. *t*-statistics appear in parentheses and are clustered by firm. *, **, *** indicate statistical significance (two-sided) at the 0.1, 0.05, and 0.01 levels, respectively.

disclosure in this context. Instead, we have simply run a canned statistical routine and "let the data speak." Without a compelling theory of why these laws might relate to voluntary disclosure, the correlations are simply empirical facts and uninterpretable in a causal manner. Importantly, our analysis does not preclude subsequent research from developing a theory that would allow us to interpret these correlations in a causal manner. This view recognizes the provisional nature of our knowledge—subsequent developments in the field may alter how we interpret facts provided previously.

Second, we might be tempted to suggest that we can draw causal inferences from these results if the statistical significance remained after a series of additional robustness tests (e.g., parallel pre-trends). Suppose the coefficients are statistically significant after conducting these additional tests: would we then be comfortable interpreting the correlations as evidence of a causal relation? For, example, as discussed and shown in Section 5.2, these relations pass common visual tests for parallel trends (see e.g., Fig. 10). Nonetheless, we are aware of no theory that explains the empirical facts in Table 1, and thus are unable to interpret the facts in a causal manner. When seeking to draw causal inferences, no amount of econometrics can substitute for a lack of theoretical foundation, and we caution against viewing causal inferences as a purely empirical endeavor.

Importantly, the more precise the theory, the greater the potential for empirical identification. If the theory is imprecise, then we might not have an *a priori* reason to expect that management forecasts is a more (or less) appropriate measure of voluntary disclosure in this setting than using Form 8-K filings, or firm-initiated press releases. Similarly, we might not have an *a priori* reason to expect that a specification based on *Firm*, *Industry* x *State*, and *Industry* x *Year* fixed effects is any more (or less) appropriate than a specification based on *Firm*, *Industry, State*, and *Year* fixed effects. When choosing a specific measure of voluntary disclosure and articulating a specific regression specification, we are implicitly invoking theory. For example, by including *Firm* fixed effects we are implicitly assuming that the omitted variable threat primarily varies across firms. If this assumption is incorrect, then the inclusion of these fixed effects is not harmless, and may in fact lead to bias in favor of rejecting a true null hypothesis (see Section 4.3).

Thus, one cannot meaningfully speak to an "identification strategy," a "tight design," or "credible evidence" without articulating a theory. On what basis is the design in Table 1, "well identified," or "tight"? For example, results in Table 3 (in Section 4.3) suggests our inferences are highly sensitive to the fixed effect structure, and Table 5 (in Section 5.1) suggests our inferences are unique to using management forecasts to measure voluntary disclosure. Because we do not have a precise theory linking smoking laws and gun laws to voluntary disclosure, we have no basis on which to prioritize the findings from one measure of voluntary disclosure, where we find results (e.g., management forecasts), over another measure of voluntary disclosure where we do not find results (e.g., firm-initiated press releases). Hence, because we have no theoretical justification to prioritize results in Table 1 over those in Table 5, we might be skeptical of the results in Table 1.

3.4. The importance of theory for generalizability

The process of generalizability refers to extrapolating inferences learned from a single empirical test (or set of tests) to circumstances occurring out of sample—and ultimately to the underlying theory being tested. On one hand, if the study has no theoretical foundation, then there is no basis on which to generalize inferences beyond the setting. On the other hand, if the empirical study closely follows and tests a general theory in a specific setting, we have good reason to believe that the inferences from that setting will generalize. Thus, the ability to generalize one's inferences beyond a specific setting hinges critically on the strength of the paper's theoretical foundation. If the theory is not compelling, then inferences are necessarily limited to the stylized setting.

Concerns about generalizability are particularly acute for studies examining stylized settings in which the setting itself is not of inherent interest, but rather is being used *exclusively* as a source of as-if random variation in a particular theoretical construct (Leuz and Wysocki, 2016; Glaeser and Guay, 2017; Leuz, 2018). However, when the setting itself is of inherent interest, inferences drawn from stylized settings can make a valuable contribution, even when they cannot be generalized (Christensen, 2019). For example, it is unquestionably important to study the Great Depression, the 2007–2008 financial crisis, and specific accounting frauds (e.g., Enron) despite the fact that inferences from these settings are unlikely to generalize. For example, in our survey, 65% of papers drawing causal inferences do so in the context of changes in regulations (e.g., Fig. 3). Changes in regulation are—by their very nature—not generalizable, as the effects of regulation depend on the institutional circumstances of the setting at hand.[17] Nevertheless, by studying changes in regulation, empirical researchers can provide valuable evidence to the public that inform policy debates—even if the inferences do not generalize (Leuz, 2018).[18]

In this section, we use a series of examples to illustrate two points: (i) concerns about generalizability depend on the objective of the research, and (ii) in settings that are not of inherent interest to accounting researchers, the contributions of many papers often rest implicitly on the ability of researchers to generalize beyond the setting examined in the study.

In our first set of examples, we consider two studies that examine the effects of IFRS on corporate policies. The first study is Barth et al. (2012), entitled “Are IFRS-based and US GAAP-based Accounting Amounts Comparable?” This study examines whether the adoption of IFRS increases accounting comparability. As the title suggests, this paper is inherently interested in the effect of IFRS *per se*, and thus concerns about generalizing results beyond IFRS adoption are minimal. This contrasts with the second study, Hail et al. (2014), entitled “Dividend Payouts and Information Shocks.” As the title suggests, this study examines whether exogenous reductions in information asymmetry between managers and shareholders cause the firm to alter its payout policy. The study draws on a broad theory of agency conflicts between managers and shareholders and uses IFRS as one setting that provides exogenous variation in the construct of interest (i.e., information asymmetry). Hail et al. (2014) explicitly seek to generalize inferences from the IFRS setting to a general theory about information asymmetry. Consequently, generalizability—i.e., the extent to which we learn about the general theory, as opposed to about IFRS *per se*—is a significant concern that this paper seeks to mitigate.

These examples illustrate that the extent to which generalizability is a concern depends on the objective of the study. In the case of the previous examples, however, the literature cares about the setting itself—IFRS—regardless of the ability to generalize. Thus, we might care about the contribution of Barth et al. (2012) and Hail et al. (2014) even if we do not believe the inferences generalize beyond IFRS. This is not always the case, and there are many examples of settings that are unlikely to be of inherent interest. In such cases, although these settings can offer interesting opportunities to discover new effects or test new theories (e.g., Leuz and Wysocki, 2016), generalizability is potentially the single-greatest concern with the study—the study must generalize beyond the setting to motivate its contribution to the accounting literature. We provide three examples where the settings themselves are unlikely to be of inherent interest to accounting researchers and the contribution hinges on their ability to generalize.

Soldier fatalities. Soldier fatalities are unlikely to be of interest to accounting scholars unless they can be used as a measure of political costs (Boland and Godsell, 2020). If one cannot generalize from the empirical relation between soldier fatalities and discretionary accruals to the theoretical relation between political costs and managers’ financial reporting decisions, then the former is unlikely to be of interest to accounting scholars.

Organ donations. Organ donations are unlikely to be of interest to accounting scholars, unless they can be used as a measures of social capital (Hasan et al., 2017). If one cannot generalize from the empirical relation between organ donations and tax avoidance to the theoretical relation between social capital and tax avoidance, then the former is unlikely to be of interest to accounting scholars.

[17] See Leuz and Wysocki (2016) and Christensen (2019, p. 1): “*it is hard to find any regulatory setting from which results can easily be generalized to other settings*.”

[18] For a recent example of research informing policy debate and the importance of generalizability (or lack thereof), see the debate over SOX 404(b) internal control audits in the context of the SEC's Revisions to Accelerated Filer and Large Accelerated Filer Definitions (SEC Release 34—88365) and accompanying statement by Commissioner Jackson: https://www.sec.gov/news/public-statement/jackson-statement-proposed-amendments-accelerated-filer-definition.

Facial features. Individuals' facial features are unlikely to be of interest to accounting scholars, unless they can be used as a measure of individuals' trustworthiness or other inherent personal traits (e.g., Jia et al., 2014; He et al., 2019; Dikolli et al., 2020; Hsieh et al., 2020; Peng et al., 2021). If one cannot generalize empirical relations between facial features and accounting outcomes to theories of individuals' personal traits, then the former are unlikely to be of interest to accounting scholars.

Because accounting researchers are unlikely to be concerned with soldier fatalities, organ donations, or facial features *per se*, the contribution of each of these studies implicitly hinges on their ability to generalize inferences to theories that interest accounting researchers. As a consequence, the theoretical foundations linking each of the empirical measures with the underlying theoretical construct is critical. In contrast, because accounting researchers are inherently interested in IFRS, the contribution of papers examining IFRS as a setting (e.g., Barth et al., 2012) does not necessarily depend on their ability to generalize beyond the setting.

We take no stance on the appropriateness of the generalizability or contributions of any of these papers, but rather use these examples to illustrate that as the literature evolves toward using more highly stylized settings further removed from traditional accounting settings, generalizability and the importance of theoretical foundations that allow for generalizability become increasingly important. Indeed, if the literature is increasingly gravitating toward the use of stylized settings for the purposes of causal inferences, as our survey suggests, and generalizability is more important in such settings, as the above discussion suggests, then we would expect to observe increasing trends in the extent to which researchers generalize their inferences.

Earlier in the literature, our survey reveals a tendency to title and motivate a paper based on the setting itself. For example, we find 18%–19% of empirical papers each year from 2005 to 2007 had the phrase “Evidence from” in their title. Between 2017 and 2019 this drops to 11%–12% of empirical papers per year. This decline (from 19% to 12%) may seem small, but that it declined at all (rather than increased) is notable—the decline occurred during a period when papers drawing causal inference using quasi-experimental settings increased over 400%. The notion that there has been a pronounced increase in the number of papers using (often highly stylized) quasi-experimental settings to draw causal inferences, but a decrease in the number of papers motivating their analysis from the setting itself suggests an increasing comfort with generalizing inferences from stylized settings.

We conclude this section with practical guidance on the appropriateness for generalizing inferences and techniques one can use to mitigate concerns about generalizability. In doing so, we make two points. First, prior literature has used two techniques to address concerns about generalizability. The first technique is to explicitly admit that the findings likely do not generalize out of sample and to make the case that generalizability is less of a concern given the specific research question or setting of interest (e.g., Jagolinzer et al., 2020; Arif et al., 2022). The second technique is for the study to analyze multiple settings in the context of the same research question. For example, Hail et al. (2014) study the relation between information asymmetry and dividend payouts using multiple quasi-experiments; and Guay et al. (2016) study the relation between financial statement complexity and voluntary disclosure using both standard panel data techniques estimated on the Compustat population and multiple quasi-experiments (see also Duguay et al., 2020; Samuels, 2021; Samuels et al., 2021).

Second, it is potentially a contribution to establish that a result from one particular setting in prior literature generalizes to multiple settings. Simply because a result exists in one particular setting or even exists in multiple settings in prior work (e.g., mandatory disclosure quality improves liquidity), does not imply that the result generalizes to all settings (e.g., a specific regulatory change). For example, the IFRS literature has long recognized that the effect of adopting IFRS in the European Union does not necessarily generalize to the effects of adopting IFRS in other settings (Christensen et al., 2016; Glaeser and Guay, 2017). Establishing that inferences in prior research do or do not generalize to other settings is often an important contribution. This is both the beauty and the curse of studying highly stylized settings: institutionally, no two settings are the same—*ex ante*, one can neither generalize from them, nor generalize to them.

4. Implementation issues surrounding quasi-experiments in the accounting literature

Given the widespread use of quasi-experimental methods documented in Section 2, and the conceptual underpinnings of quasi-experimental methods and causal inference discussed in Section 3, in this section we discuss several practical issues and associated implementation challenges common to these methods. Our literature survey reveals three core implementation challenges concerning causal inferences that researchers often face. We cover each of these in turn.

Section 4.1 highlights the distinction between an event that is exogenous and an event that provides as-if random variation. Section 2 suggests many settings studied in accounting (e.g., regulation), although arguably exogenous, often entail non-random assignment to the treatment group, which imparts a selection bias in common DiD designs. We illustrate the importance of distinguishing between these two concepts in the context of several settings studied in accounting research.

Section 4.2 reviews common diagnostic tests that are useful for assessing whether the parallel trends assumption of the DiD design holds. These diagnostic tests rely on the assumption that one can infer the unobserved counterfactual relations in the post-period from observed relations in the pre-period. Although parallel pre-trends diagnostic evidence can provide useful information, we caution that parallel pre-trends are neither necessary nor sufficient for causal inferences. We illustrate common diagnostics in the context of staggered adoption of workplace smoking laws introduced in Table 1.

Section 4.3 discusses two tradeoffs associated with high-dimensional fixed effect designs common in the literature. As discussed in Section 3.1, these designs can be particularly useful for helping rule out correlated omitted variables. However, these designs are not a panacea. First, we show that when the source of the variation in a correlated omitted variable is within-group, including group fixed effects can exacerbate omitted variable bias. Second, we show that including high-dimensional fixed effects can induce significant multicollinearity and increase the sensitivity of regression results to a handful of observations.

4.1. “Exogenous” vs. “as-if random”

4.1.1. Concepts

In this section, we highlight the distinction between an event that is exogenous and an event that provides as-if random assignment. Although this distinction is critical for causal inference, many papers still do not explicitly draw this distinction (see, e.g., Atanasov and Black, 2016; Hennessy and Strebulaev, 2020 for related discussion). We adopt the definition of exogenous in the Oxford dictionary: “Having an external cause or origin. Often contrasted with endogenous: e.g., technological changes exogenous to the oil industry.” This definition makes clear that an exogenous event (or variable) refers to something that originates outside the system being studied. This speaks to the origin of the event, but not whether it provides as-if random variation, i.e., whether the event randomly distributes firms among affected and unaffected. In this regard, an event can be plausibly exogenous to the firm being studied (e.g., the introduction of an accounting standard) but fail to provide as-if random assignment to treatment and control groups.

Section 2 suggests that the vast majority (65%) of papers using quasi-experimental methods in accounting research study regulatory settings. One commonly studied circumstance in which an exogenous event does not provide as-if random assignment is a regulation that mandates firms take an action that was previously a voluntary choice. Prominent examples include the adoption of IFRS (e.g., Daske et al., 2008), the Sarbanes-Oxley Act (Zhang, 2007), the NYSE's 2003 board independence standards (Armstrong et al., 2014), California's 2018 board gender diversity standards (Greene et al., 2020), the NASDAQ's recent board gender diversity rule, and the electronic filing of SEC forms (e.g., Samuels et al., 2021). In each case, the regulations directly affect only the subset of entities that endogenously chose not to undertake the action prior to the mandate.[19] This creates a classic selection problem, because assignment to the treatment depends on the entity's endogenous choices prior to the mandate.

For example, California's board gender diversity mandate only applies to firms that chose not to be gender-diverse before the mandate (e.g., firms without a single woman on the board as of 2019). If gender diversity provided net benefits to the firm, presumably the firm would have endogenously chosen to be gender diverse prior to being required to do so. Consequently, the regulation only treats those entities for which the costs exceeded the benefits in the voluntary regime, which creates selection bias in the assignment to treatment and control groups. Those in the treatment group are those that chose not to be gender diverse before it was required.

The implication of this selection problem—i.e., a lack of as-if random variation—is that the estimated causal effect from a DiD design will not represent an estimate of the average treatment effect, or ATE, which is the expected causal effect on an observation selected at random from the population. Instead, the estimated effect will be the sum of (1) the estimated average treatment effect on the treated or ATT, which is the expected causal effect only on those observations receiving treatment, *and* (2) a selection bias (Angrist and Pischke, 2008 p.14). This contrasts with a setting where one has truly random assignment, whereby estimates from a DiD design recover the ATE (e.g., Angrist and Pischke, 2008; Muller et al., 2014).

These concepts have important implications for how to interpret studies' results. For example, consider the studies examining Norway's 2006 requirement that firms have at least 40% female representation on the board. These standards only affected firms that did not previously choose to have at least 40% female representation. Consequently, assignment to the treatment group is not as-if random. Although the literature suggests that *mandated* diversity had a negative effect on the value of the affected firms—those with non-diverse boards (e.g., Matsa and Miller, 2013)—this effect cannot speak to the set of firms that chose to be diverse, and by extension, cannot speak to the causal effect of diversity (on the average firm in the population, i.e., the ATE).

In this case, by studying the effect of the regulation, we can learn about the causal effect of the mandate on a subset of firms in the economy; but causal inferences are unique to the *mandate* and do not generalize to the causal effect of *diversity* itself (see also Leuz and Wysocki, 2016). Although some studies are explicit in distinguishing between these two concepts (e.g., Daske et al., 2008; Christensen et al., 2013; Guest, 2021, to name just a few), this distinction is often not made in the papers appearing in our survey. For example, there is a large literature studying the 2003 NYSE and NASDAQ rules that mandate firms to have a board of at least 50% independent directors (e.g., Duchin et al., 2010; Armstrong et al., 2014; Chen et al., 2015).

[19] We use the term “direct effects” to refer to the set of firms affected by the regulation, independent of any spillovers or externalities (i.e., “indirect effects”).

Despite the fact that the treatment group is not as-if random *by construction*, most of the papers in this literature draw an equivalence between the causal effect of the mandate and the causal effect of board independence, which is incorrect.[20] For this reason, when studying regulation, we encourage the literature to limit inferences to the causal effect of the regulation itself (e.g., the causal effect of NYSE board independence rules), as opposed to generalizing inferences to the causal effect of the underlying action being altered by the regulation (e.g., the causal effect of board independence).

4.1.2. Numerical example

We illustrate the importance of these concepts in the context of a theoretical economy in which a previously voluntary behavior is made mandatory. In this theoretical economy, the effect of a mandatory disclosure (although plausibly exogenous) is explicitly conditional on choices that were previously voluntary. We then examine the potential bias imparted on popular econometric methods.

Our theoretical economy comprises 1000 firms (indexed by i) in each of three periods (indexed by t). Cash flows in each period are given by $x_{i,t} \sim N(\mu,\sigma^2)$, i.i.d. In the second period, there is an innovation (e.g., the creation of IFRS, the creation of EDGAR, or performance-based vesting). As a result of the innovation, beginning in $t = 2$, firms can choose whether to take an action, which we represent as $z_{i,t} = \{0,1\}$. For 500 firms in the economy (i.e., $i = 1, \ldots, 500$), we assume choosing $z = 1$ *increases* expected cash flows in that period by c and choosing $z = 0$ has no effect on cash flows. We refer to these as Type A firms. For the remaining 500 firms in the economy (i.e., $i = 501, \ldots, 1000$), we assume choosing $z = 1$ *decreases* net cash flows in that period by c and choosing $z = 0$ has no effect on cash flows. We refer to these as Type B firms.

Before proceeding, it is instructive to consider the average treatment effect of z on cash flows. On average, if we randomly selected firms from the economy and compelled them to take the action (i.e., exogenously set $z = 1$ for a random sample of firms), in expectation, there would be no effect on cash flows. To see this, note that exogenously setting $z = 1$ would increase cash flow by c for 50% of the economy (Type A firms), and decrease cash flow by c for the remaining 50% of the economy (Type B firms). Consequently, the average treatment effect, ATE, is zero.

Now suppose that firms can *choose* to take action z beginning in period 2. We assume firms choose the action that maximizes expected cash flow in that period. As a result, Type A firms choose to take the action (choose $z = 1$), Type B firms do not take the action (choose $z = 0$), and expected cash flows are given by:

	$t = 1$	$t = 2$	$t = 3$
Type A	μ	$\mu+c$	$\mu+c$
Type B	μ	μ	μ

Now suppose a regulator mandates all firms take the action, i.e., $z = 1$, in period $t = 3$ (e.g., mandated board diversity, mandated adoption of IFRS, or mandated climate disclosure). Because Type A firms were already taking the action, only Type B firms are directly affected by the mandate. As a result, expected cash flows are now given by:

	$t = 1$	$t = 2$	$t = 3$
Type A	μ	$\mu+c$	$\mu+c$
Type B	μ	μ	$\mu-c$

How would we estimate the causal effect of action z on cash flow? One option to estimate this effect would be to focus only on Type B firms, and estimate a simple pre-post design (i.e., compare differences in outcome for Type B firms before and after the mandate). Another option would be to estimate the classic and generalized difference-in-differences designs described in Section 3, where *Post* is the indicator variable for period 3, and *Treat* is an indicator variable if the firm is treated by the mandate (Type B firms).

Table 2 shows the results from these three estimations assuming $\mu = 0$, $\sigma^2 = 1$, and $c = 1$. Column 1 of Table 2 presents results from the simple pre/post design, and suggests the effect is negative. Note that, in this case, the simple pre/post design in column 1 recovers an unbiased estimate of the ATT, i.e, the causal effect *of the mandate* on the treated group (ATT = −1.0). In contrast, columns 2 and 3 report results from the DiD designs and estimate an effect of −1.5, which is larger than either the ATT or the ATE (ATE = 0). Here, the DiD estimator is the sum of (1) the ATT *and* (2) a selection bias (Angrist and Pischke, 2008). Thus, in a setting where a voluntary action is subsequently made mandatory, DiD designs can bias in favor of finding a result.

The intuition for why, in the example above, a more sophisticated DiD design introduces bias that is not present in the simple pre/post design is that, in the latter design, there is no (inappropriate) control group. Consequently, there is no selection bias from using a control group whose voluntary choices in the pre-period prevented treatment. This simple numerical example illustrates why the distinction between an exogenous event and an event that provides as-if random assignment is important. An exogenous event (e.g., a regulation) does not necessarily provide as-if random assignment to

[20] The 2003 NYSE/NASDAQ mandate requires listed firms to have greater than 50% independent directors. As a result, the change in the percentage of independent directors required by the rule explicitly depends on the firm's pre-existing endogenous choice of board independence. Thus, in this case, neither the level, the change, nor the required minimal change in independent directors is as-if random.

Table 2
Simulated economy: The relation between innovation and cash flows.

Dependent Variable:		*Cash Flow*		
Research Design:		Pre-Post	Classic DiD	Generalized DiD
Variable		(1)	(2)	(3)
Post	Avg coef.	−1.00	0.50	.
	Avg p-value	<0.01	<0.01	.
Treat	Avg coef.	.	−0.50	.
	Avg p-value	.	<0.01	.
*Treat * Post*	Avg coef.	.	−1.50	−1.50
	Avg p-value	.	<0.01	<0.01
Firm Fixed Effects		no	no	yes
Period Fixed Effects		no	no	yes
N		1500	3000	3000

This table presents results from regression estimates using the data from the simulated economy described in Section 4.1.2. For purposes of simulation, we set $\mu = 0$, $\sigma^2 = 1$, and $c = 1$. *Post* equals one in period $t = 3$, and zero otherwise. *Treat* equals one for Type B firms, and zero otherwise. *Firm* and *Period* fixed effects are included in column (3) and subsume the *Post* and *Treat* main effects. We present average values of the coefficients for 1000 iterations of the simulation, and average *p*-values based on standard errors clustered by firm.

treatment and control groups. If an exogenous event assigns firms to treatment and control groups as a function of a previously voluntary choice—as is often the case with regulation—then a DiD design will suffer from selection bias.[21]

4.2. Testing for parallel trends

As discussed in Section 3.2, the critical assumption in all DiD methods is the parallel trends assumption. This assumption effectively is equivalent to the "no correlated omitted variable assumption" in cross-sectional interaction design. As a result, just as one cannot test for the existence (or absence) of correlated omitted variables, one cannot test the parallel trends assumption. For example, consider the classic DiD regression specification discussed in Section 3.2.1:

$$y_{i,t} = \alpha_1 + \alpha_2 D_t + \beta_1 x_i + \beta_2 (D_t * x_i) + \varepsilon_{i,t} \tag{15}$$

The parallel trends assumption can be stated as $cov(D_t * x_i, \varepsilon) = 0$ (see, e.g., Roberts and Whited, 2013, p. 528). The reason that this is named the "parallel trends" assumption rather than the "no correlated omitted variables" assumption is because D_t is a time-indexed variable for all observations after treatment, and consequently the omitted variable would need to vary with D_t. One way that the omitted variable would vary with D_t is if x_i partitions the sample into treatment and control groups, and these groups have different temporal trends in $y_{i,t}$. Hence the label "parallel trends."

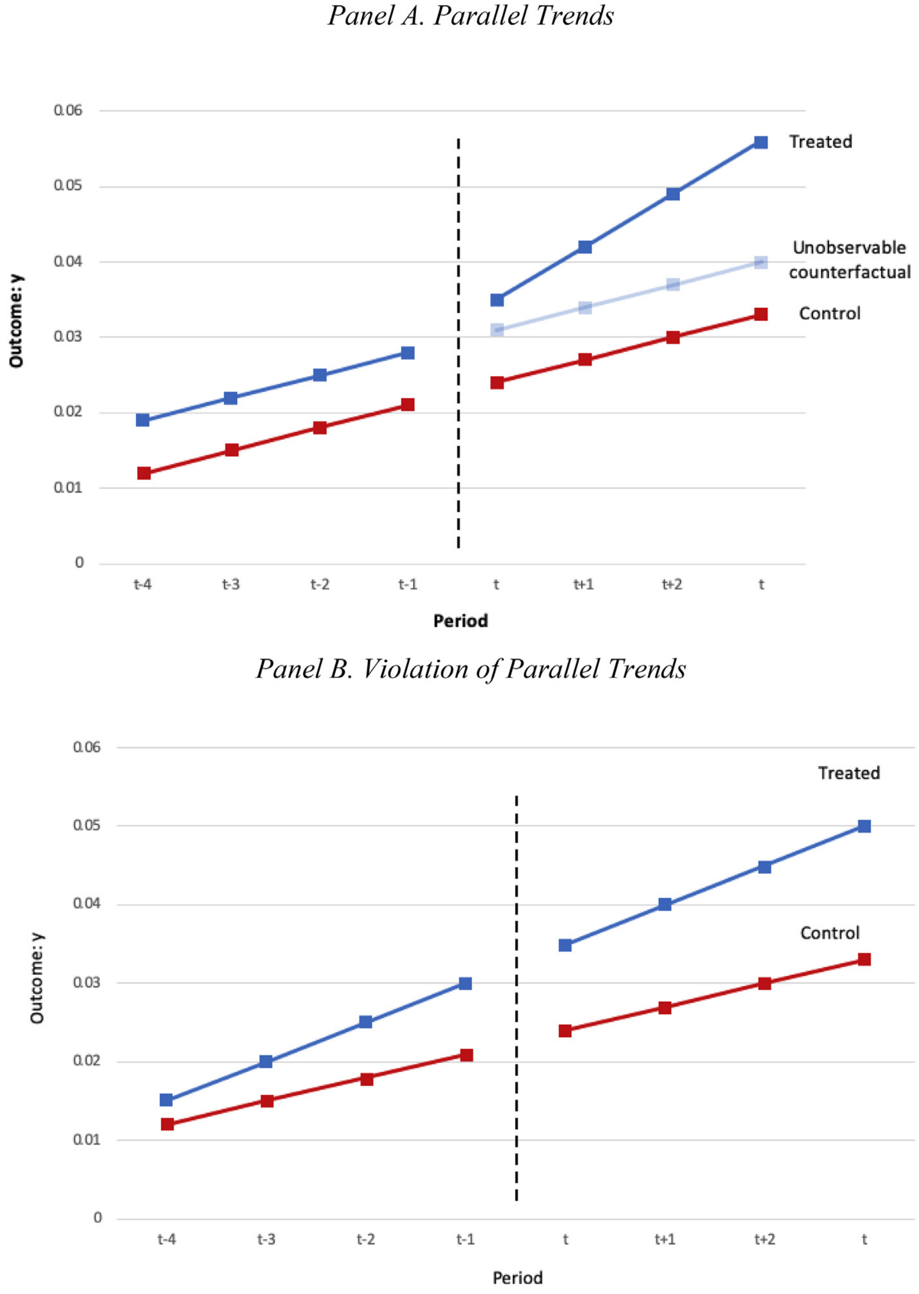
The DiD design assumes that trends in the outcome would be the same for the treatment and control groups in the absence of treatment. Panel A of Fig. 7 illustrates the parallel trends assumption of DiD designs. In particular, Panel A plots average values of the outcome y over time separately for treatment and control groups. In this example, although there is a difference in outcomes between treatment and control groups, this difference is constant over time in the absence of treatment. Before treatment, this difference between treatment and control is observable, and is commonly referred to as "pre-trends." In the post-period, however, this difference in the absence of treatment is assumed, as illustrated by the unobservable counterfactual in the figure. Note that this is inherently a counterfactual statement, and although we can test for different trends in the absence of treatment in the pre-period, we cannot test for differences in trends in the absence of treatment in the post-period.

The intuition behind most diagnostic tests of the parallel trends assumption is the additional assumption that if pre-trends are *not* parallel then the unobserved post-trends are unlikely to be parallel either. Hence, diagnostic tests focus on assessing whether the pre-trends are in fact parallel. In Panel A of Fig. 7, the pre-trends are parallel, and we might feel comfortable suggesting that the parallel trends assumption holds. In contrast, in Panel B of Fig. 7, the pre-trends are not parallel, and we might be concerned that the parallel trends assumption does not hold.

Given the focus on the temporal differences between treatment and control, a popular alternative way to present these plots is to focus on the difference between treatment and control groups over time—see Fig. 8, which plots the equivalence of Fig. 7 in differences. Note however that although the two presentations are equivalent for the purposes of assessing parallel pre-trends, there is a loss of information in Fig. 8. Fig. 8 does not inform us whether the treatment sample or the control sample drives the differences between the two groups over time.

To implement Fig. 8 in a classic DiD regression design, we replace the indicator D_t with separate indicators for *each* period in the time-series (i.e., period fixed effects) denoted by the vector $Year_t$:

[21] For those readers interested in a more detailed treatment related to issues beyond the distinction between "exogenous" and "as-if random" see Angrist and Imbens (1994) for the estimation of local average treatment effects (LATE); Heckman et al. (2001) and Heckman and Vytlacil (2001) for a related discussion including application to bounding treatment effects; and Breuer (2021) for a discussion of Bartik instruments, which can aid in reducing endogeneity concerns in common DiD designs. Throughout our discussion we have assumed "perfect compliance" (i.e., those assigned to the treatment group are in fact treated). See Armstrong and Kepler (2018), Glaeser and Guay (2017), and Armstrong et al. (2021a,b) for a discussion of this assumption as well as the stable-unit treatment value assumption.



Panel A. Parallel Trends

Panel B. Violation of Parallel Trends

Fig. 7. Parallel Trends Assumption. This figure provides a visual representation of the parallel trends assumption. Each panel plots average values of the outcome *y* over time separately for treatment and control groups.

$$y_{i,t} = \alpha_1 + \beta_1 x_i + \boldsymbol{\theta}(Year_t * x_i) + Year_t + \varepsilon_{i,t} \tag{16}$$

Note that when estimating this, the researcher must exclude one of the periods from the fixed effects to serve as the benchmark period. β_1 measures the difference between treatment and control groups in the benchmark period, and the resulting θ coefficients represent the difference in outcome between the treatment and control groups in each period in excess of the benchmark β_1. For example, suppose treatment occurs in year 2000 and the benchmark period is 1999, then the θ coefficient for each year represents the difference between treatment and control groups in that particular year *in excess of* the difference between the two groups in 1999.

In estimating this diagnostic test, the researcher has at least two important degrees of freedom beyond the choice of regression specification. First, the researcher has discretion over the choice of benchmark period. In practice the results from our survey suggest the benchmark period is commonly the period immediately preceding the treatment event. Second, the researcher might wish to aggregate observations across multiple periods beyond particular calendar time thresholds. For example, if treatment

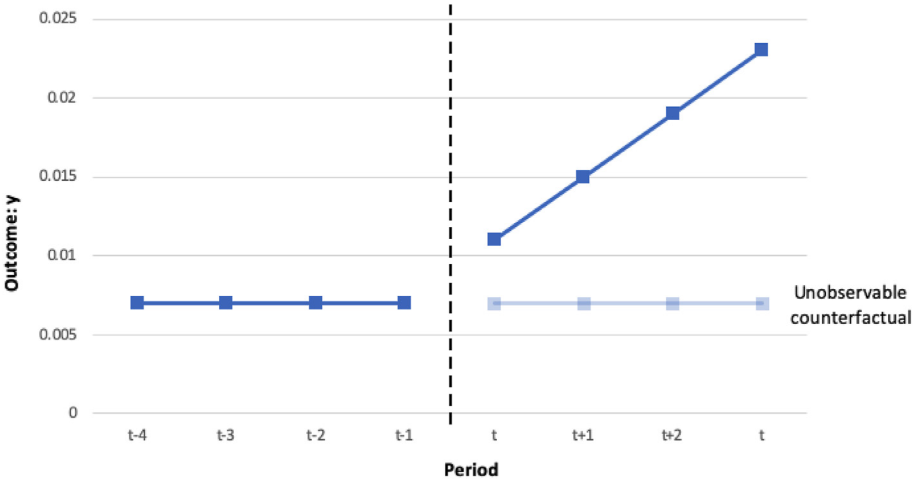


Fig. 8. Parallel Trends: Difference between Treated and Control Groups. This figure provides a visual representation of the parallel trends assumption for the average values of the outcome *y* over time for the difference between treatment and control groups.

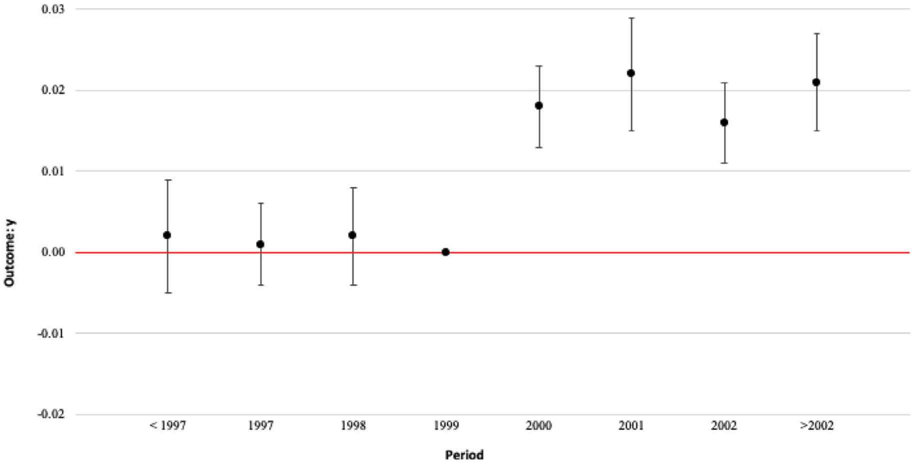


Fig. 9. Diagnostic Test Pre and Post 2000. This figure illustrates diagnostic parallel trends plots for our example discussed in Section 4.2 in which treatment occurs in the year 2000 and 1999 is chosen as the benchmark period, with separate coefficients for: (i) all years before 1997, (ii) 1997, (iii) 1998, (iv) 1999 (which will be 0 by construction), (v) 2000, (vi) 2001, (vii) 2002, and (viii) all years after 2002.

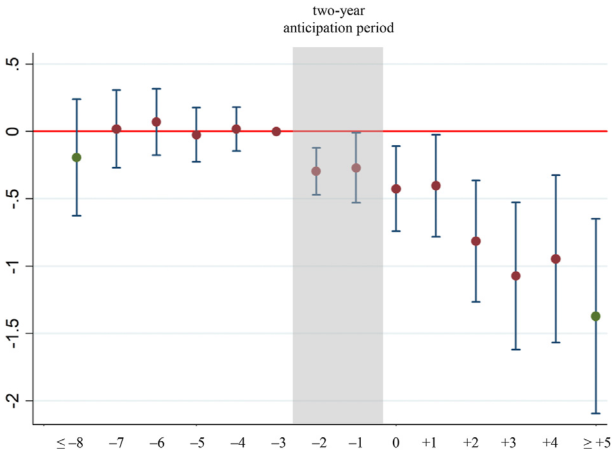


Fig. 10. Illustration of Parallel Trends Test Using Staggered Adoption of Workplace Smoking Laws. This figure illustrates diagnostic parallel trends in voluntary disclosure (*VolDisc*) for staggered adoption of workplace smoking laws presented in Panel A of Table 1. $t = 0$ is the year the law became effective, and $t = -1$ and -2 is the two-year anticipation period. We set the benchmark period to $t = -3$. Created using the *Eventdd* Stata package.

occurs in the year 2000, the researcher may choose 1999 as the benchmark period, and calculate eight separate θ coefficients: (i) for all years before 1997, (ii) 1997, (iii) 1998, (iv) 1999 (which will be 0 by construction), (v) 2000, (vi) 2001, (vii) 2002, and (viii) all years after 2002. Fig. 9 illustrates how one might provide a visual representation of these coefficients and their confidence intervals.

Although we develop the intuition for these diagnostic tests in the context of a classic DiD design, the diagnostic tests can be applied to extensions of the DiD design discussed in Section 3.2. We now discuss how one would extend this test to the staggered DiD design in which the post-period is not aligned in calendar time. Our discussion follows Barrios (2021). The first step is to restrict attention to only those entities—e.g., firms—that are treated at some point in the data (e.g., Christensen et al., 2016). For each observation, the researcher then calculates a "time to treatment" variable, which is the difference between the current period and the period in which treatment occurs. Rather than create and include time period indicator variables based on calendar time, the researcher creates and includes time period indicators based on event time based on the time to treatment variable—i.e., separate indicator variables for periods $t < -3$, $t = -3$, $t = -2$, $t = -1$, $t = 0$, $t = 1$, $t = 2$, and $t > 2$ relative to treatment (see Clarke and Schythe (2020) for the equation). Here again, the researcher must specify a choice of benchmark period, and whether and how much to aggregate periods before or after treatment (e.g., $t < -3$, $t > 2$).[22]

Fig. 10 presents results from estimating this diagnostic test for the staggered adoption of workplace smoking laws presented in Section 3.3. Similar to Barrios (2021), we allow for a two-year anticipation of the law by setting period $t = -3$ as the benchmark period (i.e., year $t = 0$ is the year the law went into effect). As a result, the coefficient estimates compare the difference between treatment and control in the respective year to that in $t = -3$. The evidence in Fig. 10 suggests that the pre-trends are parallel. Fig. 10 also highlights how, in addition to assessing pre-trends, these graphical tests can be used to assess (i) how quickly the treatment effect begins to manifest after treatment (i.e., whether there is a delay or any anticipatory effect), and (ii) whether the treatment effect is persistent (i.e., rather than being specific to a given period). Thus, these tests can be a powerful tool to complement a DiD analysis.

However, we caution that this figure, and associated diagnostic tests, rely on a correctly specified regression model and appropriate measure of the outcome of interest—something we will call into question in subsequent sections. If the regression model and/or associated measures are misspecified, so too is the resulting diagnostic test. Thus, although these graphical diagnostic tests are useful, it is important to note that they are not a panacea. First, these tests rely on the assumption that one can infer the unobserved counterfactual relations in the post-period from observed relations in the pre-period. This assumption is inherently untestable, and its validity hinges critically on the underlying theory and institutional details related to the specific setting being examined. Second, given the reliance on visual representations, and the notion that different readers can interpret the same visual representation very differently, there is often considerable subjectivity in whether a given figure does or does not support parallel pre-trends. In this regard, we caution against placing too much weight on figures alone. Thus, although parallel trends diagnostic evidence can provide useful information, the evidence is neither necessary nor sufficient for causal inferences.[23]

4.3. *Common fixed effect designs*

One of the hallmarks of many of the difference-in-differences designs used in the literature is the inclusion of high-dimensional fixed effects (e.g., a large number of fixed effects).[24] The conventional wisdom on fixed effects is that they tend to remove variation and, as a result, tend to "bias against" finding a statistically significant result (i.e., a reduction in power). Indeed, as discussed in Section 3.1.3, a key benefit of this approach is that it can remove significant problematic variation from omitted variables without the researcher actually needing to specify the omitted variables. For example, deHaan (2020) discusses how group-level fixed effects can eliminate omitted variable bias when the omitted variable does not vary within the group. In such a circumstance, the inclusion of group fixed effects is helpful in alleviating concerns about omitted variable bias. However, these methods are not a panacea, and are not without tradeoffs. In Section 3.1.3, we discussed the potential advantages of fixed effects. In this section, we discuss two potential disadvantages.

4.3.1. *Potential to exacerbate omitted variable bias*

First, fixed effects only remove variation across groups, but not within-group variation. If within-group variation is the source of the correlation with the omitted variable, then the inclusion of group-fixed effects will isolate this variation (i.e., remove all other variation), which will amplify the correlation with the omitted variable and exacerbate omitted variable bias.[25] Thus, an implicit assumption in fixed effect designs is that the data generating process for the correlated omitted

[22] Barrios (2021) provides a primer for how to implement this diagnostic test using the *eventdd* package in Stata.

[23] For those readers interested in a more detailed treatment of issues related to parallel trends see the following literature (i) Arkhangelsky et al. (2021) for available techniques (e.g., "synthetic difference-in-differences"): to apply in the event that pre-trends in outcome variables are not parallel (see also, Ben-Michael, Feller, and Rothstein, 2021a, b; Callaway and Sant'Anna, 2021; Rambachan and Roth, 2019; and Roth, 2019); and (ii) Barrios (2021) and Baker et al. (2021) for a discussion of additional issues in the context of diagnostic tests of parallel trends in the presence of staggered treatment events over time commonly studied in accounting and finance applications.

[24] See the *reghdfe* package in Stata (Correia, 2017).

[25] For example, Jennings et al. (2022) discuss a circumstance where there is within-firm, correlated measurement error in the variable of interest (state of corporate headquarters from Compustat).

variable does not entail significant within-group variation. If there is significant within-group variation in the correlated omitted variable, then a within-group analysis will potentially be more biased than a pooled (across group) analysis.

The Appendix illustrates a simple data generating process—unknown to the researcher—in a setting in which the correlated omitted variable varies within group and not across groups. In this setting, the Appendix solves analytically for the omitted variable bias, shows that the omitted variable bias is larger when fixed effects are included, and shows that the bias is increasing in the percentage of variation in the independent variable absorbed by the fixed effects.

Without knowledge of the underlying data generating process, or a precise theory that suggests the source of the variation, it is difficult to know whether to include or exclude a particular set of fixed effects. Accordingly, given the potential for bias in favor of finding a result, it is useful to motivate the inclusion of fixed effects (rather than including them "by default") and to assess the robustness of results to alternative fixed effect structures.

Consider what this might look like in practice. Table 3 repeats the analysis of Table 1, triangulating inferences across different fixed effect structures. Column (1) reports results from Table 1 that include *Firm*, *Industry* x *State*, and *Industry* x *Year* effects. Column (2) reports results for *Firm*, *Industry*, *State*, and *Year*. Column (3) removes the industry effect—firms rarely change industries and firm effects are included. Column (4) removes the firm effects. Because the underlying data generating process is unknown to the researcher, and we do not have strong theory to privilege one specification over another, the results from these additional tests are informative. In particular, the results in Table 3 suggest that we might be more skeptical of the results for gun laws (i.e., the results are not robust to alternative fixed effect structures), and more confident in the results for workplace smoking laws (i.e., the results are robust to alternative fixed effect structures).

In addition to reporting coefficient estimates, the bottom row of each panel in Table 3 reports the percentage of variation in the independent variable that is absorbed by the fixed effects (i.e., the adjusted-R^2 from a regression of the independent variable on the fixed effects). Panel A suggests the various fixed effect structures absorb about 50% of the variation in workplace smoking laws, whereas they absorb up to 90% of the variation in gun restrictions. Interestingly, we find gun laws are statistically significant only in the setting where 90% of the variation is absorbed—meaning the coefficients are effectively estimated using a small fraction of the data, which can exacerbate the fragility of the results, as discussed below.

Table 3
Triangulation: Using alternative fixed effect structures.

Panel A. Staggered Adoption of Workplace Smoking Laws

Dependent Variable:	*VolDisc*			
	Firm, *Industry* x *State*, *Industry* x *Year* effects	*Firm*, *Industry*, *State*, *Year* effects	*Firm*, *State*, *Year* effects	*State* and *Year* effects
Variable	(1)	(2)	(3)	(4)
Smoking Law$_t$	−0.48**	−0.35*	−0.34*	−0.26
	(−2.52)	(−1.87)	(−1.81)	(−1.42)
Firm Fixed Effects	yes	yes	yes	no
Industry x State Fixed Effects	yes	no	no	no
Industry x Year Fixed Effects	yes	no	no	no
Industry Fixed Effects	no	yes	no	no
State Fixed Effects	no	yes	yes	yes
Year Fixed Effects	no	yes	yes	yes
Adj R^2	51.8	49.2	49.1	13.1
N	56,516	56,516	108,381	108,381
% variation in *Smoking Law* absorbed by fixed effects	57.0	55.7	55.5	48.1

Panel B. Staggered Adoption of State Gun Laws

Dependent Variable:	*VolDisc*			
	Firm, *Industry* x *State*, *Industry* x *Year* effects	*Firm*, *Industry*, *State*, *Year* effects	*Firm*, *State*, *Year* effects	*State* and *Year* effects
Variable	(1)	(2)	(3)	(4)
Gun Law$_t$	−0.06***	−0.02	−0.02	−0.01
	(−2.78)	(−0.89)	(−0.85)	(−0.28)
Firm Fixed Effects	yes	yes	yes	no
Industry x State Fixed Effects	yes	no	no	no
Industry x Year Fixed Effects	yes	no	no	no
Industry Fixed Effects	no	yes	no	no
State Fixed Effects	no	yes	yes	yes
Year Fixed Effects	no	yes	yes	yes
Adj R^2	60.2	57.43	57.37	13.4
N	108,381	108,381	108,381	108,381
% variation in *Gun Law* absorbed by fixed effects	90.4	89.6	89.6	85.4

This table presents results from repeating the analysis in Table 1 using three alternative fixed effect structures. Column (1) repeats the estimate in Table 1 and includes untabulated *Firm*, *Industry* x *State*, and *Industry* x *Year* fixed effects. Column (2) includes untabulated *Firm*, *Industry*, *State*, and *Year* fixed effects. Column (3) includes untabulated *Firm*, *State*, and *Year* fixed effects. Column (4) includes untabulated *State* and *Year* fixed effects. Sample of 56,516 firm-years from 1996 to 2008 for workplace smoking laws; and 108,381 firm-years from 1995 to 2017 for gun laws. Industries are defined using two-digit SIC codes. *% variation absorbed by fixed effects* is the adjusted R^2 from a regression of the independent variable on the respective fixed effects. *t*-statistics appear in parentheses and are clustered by firm. *, **, *** indicate statistical significance (two-sided) at the 0.1, 0.05, and 0.01 levels, respectively.

4.3.2. *Potential to increase fragility of results*

When high dimensional fixed effects absorb an extremely high level of variation in the independent variable of interest (e.g., 99%), the remaining variation used to estimate the coefficient of interest may depend on only a small handful of observations—even when there are hundreds of thousands or millions of observations. As a result, the number of observations in the regression can provide a misleading sense of the amount of variation used to estimate the coefficient of interest (e.g., deHaan, 2020). It can also create a multicollinearity problem. In the extreme case, as the absorption rate increases to 100%, the independent variable of interest approaches a linear combination of the fixed effects, in the extreme the regression is not estimable and perfectly collinear. The problems of multicollinearity are well-known and covered in standard econometric texts (e.g., Belsley et al., 1980): regression results are fragile and coefficient estimates can swing wildly (in either direction) based on small perturbations in the included variables and sample composition.

We illustrate this point using the data from Armstrong et al. (2019, AGHT). AGHT estimate a regression of R&D expense (*RiskyInvest*$_{t+1}$) on the CEO's combined federal and state marginal tax rate (*ManagerRate*$_t$). AGHT report results from at least five different fixed effect specifications: (i) including *Year* fixed effects, (ii) including *Year* and *State* fixed effects, (iii) including *Year*, *State*, and *Firm* fixed effects, (iv) including *Year*, *State*, *Firm*, and *Manager* fixed effects, and (v) including *Industry* x *Year*, *State*, *Firm*, and *Manager* fixed effects.[26]

The results of AGHT are summarized in Panel A of Table 4. Below each regression specification we report two statistics: (i) the percentage of variation in the independent variable that was absorbed by the fixed effects (i.e., the adjusted R^2 from a regression of the independent variable on the fixed effects) and (ii) the variance inflation factor (VIF) for the variable of interest, which is a commonly used diagnostic of multicollinearity.

Table 4
Fixed effect regressions & multicollinearity: Application to Armstrong et al. (2019).

Panel A. Determining Potentially Problematic Fixed Effects

Dependent Variable:	*RiskyInvest*$_{t+1}$				
	Year effects	*Year, State* effects	*Year, State, Firm* effects	*Year, State, Firm, Manager* effects	*Industry-Year, State, Firm, Manager* effects
Variable	(1)	(2)	(3)	(4)	(5)
ManagerRate$_t$	0.203*** (2.74)	0.254*** (3.00)	0.347*** (6.51)	0.237*** (3.89)	0.290*** (3.72)
Controls	yes	yes	yes	yes	yes
Year Fixed Effects	yes	yes	yes	yes	no
State Fixed Effects	no	yes	yes	yes	yes
Firm Fixed Effects	no	no	yes	yes	yes
Manager Fixed Effects	no	no	no	yes	yes
Industry x Year Fixed Effects	no	no	no	no	yes
N	16,490	16,489	16,231	15,461	15,324
% variation in *ManagerRate*$_t$ absorbed by fixed effects	55.7	98.8	98.9	99.2	99.2
VIF	5.45	98.26	121.79	176.84	197.4

Panel B. Consequence of Problematic Fixed Effect Specifications: Fragile Results

Dependent Variable:	*RiskyInvest*$_{t+1}$				
	Year effects	*Year, State,* effects	*Year, State, Firm,* effects	*Year, State, Firm, Manager* effects	*Industry-Year, State, Firm, Manager* effects
Variable	(1)	(2)	(3)	(4)	(5)
ManagerRate$_t$	0.18** (2.54)	−0.01 (−0.08)	0.11 (1.18)	0.03 (0.21)	0.07 (0.57)
Controls	yes	yes	yes	yes	yes
Year Fixed Effects	yes	yes	yes	yes	no
State Fixed Effects	no	yes	yes	yes	yes
Firm Fixed Effects	no	no	yes	yes	yes
Manager Fixed Effects	no	no	no	yes	yes
Industry x Year Fixed Effects	no	no	no	no	yes
N	16,480	16,479	16,221	15,451	15,314

This table summarize the results from the fixed effect regressions reported in Armstrong et al. (2019) and assesses their sensitivity to multicollinearity. Panel A reproduces the results from five fixed effect specifications reported in Table 4 of Armstrong et al. (2019). The dependent variable is R&D expense (*RiskyInvest*$_{t+1}$) and the independent variable is the CEO's combined federal and state marginal tax rate (*ManagerRate*$_t$). We reproduce the results from Armstrong et al. (2019) in each column, and additionally report the amount of variation in the independent variable of interest that is absorbed by the respective fixed effect structure, and variance inflation factor. Panel B reports results from reproducing the results in Armstrong et al. (2019) after dropping ten observations. *% variation absorbed by fixed effects* is the adjusted R^2 from a regression of the independent variable on the respective fixed effects, and *VIF* is the variance inflation factor for the independent variable of interest. *t*-statistics appear in parentheses and are clustered by firm. *, **, *** indicate statistical significance (two-sided) at the 0.1, 0.05, and 0.01 levels, respectively.

[26] Many of these specifications were added at the behest of the reviewer.

Because the independent variable of interest is the applicable marginal tax rate, it should not be surprising that these fixed effect structures absorb a significant amount of its variation. Column (5) reports that the combination of *Industry* x *Year*, *State*, *Firm*, and *Manager* fixed effects absorbs 99.2% of the variation in tax rates—meaning the coefficient of interest is estimated using only 0.8% of the variation in the independent variable. Given the extreme level of absorption, it should also not be surprising to see an elevated variance inflation factor. Column (5) shows that when *Industry* x *Year*, *State*, *Firm*, and *Manager* effects are included the VIF is 197.4, approximately twice that when only *Year* and *State* effects are included (column (2) VIF = 98.26), and 40 times that when only *Year* effects are included (column (1) VIF = 5.45). For reference, Belsley et al. (1980) suggest strong multicollinearity is present when the VIF is greater than 10.

As a result of the extreme levels of induced multicollinearity, we expect the regression results to be driven by the presence of only a handful of observations. Panel B of Table 4 repeats the tests in Panel A after removing ten observations from the sample of over 16,000 observations. We find that the specifications with extreme levels of absorption—and consequently high multicollinearity—are very fragile.[27] After dropping ten observations, we find the coefficient on the variable of interest is markedly lower and statistically insignificant in four of the five specifications. This example illustrates how—in the presence of high-dimensional fixed effects—the number of observations in the sample can give a misleading sense of the amount of variation used to estimate the effect and can increase the fragility of the regression to a small handful of observations. Our findings regarding extreme levels of absorption and frailty of results are not unique to the fixed effect specifications in AGHT. As discussed above, similar extreme levels of absorption have been documented in the context of gun laws (Table 5), universal demand laws (e.g., Donelson et al., 2021), and state anti-compete laws (Jennings et al., 2022).

In sum, fixed effects can be a powerful tool for mitigating concerns about correlated omitted variables. However, like every research design choice, fixed effects are not without tradeoffs, and thus—like all control variables—should be explicitly motivated. We submit that these extreme levels of absorption are fairly commonplace, and encourage researchers who employ high-dimensional fixed effects to (1) triangulate results across different fixed effect structures, (2) report variance inflation factors on their variable of interest, and (3) report the amount of variation in the variable of interest that is absorbed by the fixed effects (e.g., the R^2 from a regression of the independent variable on fixed effects).[28]

A common practical issue arises when researchers do not know the true data generating process, and find that the empirical results hinge critically on the choice of fixed effect specification: in the extreme, the coefficient on the variable of interest flips sign between different fixed effect specifications. Here, we make two related points. (1) The specification that confirms the researcher's predictions is not necessarily the correct specification. Researchers should be aware of, and alert to the possibility of confirmation bias.[29] (2) Sign flips and other sensitivities should be something to be investigated—which may lead to novel predictions—rather than dismissed.[30] We have suggested two common diagnostic tests—the variation inflation factor and the R^2 from a regression of the independent variable on fixed effects—for assessing whether a particular specification is problematic. Armed with these diagnostic tests, researchers should be able to better justify and assess their research design choices as it relates to fixed effects.[31]

5. Can non-experimental evidence facilitate causal inference?

Given the discussion of the strengths of *quasi-experimental* methods described in Section 3, and the implementation challenges described in Section 4, we next seek to understand whether and how *non-experimental* methods and research designs—i.e., methods and designs that do not purport to emulate random assignment (and that are potentially confounded by endogeneity)—can nevertheless facilitate causal inferences. In particular, we provide a conceptual framework for researchers to assess whether and when accounting settings *without* as-if random variation are useful in addressing causal questions. We offer the view that there are some causal questions for which quasi-experiments are impractical or ill-suited.

In Section 5.1, we ask the question: does one need an experimental or quasi-experimental setting to address a causal question? We begin by discussing how—in the absence of random assignment to treatment and control groups—non-experimental evidence is particularly important for addressing causal questions. We provide several examples in which non-experimental evidence provides insight on highly practical questions that are inherently causal, but where quasi-experiments are not feasible. We use the abduction framework from Heckman and Singer (2017) to compare and contrast non-experimental and quasi-experimental approaches to addressing causal questions.

[27] Armstrong et al. (2019) report results for specifications with and without fixed effects. Inferences from the former are unaffected by these issues.

[28] deHaan (2020) provides a post-estimate Stata command, *sumhdfe*, that reports the latter.

[29] Pre-registration of research design and specification eliminates this bias (e.g., Bloomfield et al., 2018).

[30] See Bianchi et al. (2021) for an example of a paper that is transparent with respect to results being sensitive to the inclusion of firm effects, and how they alter their interpretation of the evidence as a result, i.e., that the evidence is consistent with the mafia selecting certain types of firms, rather than time-series variation in mafia involvement.

[31] For those readers interested in a more detailed treatment of the various strengths and weaknesses of fixed effects see the following literature: (i) Angrist and Pischke (2008) discuss issues when both fixed effects and lag values of the dependent variable are included in the same model; (ii) Grieser and Hadlock (2019) discuss the strict exogeneity assumption of fixed effects; (iii) Jennings et al. (2022) discuss fixed effects in the context of measurement error; (iv) Berg et al. (2021) discusses how fixed effects exacerbate coefficient bias when seeking to estimate spillover effects, and (v) Whited et al. (2021) discuss fixed effects in the more general context of the "bad controls" problem discussed in Angrist and Pischke (2008).

In Section 5.2, we ask the question: how can non-experimental and quasi-experimental evidence be combined in the context of a single study to identify causal mechanisms? As noted in Heckman and Singer (2017), the two types of methodologies are not mutually exclusive and can be complementary, as evidenced by a growing number of studies using both experimental and non-experimental designs to triangulate inferences and provide evidence of causal mechanisms. We discuss how the *combination* of quasi-experimental and non-experimental evidence can be, and has been, used to identify causal mechanisms in the literature.

In Section 5.3, we ask the question: should we prioritize quasi-experimental evidence over non-experimental evidence when the two conflict? We make the point that conflicting evidence is often an indicator that the phenomenon being studied is deeper and more complex than one might perceive and, thus, can often yield new insights. We provide an example from the literature in which dismissing a study offering non-experimental evidence as suffering from endogeneity bias—in favor of a study offering conflicting quasi-experimental evidence—would provide inferences that are at best incomplete and, at worst, misleading. We caution against dismissing non-experimental evidence.

5.1. A role for non-experimental evidence when quasi-experiments are not practical

There are an infinite number of causal questions, but only a finite number of quasi-experimental settings. For many causal questions, a setting that mimics random assignment between the treatment and control groups is simply not available. Researchers are thus faced with a choice. On the one hand, researchers could restrict their attention to the set of questions for which as-if random assignment is available. This approach espouses the belief that—in the absence of the experimental or quasi-experimental ideal—the causal question is not worth addressing. Perhaps reflecting this, in the words of Angrist and Pischke (2008, p.5): “research questions that cannot be answered by any experiment are FUQs: fundamentally unidentified questions” and associated attempts at causal inferences are “FUQ'd” (p. 7). From a philosophical perspective, there is merit to the approach of restricting consideration of questions to those that can be answered using experiments or quasi-experiments.[32] Conceivably, it restricts attention to questions that can be answered with some minimal, standard level of precision. The tradeoff is that this approach necessarily limits the literature to the set of causal questions for which quasi-experimental settings are available. As a result, the literature may miss important opportunities for growth.

On the other hand, researchers could explore causal questions even if a setting that mimics random assignment is not available, and caveat inferences appropriately. Researchers taking this approach might seek to provide evidence “consistent with” or “suggesting” causality, but that is not definitive.[33] This approach recognizes both the provisional nature of our knowledge (i.e., the process of learning is an ongoing endeavor), in addition to the inherent limitations of non-experimental methods and data. The danger in this approach is that alternative explanations might be harder to rule out.

In this section, we illustrate when the latter approach might be appropriate. We discuss settings and circumstances in which non-experimental evidence and data—even data with clear endogeneity concerns—can nevertheless shed light on causal questions. This discussion reflects the view that no single study is definitive, and alternative explanations can be ruled out over time through a collection of studies on the topic. We view this approach as consistent with the basic principles of the scientific process, as illustrated by Fig. 6, which is agnostic with respect to research method.

Within the accounting literature, examples abound of highly practical and plausibly interesting causal questions for which a quasi-experimental setting is not available. Below, we consider two papers that illustrate the value of non-experimental evidence in answering causal questions. In each case, the evidence provides compelling insight on a causal question not available from a quasi-experiment. Note that the ability to draw causal inferences is not binary (i.e., “yes causal inference,” or “no causal inference”), but rather a continuum that reflects the precision of the evidence on the causal question (i.e., from 0 to 0.999). Consequently, even if one cannot draw causal inferences with absolute certainty, one can still provide compelling evidence on causal questions. Here, we take the view that it is better to have a stream of imperfect papers with complementary research designs on an important causal question for which a quasi-experiment is not available than to ignore such questions for lack of quasi-experimental evidence.

Our first example is drawn from a classic paper in the executive compensation literature, for which no quasi-experiment with as-if random variation is available. Healy (1985) examines whether earnings-based bonus contracts incentivize managers to manipulate earnings. Note that the research question is inherently causal (e.g., the paper is entitled “The Effect of Bonus Scheme on Accounting Decisions”) and is very practical (e.g., responsible boards care about the incentive effects of bonus contracts). Healy (1985) finds evidence of a strong association between the reporting incentives of managers' bonus contracts and accruals, and that significant changes in accounting procedures follow the adoption of earnings-based bonus plans. Healy is careful not to over-interpret the results and uses language to describe the evidence as “consistent with” (or “suggests”) a causal effect.

There are significant concerns about endogeneity in this setting: in equilibrium, a rational board would anticipate the manager's response to an earnings-based bonus plan. Consequently, the decision to provide the bonus plan, its terms, and

[32] This approach dominates much of the applied microeconomics literature.

[33] In many academic disciplines, random assignment is not viewed as necessary or sufficient for causal inference. For example, based on compelling theory and empirical evidence that does not feature random assignment, the medical field has nonetheless concluded that smoking causes cancer.

the accounting response are all endogenous and jointly determined in equilibrium. This endogeneity challenge notwithstanding, it is difficult to imagine how the literature on executive compensation contracts—and contracting more generally—would have progressed if it restricted itself to studying settings with as-if random variation in contract terms. Thus, although we recognize the benefits of quasi-experiments for estimating causal effects, because as-if random variation in contracts is not available (or at least incredibly rare), non-experimental evidence is particularly valuable for addressing causal questions in this and other literatures. Indeed, despite endogeneity concerns, the theory and empirical evidence advanced in Healy (1985) has shaped the development of subsequent literature on incentive compensation (e.g., Bloomfield et al., 2021; Bushman, 2021).

Our second example is drawn from the literature on tax avoidance, where the question itself is conditioned on understanding mechanisms readily present in the real world. Asay et al. (2021) examine how a particular tax avoidance strategy affects consumer behavior. This question is inherently causal: does tax avoidance cause consumer boycotts? Using a combination of survey evidence, UPC-level product sales, and Robinhood data, Asay et al. (2021) find no evidence of a retail response to tax avoidance.

As a practical thought experiment, suppose the CEO of Starbucks asks whether their latest tax avoidance strategy will generate a consumer boycott. What analysis would one conduct to answer this question? Perhaps the ideal experiment would be one in which researchers would induce random variation in tax avoidance in a sample of firms, track consumers' learning of tax avoidance, and examine differences in their subsequent purchases. Alas, this is not feasible in practice. As an alternative experiment, one might randomly inform consumers about firms' existing tax avoidance practices (e.g., by mailing the information to a randomly selected sample of consumers), and subsequently observe these consumers' product purchase behavior. However, this would not be representative of how consumers become informed about firms' tax avoidance behavior in the real world and would provide at best partial evidence to answer the CEO's question.

As an alternative to experimental evidence, we might gather data on what happened in the past when the company or its peers applied aggressive tax avoidance strategies. Did media coverage change? Did product sales change? Perhaps we would even survey consumers about whether they knew about the firm's tax avoidance and, if so, how they responded. In any of these cases, the evidence would *not* be premised on random or as-if random variation in tax avoidance. The level of tax avoidance, media coverage, and consumer response are all endogenous choices of various agents.

Asay et al. (2021) adopt the latter approach and employ a variety of non-experimental methods to address their causal question; an approach often referred to as "identification through triangulation." Even though Asay et al. (2021) do not have a quasi-experiment, by triangulating inferences across multiple non-experimental designs and settings, they provide particularly compelling causal evidence that the average consumer in their sample does not care about tax avoidance strategies when making their purchase decisions.

Heckman and Singer (2017) refer to the notion of identification through triangulation as the "abductive" model of learning, and discuss how this approach to causal inference is not found in standard econometrics textbooks:

> *The abductive model for learning from data follows more closely the methods of Sherlock Holmes than those of textbook econometrics. The Sherlock Holmes approach uses many different kinds of clues of varying trustworthiness, weights them, puts them together, and tells a plausible story of the ensemble. [...] The abductive approach to empirical economics advocates a process and a mindset. It privileges no source of data, style of research or mode of inference for learning about the economy provided the analyst produces useful knowledge that survives critical public scrutiny. It values factually-rich descriptions as major sources of knowledge. It favors using every piece of available information, despite varying trustworthiness of parts of it. However, it asks that analysts report, in a public way, how they weigh the diverse evidence. It encourages readers of such studies to form their own opinions and justify their own weights. It recognizes the provisional nature of knowledge … the abductive mode of thought challenges the currently influential framework of the "identification problem," which underlies both treatment effect and structural approaches*. (p. 298–299)

This model of learning from data approaches causal inferences through triangulation—recognizing the strengths and weaknesses of each method, the value of multiple research designs, measures, and tests, and the transparent reporting thereof. This view recognizes that the researcher does not know the true underlying data generating process, and therefore cannot unambiguously know the correct test specification. Therefore, it prioritizes looking for patterns across multiple research methodologies, datasets, test specifications, and measures—rather than placing all evidentiary weight on a singular methodology or test specification.

Fig. 11 seeks to graphically illustrate the distinction between what Heckman and Singer (2017) refer to as the "abductive" and "treatment effect" approaches to causal inferences. This figure illustrates the abductive approach, in which the researcher is interested in the causal relation between X and Y, and articulates alternative explanations represented by Z_1, Z_2, and Z_3. Note that these alternatives could be statistical (e.g., measurement error) or economic (e.g., an alternative theoretical mechanism). The researcher's task is to conduct a set of tests that seek to rule out these explanations. This lends itself to the "specific identification" approach to omitted variables described in Section 3.1.3. To quote Sherlock Holmes, "when you have eliminated the impossible, whatever remains, however improbable, must be the truth" (Conan Doyle, 1890, The Sign of the Four).[34]

[34] The Stanford Encyclopedia of Philosophy refers to this notion as "inference to the best explanation" (Douven, 2021, p. 1).

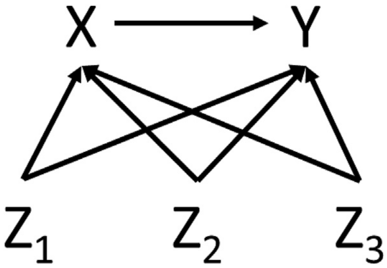


Fig. 11. Abductive Approach. This figure seeks to illustrate what Heckman and Singer (2017) refer to as the "abductive" approach discussed in Section 5.1, which articulates alternative explanations for the relation between *X* and *Y* (represented here by Z_1, Z_2, and Z_3) and then seeks to systematically rule them out.

The treatment effect approach to causal inferences is very different. Under this approach, the researcher is interested in estimating the causal relation between X and Y, and seeks to identify a setting, or quasi-experiment with random variation in X. To the extent the researcher is successful, there are no conceivable alternative explanations. Note that this approach does not require the researcher to specify alternative explanations for the relation between X and Y, but rather emphasizes the random variation provided by the setting. If the variation being studied is *not* in fact random, then it invalidates this approach to causal inference. As a result, many research studies applying this approach spend a great deal of time and effort validating the random nature of the variation provided by the setting, and comparatively little time and effort developing alternative explanations or theories.

Both approaches are equally valid for providing causal inferences. In practice, which approach is "best" depends on (i) the availability of quasi-experiments, (ii) the richness of the underlying theory for specifying alternatives, and (iii) the precision with which the researcher can rule them out. For example, if the alternatives cannot be specified or credibly ruled out, and there is an available quasi-experiment—the treatment effects approach may be the most appropriate.

Of course, triangulation and quasi-experimental designs are not mutually exclusive. Table 5 illustrates what the notion of triangulation might look like in the context of a quasi-experiment. Table 5 presents results from repeating the tests in Table 1 regarding the staggered adoption of workplace smoking laws and gun restrictions but using two alternative measures of voluntary disclosure common in the literature—firm-initiated press releases and Forms 8-K (e.g., Guay et al., 2016; He and Plumlee, 2020). Table 5 suggests the correlations regarding workplace smoking laws and gun restrictions in Table 1 are sensitive to the measure of voluntary disclosure. We find no evidence of a relation between the laws and these two alternative measures of voluntary disclosure.

Thus, because we have no *a priori* reason to prioritize one measure of voluntary disclosure (e.g., management forecasts) over another (8-K filings and press releases), one might be skeptical of interpreting the results in Table 1 as indicative of a causal effect of these state laws on voluntary disclosure. This is especially the case when researchers have access to dozens of measures of voluntary disclosure. The fact that relations are specific to one measure, when there is no *a priori* reason to expect them to be, suggests the results might be spurious—occurring randomly in the data and without any particular meaning. Thus, unless one triangulated results across multiple measures of the theoretical construct of interest, one would have no way of knowing that the results were so fragile that they hinged on a specific measure of voluntary disclosure.

5.2. Combining quasi-experimental and non-experimental evidence to identify causal mechanisms

In this section, we discuss how the *combination* of quasi-experimental and non-experimental evidence can be, and has been, used to identify causal mechanisms in the literature.

We follow Kahn and Whited (2018) and draw a distinction between "estimation" and "identification" in the context of causal inference. This distinction is important, because estimation of a causal effect is a statistical process entailing as-if random variation, whereas identification involves drawing inferences about the source of the underlying causal effect.[35] Successful estimation answers the question: *what* is the effect? Successful identification answers the question: *why* is the effect? We illustrate this distinction in Fig. 12, which shows a causal diagram linking two constructs, X and Y. In this diagram, X causally affects Y through A and B, *which are both unobservable*. Using (as-if) random variation in X, it is possible to empirically *estimate* the causal effect of X on Y (i.e., *what* is the effect), without *identifying* the underlying mechanisms, A and B, that produce the effect (i.e., *why* is the effect).

[35] Kahn and Whited (2018) recognize that many studies commingle estimation and identification: "*[I]t [is] easy to confuse identification with the establishment of causality through exogenous variation. In fact,* Angrist and Pischke (2008) *present the issue of identification entirely as a search for an approximation to an ideal experiment*" (p. 3).

Table 5
Triangulation: Using alternative measures of voluntary disclosure.

Panel A. Staggered Adoption of Workplace Smoking Laws

Dependent Variable:	Table 1 results *VolDisc* = Number of Management Forecasts	Alt. Measure *VolDisc* = Number of 8-K Filings	Alt. Measure *VolDisc* = Number of Firm-Initiated Press Releases
Variable	(1)	(2)	(3)
Smoking Law$_t$	−0.48**	−0.05	−0.25
	(−2.52)	(−0.33)	(−0.60)
Firm Fixed Effects	yes	yes	yes
Industry x State Fixed Effects	yes	yes	yes
Industry x Year Fixed Effects	yes	yes	yes
Adj R^2	51.8	58.38	91.07
N	56,516	56,516	19,477

Panel B. Staggered Adoption of State Gun Laws

Dependent Variable:	Table 1 results *VolDisc* = Number of Management Forecasts	Alt. Measure *VolDisc* = Number of 8-K Filings	Alt. Measure *VolDisc* = Number of Firm-Initiated Press Releases
Variable	(1)	(2)	(3)
Gun Law$_t$	−0.06***	−0.02	0.91
	(−2.78)	(−1.53)	(1.32)
Firm Fixed Effects	yes	yes	yes
Industry x State Fixed Effects	yes	yes	yes
Industry x Year Fixed Effects	yes	yes	yes
Adj R^2	60.2	56.96	78.89
N	108,381	108,381	57,300

This table presents results from repeating the analysis in Table 1 using two alternative measures of voluntary disclosure, the number of 8-K filings during the year (excluding Item 2.02, Results of Operations) and the number of firm-initiated press releases during the year. See Guay et al. (2016) and He and Plumlee (2020). Column (1) repeats the estimate in Table 1, columns (2) and (3) present results for the alternative measures of voluntary disclosure. Each column includes untabulated *Firm*, *Industry* x *State*, and *Industry* x *Year* fixed effects. Sample of 56,516 firm-years from 1996 to 2008 for workplace smoking laws; and 108,381 firm-years from 1995 to 2017 for gun laws. Data on press releases begins in 2004. Industries are defined using two-digit SIC codes. *t*-statistics appear in parentheses and are clustered by firm. *, **, *** indicate statistical significance (two-sided) at the 0.1, 0.05, and 0.01 levels, respectively.

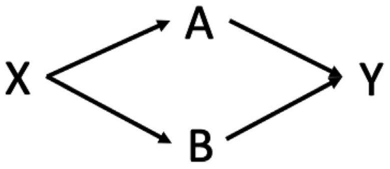


Fig. 12. Estimation vs. Identification. This figure presents a hypothetical causal diagram that illustrates the causal path between X and Y discussed in Section 5.2. Each letter represents a distinct theoretical construct. In our example, X and Y are observable and A and B are not.

Hypothetically, if one were interested not just in estimating the causal effect of X on Y, but also in understanding why X causes Y (i.e., identification of the causal mechanism), one would also need random variation in the intermediate constructs along the causal chain. That is, one would need random variation in X to estimate its effect on A and B, and random variation in A and B to estimate their effect on Y. Moreover, to rule out that there is no direct effect of X on Y (independent of A and B), one would need to show that the causal effect of X on Y does not exist after controlling for A and B. This set of tests would identify the entire causal chain between X and Y.

Given the difficulty in finding as-if random variation in one of the theoretical constructs—let alone all three of them—and the difficulty in developing sufficient theory to specify the entire causal chain, the task of successfully identifying the causal mechanism(s) underlying the causal effect of X on Y is extremely challenging—especially in the context of a single study. Although one view of this example is that it is complex, and involves multiple forces and channels, it is likely a vast oversimplification of the complex forces at work in capital markets and corporate decisions.

In the absence of as-if random variation in every link in the causal chain, one can combine non-experimental methods with quasi-experimental methods in an effort to triangulate inferences—to estimate the causal effect, while providing evidence consistent with (but not definitive of) a particular channel. Below, we illustrate how non-experimental evidence—even from data with pervasive endogeneity issues—can be combined with quasi-experimental evidence to facilitate identification and discuss how this combination is implemented in practice.

Consider how a researcher might proceed to test for a particular causal channel in the diagram indicated by Fig. 12 in the absence of as-if random variation in every link in the causal chain. Suppose the researcher finds a setting with random variation in X to estimate the causal effect of X on Y. The researcher conjectures that the underlying mechanism operates through A, yet A is unobserved. To test the causal channel, the researcher could estimate the causal effect of X on Y in a setting in which theory suggests X *does not* affect A and compare it to the estimated causal effect in a setting where theory suggests X *does* affect A. The researcher could then compare the estimated causal effects across these two settings. This amounts to a cross-sectional test, or a triple differences design (DiDiD), and allows for what is known as "heterogenous treatment effects"—the notion that the causal effect of X on Y varies with characteristics of the underlying observations. Evidence that the causal effect of X on Y varies in the predicted manner can provide compelling evidence of the causal mechanism. This methodology effectively combines the cross-sectional interaction design and the DiD method described in Section 3.

In practice, if the variable used to partition the sample and estimate cross-sectional differences in causal effects is itself endogenous, then the tests are inherently a departure from the experimental ideal (e.g., it is well known that partitioning by, or interacting with, an endogenous variable introduces bias; e.g., Wooldridge, 2000). Nevertheless, despite the endogenous nature of the partitioning variable, the tests can conceptually facilitate identification of the causal mechanism—and are commonly used in the literature. We briefly discuss two seminal studies that partition the sample on endogenous variables to provide evidence on causal mechanisms.

In our first example, Balakrishnan et al. (2014) examine whether managers seek to offset an exogenous reduction in liquidity following analyst brokerage closures by increasing voluntary disclosure and, if so, whether these efforts lead to a recovery in liquidity. In the study's theoretical framework, voluntary disclosure and liquidity are explicitly endogenous: there is a drop in liquidity because of an exogenous event (brokerage closure), and managers choose to issue disclosures to provide more information and offset the anticipated drop in liquidity. Recognizing that the voluntary disclosure decision is endogenous, Balakrishnan et al. (2014) estimate the causal effect of the brokerage closure on liquidity separately in the sample of firms that *did* provide voluntary disclosure, and the sample of firms that *did not* provide voluntary disclosure. For the firms that did provide voluntary disclosure, the study finds that the negative effect of the brokerage closure on liquidity reverses in the subsequent quarter, whereas the effect on the firms that did not provide voluntary disclosure persists. Note that the partitioning variable in this example—the choice of voluntary disclosure—is not as-if random, and is endogenous to liquidity. Nonetheless, this evidence helps rule out alternative explanations and provides compelling evidence on the causal mechanism of the paper's central research question.

In our second example, Christensen et al. (2016) estimate the causal effect of changes in European Union securities regulations on liquidity using variation in the countries' adoption dates of the securities regulation. The study finds significant increases in liquidity following countries' adoption of the regulation, and conjectures that the effect of the regulation depends on "endogenous prior conditions" (i.e., existing country-level attributes). They consider two alternative hypotheses: (i) the effect of new securities regulations is larger in countries where existing securities regulations are weak; and (ii) the same institutional, market, and political forces that limit the effectiveness of existing securities regulations also limit the effectiveness of new securities regulations. The study finds evidence of the latter: the causal effect of new securities regulations is weaker in countries that already have weak securities regulations. Similar to Balakrishnan et al. (2014), the variable(s) used to partition the sample and estimate heterogeneous treatment effects are proxies for the countries' existing conditions and therefore are endogenous. This endogeneity notwithstanding, the tests provide compelling evidence on the conditions that the causal mechanisms depend on and represent a valuable part of the study's contribution.

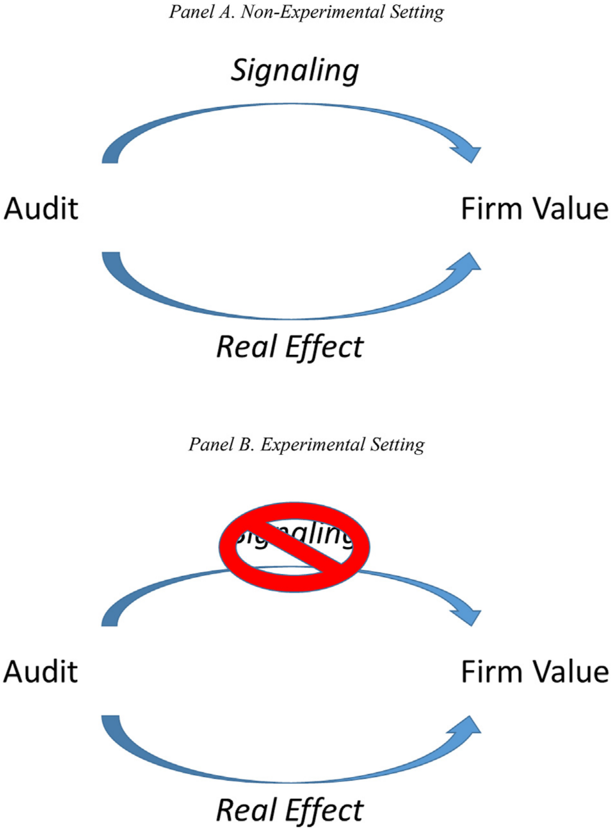
Although there are many other examples we could offer, we will not belabor the point. Even non-experimental data—and data with pervasive endogeneity issues—can be combined with compelling theory and quasi-experimental evidence to provide stronger and more nuanced inferences.

5.3. When quasi-experimental and non-experimental evidence conflict

Although the combination of the two methods—and consistency in results across them—can provide credible inferences, occasionally evidence across these methods can conflict. In such cases, how should researchers resolve these conflicts? For instance, should a study offering non-experimental evidence that suffers from endogeneity issues be dismissed in favor of a study offering quasi-experimental evidence that does not suffer from such issues? In this section, we caution against always prioritizing inferences from quasi-experimental settings over non-experimental settings and dismissing the latter as "subject to endogeneity." Seemingly conflicting evidence is often an indicator that the phenomenon being studied is deeper and more complex than one might perceive and can often yield new insights.

We use a simple example drawn from the audit literature to illustrate the potential for conflicting evidence and the danger in dismissing inferences from any one approach when the evidence from different approaches conflicts. We show that considering either approach in isolation would provide inferences that are at best incomplete and at worst misleading. Throughout this example, we discuss how a combination of types of evidence is often necessary to fully understand real-world phenomena and caution against a strict preference for one type of evidence over another.

Our example relies on the notion that managers often undertake a specific action with the explicit purpose of communicating—or "signaling"—their private information. Signaling is a pervasive phenomenon in the accounting literature, which has its origins with Spence (1978). In the accounting literature, signaling has been used to explain patterns in voluntary disclosure (Trueman, 1986), accounting choices (Myers, 1989), dividend payouts (DeAngelo et al., 1996), insider stock purchases (Armstrong et al., 2021a,b), and corporate social responsibility (Lys et al., 2015), among others. We draw our example from the literature on the signaling value of obtaining an audit (Kausar et al., 2016).



Panel A. Non-Experimental Setting

Panel B. Experimental Setting

Fig. 13. Mechanisms for Audits on Firm Value. This figure illustrates two channels through which audits can affect firm value. Panel A shows the channels at work in the non-experimental setting discussed in Section 5.3, and Panel B shows the channels at work in the experimental setting discussed in Section 5.3.

Consider the question of whether audits affect firm value. To answer this question, one might study a setting in which firms can choose to obtain an audit. In this setting, there are two conceivable channels through which audits can affect firm value (Minnis, 2011; Kausar et al., 2016). First, firms can choose to obtain an audit to signal positive private information about future cash flows to investors. We refer to this as the "signaling channel." Second, the choice of audit can have a real effect on the level of investment and firm cash flow independent of signaling. We refer to this as the "real effects" channel (e.g., Roychowdhury et al., 2019). Panel A of Fig. 13 illustrates these two channels. Theoretically, both channels would generate a positive correlation between the audit choice and future cash flows and, as a result, a positive market reaction to the disclosure of the choice to obtain an audit.[36] Notwithstanding the inherent endogeneity concerns of this non-experimental setting, or the inability to disentangle the two channels, the positive market reaction would suggest that investors revise their beliefs about firm value when the audit choice is disclosed, which is consistent with the audit choice causing changes in firm value.

As another approach to addressing this question, one might study a setting in which a regulator randomly assigns firms to "audit" and "no audit" groups, but otherwise identical to the above. This would appear to be an ideal setting to study the causal effect of audits on firm valuation. However, in this case, because the choice of whether to obtain an audit is removed, the firm cannot use the audit to signal future cash flows to investors. Consequently, in the presence of random assignment, the signaling channel is eliminated and only the real effect channel remains.[37] Panel B of Fig. 13 illustrates this scenario. In this setting, the experiment allows us to precisely estimate the causal effect of the audit on firm value through the real effects channel but is not capable of providing inferences regarding the signaling channel.

Now, consider a circumstance in which the findings *seemingly* conflict across these two approaches. Suppose that in the non-experimental setting, we find that investors react positively to learning the firm will be audited, but in the experimental setting we find that investors do not react at all. In this circumstance, the estimates of the causal effect clearly conflict across the two approaches. Given this conflict, and the lack of random variation in the non-experimental approach, it might be tempting to dismiss the causal inferences from the first approach as confounded by endogeneity. However, because the two

[36] In a signaling equilibrium, investors rationally interpret the audit choice as a signal of the manager's positive private information about future cash flows, which generates a positive market reaction to the choice to obtain an audit.

[37] A common theme throughout the literature is that, to signal private information, the action is necessarily an endogenous choice. Consequently, as-if random variation in the action, by construction, removes the element of choice from the decision-maker and eliminates any signaling ability.

approaches feature different causal channels, the inferences are not necessarily in conflict. It is only when the evidence from the non-experimental approach is taken in conjunction with the quasi-experimental approach—i.e., when the signaling channel is eliminated—that we learn that the relation in the former is attributable to signaling rather than real effects.

This example illustrates that if one were to simply dismiss the non-experimental evidence as confounded by endogeneity, there would be an important loss of information: we would be unaware of the existence and importance of the signaling channel. When interpreted in isolation, evidence from either approach provides an incomplete picture; it is only in comparing and contrasting evidence across the two approaches that the full picture emerges. As a result, we caution against a strict preference for one type of evidence (e.g., quasi-experimental) over another (e.g., non-experimental). We learn more from the combination of several studies than from each individual study in isolation. Practically speaking, when researchers try to assess how to reconcile or dismiss conflicting evidence from non-experimental and quasi-experimental settings, we recommend they consider whether there are different economic forces at work across the two settings.

6. Conclusion

We conclude with a brief summary and an important caveat regarding practical considerations for researchers interested in drawing causal inferences. Our review of the accounting literature leads us to conclude that drawing reliable causal inferences is very challenging—and much more challenging than simply the choice of method. Reliable causal inferences require compelling economic theory, methods that make assumptions that comport with the institutional setting being studied, and a plethora of robustness tests to triangulate inferences across (often implicit) theoretical assumptions.

Despite their best efforts, sometimes researchers cannot find a quasi-experiment, in these cases, it is acceptable—even desirable—to provide evidence using non-experimental methods with appropriate caveats. We caution against the idea that one should restrict attention to only those causal questions for which there are settings conducive to quasi-experimental methods, and we caution against the dogmatic application of any method. Regardless of whether a researcher employs a quasi-experimental method, because the true data generating process is unknown, generalizable causal inferences necessarily require using more than one regression specification, and more than one measure of a specific theoretical construct.

We caveat that although our review focuses on the methods accounting researchers use to draw causal inferences, there are at least three important caveats to our survey.

(1) The choices of method and setting are irrelevant if empirical results do not replicate (Hail et al., 2020). Replicability is the minimum quality standard of any credible scientific work. In some sense, a poorly executed study that *is* replicable provides more information, and a greater building block for future work, than a seemingly well-executed study that *is not* replicable. Angrist and Pischke (2010) refer to the use of quasi-experiments as a “credibility revolution” in causal inference. However, without replicability, there can be no credibility.

(2) Throughout our review, we have encouraged transparent reporting of results across multiple specifications, settings, and methods. In making these recommendations, we assume researchers, reviewers, and editors are just as satisfied with reporting null results as with reporting positive results: that the incentives for transparency are greater than the incentives for selective reporting. However, many scholars have observed the opposite (Brodeur et al., 2016; Ohlson, 2022). Several studies in statistics (Gelman and Loken, 2014, 2017), psychology (Simmons et al., 2011), the sciences (Smaldino and McElreath, 2016), and economics (Brodeur et al., 2020) have documented how researchers' incentives to find positive results and publish their paper can distort causal inferences—either through selective reporting or ex post justification of research design choices. These are not easy issues to tackle, and are beyond the scope of our review. Nevertheless, they are of utmost importance when seeking to draw causal inferences from published research.

(3) Many papers in the accounting literature do not seek to address causal questions, but nonetheless make important contributions. For example, a growing body of academic research consists of ‘forensic studies’ that seek to document patterns in the data that are consistent with suspicious—if not outright illicit—behavior in capital markets.[38] While such studies are beyond the purview of this survey, and are ultimately descriptive, they play an important role in drawing attention to patterns in the data that, at best, are inconsistent with good corporate governance, and, at worst, are evidence of violations of securities laws. This body of work illustrates that while causal inferences are important, they not always necessary to provide a contribution and advance the frontiers of knowledge.

Data availability

Data will be made available upon request. Code is available in the Internet Appendix on SSRN.

[38] Examples include Lie (2005) and Heron and Lie (2007) studying option backdating; Dechow et al., 2015 studying insider trading related to Correspondence with the SEC related to revenue recognition; Heitzman and Klasa (2021) studying abnormal options trading activity around non-public merger negotiations; Mehta et al. (2021) studying “shadow trading” in peer firms prior to public announcements of focal firms; Blackburne et al. (2021) studying trading around non-public SEC investigations; Bianchi et al. (2021) studying firms with board connections to organized crime; and Haselmann et al. (2021) studying whether banks trade on private information about their borrowers.

Appendix. Fixed Effects and Omitted Variable Bias

We consider a circumstance where a researcher is interested in estimating a regression of an outcome variable (Y) on an independent variable of interest (X), but is concerned about the existence of an unknown correlated omitted variable (Z). We assume the researcher does not know the true data generating process for any of these variables, and that because the omitted variable is unknown to the researcher it cannot be included in the regression—precluding the "specific identification" approach discussed in Section 3.1.

We further assume the unknown correlated omitted variable varies within-group. We solve for the omitted variable bias when group fixed effects are excluded from the regression model, and compare it to the bias when group fixed effects are included in the regression. We show the omitted variable bias is larger in the latter, and that the bias is increasing in the percentage of variation in the independent variable that is absorbed by firm fixed effects.

Assume the true data generating process (unknown to the researcher) is given by:

$$Y_{i,t} = \beta X_{i,t} + \theta Z_{i,t} \tag{A1}$$

where i indexes firms and t indexes years, and throughout our analysis we assume all variables are mean-zero and normally distributed, i.i.d. We assume the variable of interest, $X_{i,t}$, has two components: a random within-firm component, $x_{i,t}$ and a firm-fixed effect component, F_i:

$$X_{i,t} = x_{i,t} + F_i \tag{A2}$$

The unknown correlated omitted variable $Z_{i,t}$ also has two components. The first component is correlated with the within-firm component in the independent variable of interest, $\rho x_{i,t}$. The second component is uncorrelated with the independent variable of interest, $z_{i,t}$:

$$Z_{i,t} = \rho x_{i,t} + z_{i,t} \tag{A3}$$

Now suppose the researcher estimates two regressions: a regression of $Y_{i,t}$ on $X_{i,t}$ excluding firm effects, and a regression of $Y_{i,t}$ on $X_{i,t}$ including firm fixed effects. Let β^{NoFE} (β^{FE}) denote the slope coefficient in the former (latter) regression. Solving for the respective slope coefficients yields:

$$\beta^{NoFE} = \frac{cov(Y_{i,t}, X_{i,t})}{var(X_{i,t})} = \beta + \frac{\theta cov(Z_{i,t}, X_{i,t})}{var(X_{i,t})} = \beta + \theta\rho\frac{var(x_{i,t})}{var(x_{i,t}) + var(F_i)} \tag{A4}$$

and

$$\beta^{FE} = \frac{cov(Y_{i,t}, x_{i,t})}{var(x_{i,t})} = \beta + \frac{\theta cov(Z_{i,t}, x_{i,t})}{var(x_{i,t})} = \beta + \theta\rho \tag{A5}$$

Equation (A4) shows that the magnitude of the bias when fixed effects are *excluded* from the regression is given by $|\theta\rho|\frac{var(x_{i,t})}{var(x_{i,t})+var(F_i)}$. The latter term in this expression represents the percentage of variation in the independent variable, $X_{i,t}$, that is attributable to $x_{i,t}$, and correlated with the omitted variable. As $x_{i,t}$ becomes a proportionately smaller component of $X_{i,t}$, the omitted variable bias decreases.

Equation (A5) shows that the magnitude of the bias when fixed effects are *included* in the regression is simply $|\theta\rho|$. The intuition for this result is that the inclusion of firm fixed effects in the regression,removes the influence of F_i, and isolates only the within-firm variation in $X_{i,t}$ stemming from $x_{i,t}$. In this setting, $x_{i,t}$ is the portion on the independent variable that is correlated with the omitted variable.

To analyze which specification has larger bias, we can express the difference in magnitude of the omitted variable bias between β^{FE} and β^{NoFE} as:

$$\Delta = |\theta\rho| - |\theta\rho|\frac{var(x_{i,t})}{var(x_{i,t}) + var(F_i)} = |\theta\rho|\frac{var(F_i)}{var(X_{i,t})} > 0 \tag{A6}$$

We make three points based on the preceding analysis:

(1) Similar to the classic omitted variable bias described in Section 3.1, in either regression model, the magnitude of the omitted variable bias increases in the magnitude of θ and ρ. The stronger the correlation between the omitted variable ($Z_{i,t}$) and either the outcome variable ($Y_{i,t}$) or the independent variable of interest ($X_{i,t}$), the greater will be the magnitude of the omitted variable bias.

(2) The magnitude of the omitted variable bias is greater in the fixed effect specification. That is, $\Delta > 0$ because both $|\theta\rho|$ and $\frac{var(F_i)}{var(X_{i,t})}$ are strictly positive.

(3) The difference in magnitude of the omitted variable bias between β^{FE} and β^{NoFE} is increasing in the fraction of variation in the independent variable that is attributable to fixed effects. To see this, note that Δ is increasing in $\frac{var(F_i)}{var(X_{i,t})}$, where the latter term represents the fraction of variation in $X_{i,t}$ that is attributable to F_i, and is bounded between 0 and 1.

References

Ai, C., Norton, E.C., 2003. Interaction terms in logit and probit models. Econ. Lett. 80 (1), 123–129.

Angrist, J., Imbens, G., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475.

Angrist, J., Pischke, J., 2008. Mostly Harmless Econometrics. Princeton university press.

Angrist, J., Pischke, J., 2010. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J. Econ. Perspect. 24 (2), 3–30.

Arellano, M., Hahn, J., 2007. Understanding bias in nonlinear panel models: some recent developments. Econometric Soc Monographs 43, 381.

Arif, S., Kepler, J., Schroeder, J., Taylor, D., 2022. Audit process, private information, and insider trading. Rev. Account. Stud. 1–32.

Arkhangelsky, D., Athey, S., Hirshberg, D.A., Imbens, G.W., Wager, S., 2021. Synthetic difference-in-differences. Am. Econ. Rev. 111 (12), 4088–4118.

Armstrong, C., Blackburne, T., Quinn, P., 2021a. Are CEOs' purchases more profitable than they appear? J. Account. Econ. 71, 2–3, 101378.

Armstrong, C.S., Core, J.E., Guay, W.R., 2014. Do independent directors cause improvements in firm transparency? J. Financ. Econ. 113 (3), 383–403.

Armstrong, C.S., Glaeser, S., Huang, S., 2021b. Contracting with controllable risk. Account. Rev. Forthcoming.

Armstrong, C., Glaeser, S., Huang, S., Taylor, D., 2019. The economics of managerial taxes and corporate risk-taking. Account. Rev. 94 (1), 1–24.

Armstrong, C., Kepler, J., 2018. Theory, research design assumptions, and causal inferences. J. Account. Econ. 66 (2–3), 366–373.

Atanasov, V., Black, B., 2016. Shock-based causal inference in corporate finance and accounting research. Critical Finance Review 5, 207–304.

Asay, H., Hoopes, J., Thornock, J., Wilde, J., 2021. Tax Boycotts. Working paper.

Baker, A., Larcker, D., Wang, C., 2021. How Much Should We Trust Staggered Difference-In-Differences Estimates? Working paper.

Balakrishnan, K., Billings, M., Kelly, B., Ljungqvist, A., 2014. Shaping liquidity: on the causal effects of voluntary disclosure. J. Finance 69 (5), 2237–2278.

Barrios, J., 2021. Staggeringly Problematic: A Primer on Staggered DiD for Accounting Researchers. Working paper.

Barrios, J., 2022. Occupational licensing and accountant quality: evidence from the 150-hour rule. J. Account. Res. 60 (1), 3–43.

Barth, M., Clinch, G., 2009. Scale effects in capital markets-based accounting research. J. Bus. Finance Account. 36 (3-4), 253–288.

Barth, M., Landsman, W., Lang, M., Williams, C., 2012. Are IFRS-based and US GAAP-based accounting amounts comparable? J. Account. Econ. 54 (1), 68–93.

Belsley, D., Kuh, E., Welsch, R., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons.

Bem, D.J., 2011. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. J. Pers. Soc. Psychol. 100 (3), 407.

Belnap, A., 2020. The Effect of Public Scrutiny on Mandatory and Voluntary Disclosure: Evidence from a Randomized Field Experiment. Working paper.

Belnap, A., Hoopes, J., Maydew, E., Turk, A., 2020. Real Effects of Tax Audits: Evidence from Firms Randomly Selected for IRS Examination. Working paper.

Berg, R., Reisinger, M., Streitz, D., 2021. Spillover effects in empirical corporate finance. J. Financ. Econ. 142 (3), 1109–1127.

Ben-Michael, E., Feller, A., Rothstein, J., 2021a. The augmented synthetic control method. J. Am. Stat. Assoc. 116 (536), 1789–1803.

Ben-Michael, E., Feller, A., Rothstein, J., 2021b. *Synthetic Controls with Staggered Adoption* (No. W28886). National Bureau of Economic Research.

Bertomeu, J., Beyer, A., Taylor, D.J., 2016. From casual to causal inference in accounting research: the need for theoretical foundations. Foundations and Trends in Accounting 15–63.

Beyer, A., Cohen, D., Lys, T.Z., Walther, B.R., 2010. The financial reporting environment: review of the recent literature. J. Account. Econ. 50 (2–3), 296–343.

Bianchi, P., Marra, A., Masciandaro, D., Pecchiari, N., 2021. Organized Crime and Firms' Financial Statements: Evidence from Criminal Investigations in Italy. *The Accounting Review.* forthcoming.

Blackburne, T., Kepler, J.D., Quinn, P.J., Taylor, D., 2021. Undisclosed SEC investigations. Manag. Sci. 67 (6), 3321–3984.

Bloomfield, M., Gipper, B., Kepler, J., Tsui, D., 2021. Cost shielding in executive bonus plans. J. Account. Econ. 72 (2–3), 101428.

Bloomfield, R., Rennekamp, K., Steenhoven, B., 2018. No system is perfect: understanding how registration-based editorial processes affect reproducibility and investment in research quality. J. Account. Res. 56 (2), 313–362.

Boland, M., Godsell, D., 2020. Local soldier fatalities and war profiteers: new tests of the political cost hypothesis. J. Account. Econ. 70 (1), 101316.

Bowen III, D., Frésard, L., Taillard, J., 2017. What's your identification strategy? Innovation in corporate finance research. Manag. Sci. 63 (8), 2529–2548.

Breuer, M., 2021. Bartik Instruments: an Applied Introduction. J Financ Rep forthcoming.

Brodeur, A., Le, M., Sangnier, M., Zylberberg, Y., 2016. Star Wars: the empirics strike back. Am. Econ. J. Appl. Econ. 8 (1), 1–32.

Brodeur, A., Cook, N., Heyes, A., 2020. Methods matter: P-hacking and publication bias in causal analysis in economics. Am. Econ. Rev. 110 (11), 3634–3660.

Brown, N.C., Stice, H., White, R.M., 2015. Mobile communication and local information flow: evidence from distracted driving laws. J. Account. Res. 53 (2), 275–329.

Bushman, R., 2021. Cash-based bonus plans as a strategic communication, coordination and commitment mechanism. J. Account. Econ. 101–447.

Callaway, B., Sant'Anna, P.H., 2021. Difference-in-differences with multiple time periods. J. Econom. 225 (2), 200–230.

Campbell, N., Williamson, B., Heyden, R., 2009. Biology: exploring life. Savvas Learning Co.

Chen, X., Cheng, Q., Wang, X., 2015. Does increased board independence reduce earnings management? Evidence from recent regulatory reforms. Rev. Account. Stud. 20 (2), 899–933.

Christensen, H., 2019. Broad-versus Narrow-Sample Evidence in Disclosure Regulation Studies: A Discussion of Badia, Duro, Jorgensen, and Ormazabal (2018). Contemporary Accounting Research. Forthcoming.

Christensen, H., Floyd, E., Liu, L.Y., Maffett, M., 2017. The real effects of mandated information on social responsibility in financial reports: evidence from mine-safety records. J. Account. Econ. 64 (2–3), 284–304.

Christensen, H., Hail, L., Leuz, C., 2016. Capital-market effects of securities regulation: prior conditions, implementation, and enforcement. Rev. Financ. Stud. 29 (11), 2885–2924.

Christensen, H., Hail, L., Luez, C., 2013. Proper inferences or a market for excuses? The Capital-Market Effects of Mandatory IFRS Adoption.

Clarke, D., Schythe, K., 2020. Implementing the panel event study. Working paper.

Conan Doyle, Arthur, 1890. The Sign of the Four. Lippincott's Monthly Magazine.

Correia, S., 2017. A feasible estimator for linear models with multi-way fixed effects. Working paper.

Daske, H., Hail, L., Leuz, C., Verdi, R., 2008. Mandatory IFRS reporting around the world: early evidence on the economic consequences. J. Account. Res. 46 (5), 1085–1142.

DeAngelo, H., DeAngelo, L., Skinner, D.J., 1996. Reversal of fortune dividend signaling and the disappearance of sustained earnings growth. J. Financ. Econ. 40 (3), 341–371.

Deaton, A., 2009. *Instruments Of Development: Randomization In the Tropics, and the Search for the Elusive Keys to Economic Development* (No. W14690). National Bureau of Economic Research.

Dechow, P., Lawrence, A., Ryans, J., 2015. SEC comment letters and insider sales. Account. Rev. 91 (2), 401–*439*.

deHaan, E., 2020. Practical guidance on using and interpreting fixed effects models. Working paper.

Dikolli, S., Keusch, T., Mayew, W., Steffen, T.D., 2020. CEO behavioral integrity, auditor responses, and firm outcomes. Account. Rev. 95 (2), 61–88.

Donelson, D., Kettell, L., McInnis, J., Toynbee, S., 2021. The need to validate exogenous shocks: shareholder derivative litigation, universal demand laws and firm behavior. J. Account. Econ. 73 (1), 101427.

Douven, I., 2021. In: Zalta, E. (Ed.), “Abduction”, the Stanford Encyclopedia of Philosophy. Available online at: plato.stanford.edu/archives/sum2021/entries/abduction.

Duchin, R., Matsusaka, J.G., Ozbas, O., 2010. When are outside directors effective? J. Financ. Econ. 96 (2), 195–214.

Duguay, R., Rauter, T., Samuels, D., 2020. The Impact of Open Data on Public Procurement. Working paper.

Edwards, G., Nesson, E., Robinson, J., Vars, F., 2018. Looking down the barrel of a loaded gun: the effect of mandatory handgun purchase delays on homicide and suicide. Econ. J. 128 (616), 3117–3140.

Fang, V., Huang, A., Karpoff, J., 2016. Short-selling and earnings management: a controlled experiment. J. Finance 71 (3), 1251–1294.

Gao, H., Hsu, P., Li, K., Zhang, J., 2020. The real effect of smoking bans: evidence from corporate innovation. J. Financ. Quant. Anal. 55 (2), 387–427.

Gelman, A., Loken, E., 2014. The statistical crisis in science: data-dependent analysis–a" garden of forking paths"–explains why many statistically significant comparisons don't hold up. Am. Sci. 102 (6), 460–466.

Gelman, A., Loken, E., 2017. Measurement error and the replication crisis the assumption that measurement error always reduces effect sizes is false. Science 355/6325, 584–585.

Glaeser, S., Guay, W., 2017. Identification and generalizability in accounting research: a discussion of Christensen, Floyd, Liu, and Maffett (2017). J. Account. Econ. 64 (2–3), 305–312.

Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. J. Econom. 225 (2), 254–277.

Gow, I., Larcker, D., Reiss, P., 2016. Causal inference in accounting research. J. Account. Res. 54 (2), 77–523.

Greene, W., 2004. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. Econom. J. 7 (1), 98–119.

Greene, D., Intintoli, V.J., Kahle, K.M., 2020. Do board gender quotas affect firm value? Evidence from California Senate Bill No. 826. J. Corp. Finance 60, 101526.

Grieser, W., Hadlock, C., 2019. Panel-data estimation in finance: testable assumptions and parameter (in) consistency. J. Financ. Quant. Anal. 54 (1), 1–29.

Guay, W., Samuels, D., Taylor, D., 2016. Guiding through the fog: financial statement complexity and voluntary disclosure. J. Account. Econ. 62 (2–3), 234–269.

Guest, N., 2021. The information role of the media in earnings news. J. Account. Res. 59 (3), 1021–1076.

Hail, L., Lang, M., Leuz, C., 2020. Reproducibility in accounting research: views of the research community. J. Account. Res. 58 (2), 519–543.

Hail, L., Tahoun, A., Wang, C., 2014. Dividend payouts and information shocks. J. Account. Res. 52 (2), 403–456.

Hasan, I., Hoi, C., Wu, Q., Zhang, H., 2017. Does social capital matter in corporate decisions? Evidence from corporate tax avoidance. J. Account. Res. 55 (3), 629–668.

Hansen, C., 2007. Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. J. Econom. 140 (2), 670–694.

Haselmann, R., Leuz, C., Schreiber, S., 2021. Know your customer: relationship lending and bank trading. Working paper.

He, J., Plumlee, M.A., 2020. Measuring disclosure using 8-K filings. Rev. Account. Stud. 25 (3), 903–962.

He, X., Yin, H., Zeng, Y., Zhang, H., Zhao, H., 2019. Facial structure and achievement drive: evidence from financial analysts. J. Account. Res. 57 (4), 1013–1057.

Healy, P., 1985. The effect of bonus schemes on accounting decisions. J. Account. Econ. 7 (1–3), 85–107.

Heckman, J., 2005. The scientific model of causality. Socio. Methodol. 35 (1), 1–97.

Heckman, J., Urzua, S., 2010. Comparing IV with structural models: what simple IV can and cannot identify. J. Econom. 156 (1), 27–37.

Heckman, J., Singer, B., 2017. Abducting economics. Am. Econ. Rev. 107 (5), 298–302.

Heckman, J., Tobias, J.L., Vytlacil, E., 2001. Four parameters of interest in the evaluation of social programs. South. Econ. J. 68 (2), 210–223.

Heckman, J., Vytlacil, E., 2001. Policy-relevant treatment effects. Am. Econ. Rev. 91 (2), 107–111.

Heckman, J., Vytlacil, E., 2007. Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. Handb. Econom. 6, 4779–4874.

Heinle, M., Samuels, D., Taylor, D.J., 2020. Disclosure Substitution. Management Science. Forthcoming.

Heitzman, S., Klasa, S., 2021. Informed trading reactions to new private information: evidence from nonpublic merger negotiations. Manag. Sci. 67 (4), 2630–2656.

Hennessy, C., Strebulaev, I., 2020. Beyond random assignment: credible inference and extrapolation in dynamic economies. J. Finance 75 (2), 825–866.

Heron, A.R., Lie, E., 2007. Does backdating explain the stock price pattern around executive stock option grants? J. Financ. Econ. 83 (2), 271–295.

Hope, O.-K., Danqi, H., Zhao, W., 2017. Third-party consequences of short-selling threats: the case of auditor behavior. J. Account. Econ. 63 (2–3), 479–498.

Hsieh, T., Kim, J., Wang, R., Wang, Z., 2020. Seeing is believing? Executives' facial trustworthiness, auditor tenure, and audit fees. J. Account. Econ. 69 (1), 101260.

Huang, Y., Jennings, R., Yong, Y., 2017. Product market competition and managerial disclosure of earnings forecasts: evidence from import tariff rate reductions. Account. Rev. 92 (3), 185–207.

Jagolinzer, A., Larcker, D., Ormazabal, G., Taylor, D., 2020. Political connections and the informativeness of insider trades. J. Finance 75 (4), 1833–1876.

Jennings, J., Kim, J., Lee, J., Taylor, D., 2022. Measurement Error, Fixed Effects, and False Positives in Accounting Research. Working paper.

Jia, Y., Lent, L.V., Zeng, Y., 2014. Masculinity, testosterone, and financial misreporting. J. Account. Res. 52 (5), 1195–1246.

Kahn, R., Whited, T., 2018. Identification is not causality, and vice versa. Review of Corporate Finance Studies 7 (1), 1–21.

Kausar, A., Shroff, N., White, H., 2016. Real effects of the audit choice. J. Account. Econ. 62 (1), 157–181.

Kecskés, A., Mansi, S., Zhang, A.J., 2013. Are short sellers informed? evidence from the bond market. Account. Rev. 88 (2), 611–639.

Kelly, B., Ljungqvist, A., 2012. Testing asymmetric-information asset pricing models. Rev. Financ. Stud. 25 (5), 1366–1413.

Kyle, A., 1985. Continuous auctions and insider trading. Econometrica: J. Econom. Soc. 1315–1335.

Larcker, D., Rusticus, T., 2010. On the use of instrumental variables in accounting research. J. Account. Econ. 49 (3), 186–205.

Lawrence, A., Ryans, J., Sun, E., Laptev, N., 2018. Earnings announcement promotions: a Yahoo Finance field experiment. J. Account. Econ. 66 (2–3), 399–414.

Lee, D., Lemieux, T., 2010. Regression discontinuity designs in economics. J. Econ. Lit. 48 (2), 281–355.

Leuz, C., 2018. Evidence-based policymaking: promise, challenges and opportunities for accounting and financial markets research. Account. Bus. Res. 48 (5), 582–608.

Leuz, C., Wysocki, P., 2016. The economics of disclosure and financial reporting regulations: evidence and suggestions for future research. J. Account. Res. 54 (2), 525–622.

Li, Y., Zhang, L., 2015. Short selling pressure, stock price behavior, and management forecast precision: evidence from a natural experiment. J. Account. Res. 53 (1), 79–117.

Lie, E., 2005. On the timing of CEO stock option awards. Manag. Sci. 51 (5), 802–812.

Lys, T., Naughton, J., Wang, C., 2015. Signaling through corporate accountability reporting. J. Account. Econ. 60 (1), 56–72.

Matsa, D., Miller, A., 2013. A female style in corporate leadership? Evidence from quotas. Am. Econ. J. Appl. Econ. 5 (3), 136–169.

Mehta, N.M., Reeb, M.D., Zhao, W., 2021. Shadow trading. Account. Rev. 96 (4), 367–404.

Minnis, M., 2011. The value of financial statement verification in debt financing: evidence from private US firms. J. Account. Res. 49 (2), 457–506.

Muller, C., Winship, C., Morgan, S.L., 2014. Instrumental Variables Regression. Sage, London, pp. 277–300.

Myers, S., 1989. Signaling and Accounting Information.

Ohlson, J., 2022. Researchers' data analysis choices: an excess of false positives? Rev. Account. Stud. 27, 649–667.

Panhans, M., Singleton, J., 2017. The empirical economist's toolkit: from models to methods. Hist. Polit. Econ. 49 (Suppl. ment), 127–157.

Peng, L., Teoh, S., Wang, Y., Yan, J., 2021. Face Value: Trait Inference, Performance Characteristics, and Market Outcomes for Financial Analysts. Working Paper.

Rambachan, A., Roth, J., 2019. An Honest Approach to Parallel Trends. Harvard University. Unpublished manuscript.

Roberts, M., Whited, T., 2013. Endogeneity in empirical corporate finance1. In: Handbook of the Economics of Finance, vol. 2. Elsevier, pp. 493–572.

Roth, J., 2019. Pre-test with Caution: Event-Study Estimates after Testing for Parallel Trends. Department of Economics, Harvard University. Unpublished manuscript.

Roychowdhury, S., Shroff, N., Verdi, R.S., 2019. The effects of financial reporting and disclosure on corporate investment: a review. J. Account. Econ. 68 (2–3), 101246.

Samuels, D., 2021. Government procurement and changes in firm transparency. Account. Rev. 96 (1), 401–430.

Samuels, D., Taylor, D., Verrecchia, R., 2021. The economics of misreporting and the role of public scrutiny. J. Account. Econ. 71 (1), 101–340.

Smaldino, P., McElreath, R., 2016. The natural selection of bad science. R. Soc. Open Sci. 3 (9), 1–17.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366.

Stock, J., Watson, M., 2003. Introduction to Econometrics, vol. 1. Pearson, New York.

Spence, M., 1978. Job market signaling. Q. J. Econ. 87, 355–374.

Trueman, B., 1986. Why do managers voluntarily release earnings forecasts? J. Account. Econ. 8 (1), 53–71.

Umar, T., 2020. Complexity aversion when seeking alpha. Working paper.

Verrecchia, R., 1983. Discretionary disclosure. J. Account. Econ. 5, 179–194.

Vigen, T., 2015. Spurious Correlations. Hachette UK.

Watts, R., Zimmerman, J., 1978. Towards a positive theory of the determination of accounting standards. Account. Rev. 112–134.

White, R., Webb, M., 2021. Randomization inference for accounting researchers. J Financ Rep 6 (2), 129–141.

Whited, R., Swanquist, Q., Shipman, J., Moon, J., 2021. Out of control: the (over) use of controls in accounting research. Account. Rev. Forthcoming.

Wooldridge, J.M., 2000. Econometric Analysis of Cross Section and Panel Data. MIT press.

Zhang, I., 2007. Economic consequences of the sarbanes–oxley act of 2002. J. Account. Econ. 44 (1–2), 74–115.