Critical Finance Review, 2016, 5: 207–304

Shock-Based Causal Inference in Corporate Finance and Accounting Research

Vladimir Atanasov
Bernard Black

[1]*College of William and Mary, USA; vladimir.atanasov@mason.wm.edu*
[2]*Northwestern University, USA; bblack@northwestern.edu*

ABSTRACT

We study shock-based methods for credible causal inference in corporate finance research. We focus on corporate governance research, survey 13,461 papers published between 2001 and 2011 in 22 major accounting, economics, finance, law, and management journals; and identify 863 empirical studies in which corporate governance is associated with firm value or other characteristics. We classify the methods used in these studies and assess whether they support a causal link between corporate governance and firm value or another outcome. Only a stall minority of studies have convincing causal inference strategies. The convincing strategies largely rely on external shocks – usually from legal rules – often called "natural experiments". We examine the 74 shock-based papers and provide a guide to shock-based research design, which stresses the common features across different designs and the value of using combined designs.

Keywords: Causal inference, Shock-based inference, Difference-in-differences, Natural experiments, Regression discontinuity, Instrumental variables, Placebo tests, Corporate governance, Identification

JEL Codes: C18, G34, G38, M48

ISSN 2164-5744; DOI 10.1561/104.00000036

1 Introduction

Much corporate finance research is concerned with causation – does a change in some input *cause* a change in some output?[1] Does corporate governance affect firm performance? Does capital structure affect firm investments? How do corporate acquisitions affect the value of the acquirer, or the acquirer and target together? Without a causal link, we lack a strong basis for recommending that firms change their behavior or that governments adopt specific reforms.

Consider, for example, corporate governance research. Decisionmakers – corporate boards, investors, regulators – want to know whether a change in governance will cause a change in firm value or performance. To provide a credible basis for "causal inference" (sometimes called "identification", a term we will avoid because it means different things to different people), a research design must address multiple econometric concerns.[2] Some of these are referred to as "endogeneity" – another term with multiple meanings that we will avoid.

Most corporate finance research does not directly address causal inference. Among the minority of papers that address this issue, an even smaller minority use credible causal research designs. We study what researchers do in major journals, and then build on this survey to provide an overview of "shock-based" research designs, which rely on an external shock as a basis for causal inference. These shock-based designs are sometimes called "natural experiments" or "quasi-experiments." We will avoid these terms also, partly because different authors use them with different meanings, and partly because they are misleading – a typical "natural experiment" is neither natural nor an experiment.[3]

"Non-causal" research designs can also be valuable, especially when they are the best available. For example, a panel data design, with firm fixed or random effects and extensive covariates, does not lock down causation, but it provides a clue, and sometimes a strong clue. The correlations provided by a

[1] We use "causation" in this restricted sense, often called the Rubin Causal Model (sometimes the Neyman-Rubin Causal Model); see Rubin (1974); Holland (1986).

[2] Many applied researchers use "identification" loosely to mean something very close to what we mean by "causal inference". Econometricians try to be more precise, but they do not use a single definition, and often tie identification to a particular regression model. See, for example, Wooldridge (2010, §4.2.1) ((In the context of [regression] models that are linear in the parameters [such as OLS] under random sampling, identification of [the coefficient] β [on an independent variable] simply means that β can be written in terms of population moments in observable variables. (Later, when we consider nonlinear models, the notion of identification will have to be more general. Also, special issues arise if we cannot obtain a random sample from the population.) In contrast, causal inference should ideally *not* depend strongly on researchers' choice of a particular model.

[3] Many applied researchers use the terms "natural experiment" and "quasi experiment" as synonyms, with meanings close to what we mean by "shock-based." But some give these terms different meanings, including Shadish *et al.* (2002) and Dunning (2012). Also, neither of these books, one on "quasi-experiments" and the other on "natural experiments," addresses difference-in-differences designs. Many applied researchers would also not see these terms as encompassing event studies.

"pure observational study," with careful matching of treated and control firms, can be valuable as well. These designs are outside the scope of this project.[4]

To study all shock-based papers in corporate finance is an unmanageable task. We therefore narrow the scope of our assessment, and study what researchers do in corporate governance studies – a still large but (barely) manageable job. We survey 13,461 articles in 22 major journals in accounting, economics, finance, law, and management over 2001-2011, and identify 863 empirical corporate governance papers, which study whether corporate governance predicts firm value or another dependent variable. Many of these papers do not directly discuss causation, but we care about the results principally because we care about causation. We classify the strategies these papers use, identify 74 papers with shock-based research designs (involving 40 distinct shocks), and study these papers to provide a guide to shock-based design. While our focus is on corporate governance research, the lessons on research design apply to research in accounting and corporate finance more generally.

We focus on corporate governance for several reasons. One is manageability. A second is the availability of shocks. Governments regularly change corporate governance rules; some of these changes provide useful shocks. Third, policymakers need to know whether governance *causes* value. If researchers provide evidence only on association, policymakers may adopt rules based on flawed data. Fourth, by examining an (important) area that we know, we can provide more focused analysis of good and less-good shock-based designs.

A central theme of this paper is that credible causal inference strategies often rely on "shocks" to governance. These shocks can provide reason to believe that a change in governance *causes* a change in the firm's value or behavior. Here, a "shock" is a discrete, external event that causes some firms to be treated; the others become "controls." The assignment of firms to treatment versus control should be plausibly exogenous – not chosen by the firm, and ideally uncorrelated with firm characteristics (observed or unobserved) that might predict response to the shock or other changes in the world. Usually, and ideally, we can measure outcomes both before and after the shock. Most convincing shocks, in turn, come from legal rules, rule changes, and law-based discontinuities (together, "legal shocks").

A second central theme is a focus on *shocks* and on common themes in shock-based design, which apply across the methods that are used to exploit shocks. Difference-in-differences (DiD), regression discontinuity (RD), event study (ES), and instrumental variable (IV) designs can all be used to exploit shocks. To be credible, these designs must satisfy similar exogeneity, relevance, covariate balance, and "only through" conditions. These common elements of shock-based

[4] For examples of our own work using non-shock-based research designs, see Atanasov *et al.* (2012) (pure observational study of impact of litigation on the reputation of venture capitalists); Black *et al.* (2014) (study, using firm fixed and random effects, of the impact of firm-level corporate governance on firm value in emerging markets). For a skeptical assessment of how strong the clue to causation is, from a "classic panel data" design with firm fixed or random effects, see Nasev *et al.* (2016).

design have been obscured because most of the causal inference literature treats each design separately, and the literature on particular designs often glosses over one or more of these requirements. For example, the DiD, IV, and ES literature rarely stresses the need for covariate balance between treated and control firms, and the DiD literature rarely stresses the need to satisfy an only through condition.

A third theme is the value of using multiple shock-based designs and, where feasible, using combined designs. If a "credible shock" (one which provides a credible basis for causal inference) exists, several designs can often be used to exploit it. For example, shock-based IV designs can often be recast as DiD. Frequently, shock-based methods can be combined. For example, if a shock involves a discontinuity, a combined DiD/RD design can be attractive. Often, covariate balance can be improved by combining a shock-based design with "balancing methods" adapted from pure observational studies.

We seek to provide guidance on how to *improve* shock-based causal inference, even if inference remains imperfect. We share neither the perspective of some researchers, whose view can be caricatured as "endogeneity is everywhere, one can never solve it, so let's stop worrying about it"; nor the "endogeneity police", whose attitude is that "if causal inference isn't (nearly) perfect, a research design is (nearly) worthless"; nor that of authors who know they have an endogeneity problem, but say little or nothing about it in their paper, hoping the referee won't notice, or else use a weak instrument to address endogeneity and hope the referee won't object. Our anecdotal sense is that paper acceptance and rejection decisions often turn on which position – endogeneity is everywhere, endogeneity police, or our middle ground – best describes the referee and the editor.

As part of providing advice on better shock-based design, we (unavoidably) criticize many of the shock-based papers we study. These criticisms should not obscure the value of exploiting shocks, when they can be found. An imperfect shock-based paper will often be more convincing than the non-shock alternatives.

We believe that useful shocks can often be found. Even true randomized trials can sometimes be found or created. We are collecting the shocks used in our sample in a public database that we plan to post on the Social Science Research Network (SSRN) (Atanasov and Black, 2016a). We plan to update this database to include additional shocks. Many of these shocks can be put to additional uses. Many more useful shocks surely exist, but have not yet been exploited.

Issues of causal inference are receiving increased attention in finance and accounting. Three recent papers overlap with ours, but none focuses on shock-based designs. Roberts and Whited (2013) review endogeneity issues in corporate finance research generally. Larcker and Rusticus (2010) criticize the IVs used in accounting research. Bowen *et al.* (2016), study the evolution of researcher attention to endogeneity in corporate finance over 1970-2012, but focus on which methods are used, not whether they are used well.[5]

[5]In other related work, Claessens and Yurtoglu (2013) survey empirical corporate governance

This paper proceeds as follows. Section 2 discusses the principal causal inference challenges in corporate governance research, presents our notation, and provides an overview of shock-based research designs, stressing their common features. Section 3 describes our data and methodology. Sections 4-7 discuss, respectively, DiD, ES, IV, and RD designs. Section 8 concludes. We borrow liberally from the general causal inference literature, often without citation.

2 Shock-Based Research Design: Overview

Sections 2.1–2.3 provide background: We review the challenges to causal inference in corporate governance studies, and corporate finance more generally; present our causal inference notation; and summarize why randomized trials can produce unbiased causal estimates. There will be little new here for readers familiar with causal inference. Section 2.4 provides an overview of shock-based inference; the remaining sections provide details on particular methods.

2.1 Empirical Challenges to Causal Inference in Corporate Governance Research

We review briefly here the principal reasons why one cannot regress an outcome variable (say Tobin's q) on a governance variable *gov*, a constant term (which we assume below, but do not repeat), and a vector of controls **x**, and infer that a change in *gov* will *cause* a change in q.

Suppose we run such a regression, using ordinary least squares (OLS):

$$q_i = a + b * gov_i + \boldsymbol{c} * \boldsymbol{x_i} + \epsilon_i \tag{1}$$

and observe a positive (and statistically significant, which we assume below, but do not repeat) coefficient b on *gov*. This tells us that, conditioned on covariates (which we assume below, but do not repeat), higher *gov predicts* higher q. We cannot infer that a change in *gov* will *cause* a change in q (on average, which we assume below, but do not repeat) – using "cause" to mean that, if one increases *gov*, changing nothing else, q will increase.

One problem is *reverse causation*. Perhaps q causes *gov*. Regression cannot tell us the direction of the causal arrow. After all, we could have instead regressed

research in emerging markets, and note the trend toward greater attention to causal inference. Brown *et al.* (2011) offer a broad review of corporate governance research, but pay limited attention to IV and none to other causal inference methods. Gassen (2014) studies the use of causal methods in accounting research, but does not assess whether the methods are used well. Lennox *et al.* (2012) study the use of Heckman selection models in accounting research. Gippel *et al.* (2015) find frequent shortfalls in addressing endogeneity in a small sample of finance and accounting papers published in Asia-Pacific journals and suggest greater use of natural experiments. Catan and Kahan (2016) and Karpoff and Wittry (2015) criticize DiD studies of the impact of state adoptions of antitakeover statutes.

gov on *q*:

$$gov_i = a_r + b_r * q_i + \mathbf{c_r} * \boldsymbol{x_i} + \epsilon_{i,r} \tag{2}$$

Usually, if b is positive and significant, the coefficient b_r from this reversed regression will be as well.

A second problem is *omitted variable bias*. Perhaps one or more unobserved variables **u** cause both q and *gov*, or mediate the relationship between q and *gov*. Without them, the coefficient b is a biased estimate of the true causal effect of *gov* on q. Consider a single omitted u. If we observed u, the "long" regression model would be:

$$q_i = a_{long} + b_{long} * gov_i + \mathbf{c_{long}} * \boldsymbol{x_i} + d_{long} * u_i + \epsilon_{i,r} \tag{3}$$

The coefficient b from the "short" regression (1) equals the coefficient $\mathrm{b_{long}}$ from the long regression (3), plus an omitted bias term equal to the product of $\mathrm{d_{long}}$ and f = coefficient on *gov* from regressing u on *gov)* (Wooldridge, 2012, § 5.1):

$$\mathrm{b} = \mathrm{b_{long}} + \mathrm{d_{long}} * \mathrm{f} \tag{4}$$

In much governance research, there can be multiple omitted variables and we aren't sure what they are, so we don't even know the sign of the bias.

Given panel data on firms, plus sufficient within-firm time variation in *gov*, one can use a firm fixed effects (FE) specification to partly address omitted variable bias. Let f_i be firm effects and g_t be time effects, and replace the *OLS* specification in eqn. (1) with:

$$q_{it} = a + f_i + g_i + b * gov_{it} + \mathbf{c} * \mathbf{x_{it}} + \varepsilon_{it} \tag{5}$$

This specification controls for unobserved time-invariant firm factors. This helps, but only so much. Unobserved time-varying covariates will still lead to omitted variable bias. And governance often changes slowly over time, so FE may have low power.

A third, related concern is *specification error*. Even if we could perfectly measure *gov* and all relevant covariates, we would not know the functional form through which each influences q. Misspecification of *gov* or the **x**'s is similar to omitted variable bias. The missing part of the correct specification leads to biased coefficients, just like any other omitted variable.

A fourth concern, also related to omitted variable bias, is that firms may change *gov* to *signal* to investors something about management attitudes, or other factors which investors can't readily observe.[6] Conversely, firms may appear not to benefit

[6]An analogy may help to illustrate the difference between omitted variable bias and signaling. Consider the classic labor economics problem of measuring the returns to education. Students may obtain more education *because* they are smarter and thus learn faster or enjoy learning more (ability is an omitted variable, that is correlated with both education and return to education), or they may obtain more education to signal to employers that they are smarter (ability is correlated with education, but additional education may have no actual value).

from a governance reform because the impact of the reform on value is offset by a negative signal from its adoption.[7]

A fifth concern is *simultaneity*, in which q, *gov*, and $\mathbf{x}$ are determined simultaneously. Perhaps there is bidirectional causation, with q causing *gov* and *gov* also causing q. OLS regression will provide a biased estimate of the magnitude and perhaps the sign of the effect.[8]

A sixth problem is *heterogeneous effects*, with the causal effect of *gov* on q depending on both observed and unobserved firm characteristics. Assume that firms seek to maximize q; different firms have different optimal *gov*'s; and firms know their optimal *gov*'s. If we observed all factors that affect q, each firm would be at its own optimum and OLS regression would give a zero coefficient on *gov*, which would misleadingly suggest no relationship between *gov* and q. If some $\mathbf{u}$'s are unobserved, we could find a positive or negative relationship; but would know neither the true causal relationship nor how it is mediated by the $\mathbf{x}$'s and $\mathbf{u}$'s.

A seventh problem, especially relevant for corporate governance research, is *construct validity*. Corporate governance involves a complex system of diverse mechanisms, serving a number of goals. We usually don't know what is "good" governance, either in general or for particular firm goals. Many studies build *gov* by summing scores on a variety of "elements" (e.g., Gompers *et al.*, 2003). Some features of the multi-element index may be important, others may not; some may be complements, while others may be substitutes. The construct that we call *gov* may poorly fit the underlying concept. Tobin's q is also a construct, which imperfectly measures many things.[9]

An eighth problem is *measurement error*. "Classical" random measurement error in *gov* or the $\mathbf{x}$'s will bias coefficient estimates toward zero. Classical measurement error for the outcome will inflate standard errors but will not lead to biased coefficients. The consequences of non-random measurement error are similar to specification error.

A ninth factor, which one might call *observation bias,* is analogous to the Hawthorne effect, in which observed subjects behave differently because they are observed. Firms which change *gov* may behave differently because their managers or employees *think* the change in *gov* matters, when in fact it has no direct effect.[10]

[7]An example is stock price reaction to a firm replacing its CEO. The governance action is inextricably bundled with the release of information about the firm's performance under the old CEO, which led to the replacement, and about the quality of the board.

[8]Roberts and Whited (2013) provide a formula for the bias from bidirectional causation for the simple case with no control variables.

[9]Black *et al.* (2014) discuss construct validity for governance studies. For an overview of construct validity issues and responses, see Shadish *et al.* (2002).

[10]A recent example is Grullon *et al.* (2015) who study a randomized experiment in which the SEC assigned one-third of the Russell 3000 firms to be treated by relaxing short sale restrictions, with the other two-thirds as a control group. The authors find a small rise in short selling and fall in share price at treated firms, and a much larger decrease in real investment. An "observation bias" story for the drop in investment (suggested by Holger Spamann at a conference): treated firms might believe they

A tenth factor involves *interdependent effects* on firms which adopt a reform. For example, a governance reform that will not affect share price for a single firm might be effective if adopted widely, because investors will then appreciate the reform's impact. Conversely, a reform which improves efficiency for a single firm might not improve profitability if adopted widely, because the gains will be competed away.

These obstacles to credible causal inference, plus others put aside below as beyond scope, suggest the challenges facing empirical corporate governance researchers. We turn next to the terminology of causal inference and some responses to these challenges.

2.2 Causal Inference Notation and Terminology

We summarize here the causal inference notation and terminology we will use. We work primarily within the "Rubin Causal Model" (so termed in Holland (1986)), in which causal inference is centrally a missing data problem, and follow the notation in Imbens and Rubin (2015). For an observed firm i, we would like to know how an outcome y_i would change if we "treated" firm i in some way, while leaving all else unchanged (other than follow-on changes caused by the treatment). We assume a binary "treatment", ($w_i = 1$ if treated; 0 if not), but consider continuous treatments below. The firm-level causal effect τ_i of treatment on y_i is defined as the value of y_i if firm i is treated, minus the value of y_i if firm i is not treated:

$$\tau_i = y_i(w_i = 1) - y_i(w_i = 0), \text{ or, more compactly: } \tau_i = y_{i1} - y_{i0}$$

The "fundamental problem of causal inference" (Holland, 1986) is that we observe only one of the two potential outcomes, y_{i1} and y_{i0}. If firm i is treated, we observe y_{i1} but not y_{i0}; while if firm i is not treated, we observe y_{i0} but not y_{i1}. The causal inference strategy is to impute the missing potential outcome for the treated firms from the control firms (and vice versa, but we focus here on treatment effects for treated firms). The central challenge to imputation is "selection bias": the treated and control firms differ in some way, perhaps unobserved, which will bias the estimated treatment effects.

2.3 Randomized Trials (RT)

One way to ensure similar treatment and control groups is to conduct a randomized trial (also called a randomized experiment). If the treatment is truly random, then all variables of interest – the potential outcomes, observed pretreatment covariates $\mathbf{x}$, and unobserved pretreatment covariates $\mathbf{u}$, will all be independent of whether a firm is treated. Any differences in covariates in a finite sample will be random

are vulnerable to "bear raids", and cut or defer their investments, perhaps only until the experiment expired.

and tend to zero as the sample size increases. Denote the period after treatment as a (for after), and the period before treatment as b (for before). A statement of random assignment to treatment that allows for both periods is:

$$\mathrm{w_i} \perp \left(y_{0i,a}; y_{0i,b}; y_{1i,a}; \mathbf{x}_{i,b}; \mathbf{u}_{i,b}\right) \tag{6}$$

Because an RT eliminates selection bias, it is often considered the gold standard for causal inference. Thus, it is useful to understand which empirical challenges it does and does not address. An RT addresses reverse causation and simultaneity because the treatment changes only *gov*, without affecting other variables that could influence q. It addresses omitted variable bias and measurement error for covariates, because the shock is independent of all covariates, observed or not, and therefore is also independent of any function f(**x**) of the covariates, including a noisy or incorrect measurement of these variables. The shock is applied at random, which precludes a signaling effect.

An RT can let us estimate the average treatment effect for the treated (ATT) even without data for the "before" period. We want to estimate ATT $= \mathrm{E_{treated}}[\mathrm{y_{1,a}}] - \mathrm{E_{treated}}[\mathrm{y_{0,a}}]$ We can estimate the "after" expectations $E_{\mathrm{treated}}[\mathrm{y_{1,a}}]$ and $\mathrm{E_{controls}}[\mathrm{y_{i0,\,a}}]$. Randomization ensures that the observed estimate for the control group, $\overline{y}_{0,a}^{\mathrm{controls}} = \frac{1}{n_c} \sum_{controls\, j} y_{j0,\,a}$ is an unbiased estimate of the unobserved expectation for the treated $\mathrm{E_{treated}}[\mathrm{y_{0,\,a}}]$, which we want to know. Thus, the naïve estimate of ATT:

$$\hat{\tau}_{ATT}^{naive} = \overline{y}_{1,a}^{\mathrm{treated}} - \overline{y}_{0,a}^{\mathrm{controls}} \tag{7}$$

is unbiased. In addition, since the treated and the controls are the same in expectation, ATT = ATC (average treatment effect for the controls) = ATE (average treatment effect for the entire sample). One can also estimate the treatment effect from a simple regression:

$$y_i = \alpha + (\delta_{\mathrm{RT}} * w_i) + \varepsilon_i \tag{8}$$

Here $\hat{\delta}_{RE}$ is the estimate of ATT. Pretreatment covariates $\boldsymbol{x}_{\mathrm{i,b}}$ can be added to eqn. (8); this can improve the precision of the estimate, but the estimate is unbiased even without them.

Even an RT, however, cannot address construct validity, measurement error for the dependent variable, or specification error for the dependent variable. To offer an example, suppose we want to test whether an audit committee will reduce the likelihood that firms will commit financial fraud. To test this hypothesis, we assign firms at random to be treated with an audit committee or not, and find no effect. There might be an effect in fact, which we missed because we erred in defining financial fraud (construct validity), measuring it (measurement error), or in picking a threshold materiality level above which we counted an event as financial fraud (specification error). An RT can address observation bias only if the controls receive a placebo. In corporate governance research, randomized trials are rare (there are none in our sample) and placebos are hard to imagine.

An RT also cannot address interdependent effects. The Rubin causal model excludes interdependence through the "stable unit treatment value assumption" (SUTVA). SUTVA has two aspects: there is only one level of treatment, and treating one firm does not affect other treated or control firms.[11] The "one level of treatment" assumption can be relaxed. The "SUTVA independence" assumption is crucial, yet may be violated in corporate governance research.

Some notes on the RT design: First, ATT, ATC, and ATE are *average* effects. We can estimate averages for subsamples by conditioning on covariates. Second, if some firms are assigned to treatment but do not comply with the treatment, or some control firms voluntarily adopt the treatment, one has an "intent to treat" or "encouragement" design. One can estimate the treatment effect for "compliers" – firms whose behavior is changed by being assigned to treatment – by using assignment to treatment as an instrument for actual treatment. Third, if data is available both before and after treatment, one can use a DiD design to reduce the risk of bias due to imperfect randomization.

2.4 Shock-Based Causal Inference

Shock-based designs use an external shock to limit selection bias. At their best, they can approach, but never achieve, fully random assignment. Different designs – DiD, ES, IV, and RD – appear to rely on different assumptions. However, we will argue, these designs share core elements. All rely on a "good shock" – one which permits credible causal inference. A good shock should satisfy five conditions:

(1) *Shock Strength*: The shock is strong enough to significantly change firm behavior or incentives.

(2) *Exogeneous Shock*. The shock came from "outside" the system one is studying. Treated firms did not choose whether to be treated, cannot anticipate the shock, the shock is expected to be permanent, and there is no reason to believe that which firms were treated depends on unobserved firm characteristics. If the shock is exogenous, or appears to be, we are less worried that unobservables might be correlated with both assignment to treatment and the potential outcomes, and thus generate omitted variable bias. Shock exogeneity should be defended, not just assumed.

(3) *"As If Random" Assignment:* The shock must separate firms into treated and controls in a manner which is close to random. One often needs to allow an exception for the variable which determines which firms are affected by the shock, which we will call the "forcing variable ($x^{forcing}$)," any related

[11] For example, Imbens and Rubin (2015, §1.6).

variables, and, in some studies, a variable which is changed by the shock (x^{forced}).[12] Different research designs can accommodate different departures from random assignment, but the closer the shock comes to random assignment, the more credible it will be.

(4) *Covariate balance*. The forcing and forced variables aside, the shock should produce reasonable covariate balance between treated and control firms, including "common support" (reasonable overlap between treated and control firms on all covariates). Somewhat imperfect balance can be address with balancing methods, but severe imbalance undermines shock credibility, even if the reason for imbalance is not obvious. Covariate balance should be reported.

(5) *Only-Through Condition(s)*: We must have reason to believe that the apparent effect of the shock on the outcome came *only through* the shock (sometimes, through a specific channel). The shock must be "isolated" – there must be no other shock, at around the same time, that could also affect treated firms differently than control firms. And if one expects the shock to affect outcomes through a particular channel, the shock must also affect the outcome only through that channel. In IV analysis, this is called an "exclusion restriction," because one assumes away (excludes) other channels; we prefer the more descriptive term "only-through condition."

These conditions are related. A truly exogenous shock will tend to produce as-if random assignment to treatment. As-if-random assignment, if close to truly random, should produce reasonable covariate balance and support the credibility of the only-through condition. Conditions (1) and (2) are part of standard discussions of DiD, and (1), (2) and (5) are well-known for IV. But standard statements of the conditions for DiD do not address the remaining conditions, and standard statements for IV often do not discuss as-if random assignment or its corollary, covariate balance.

All shock-based designs provide a "local" estimated treatment effect. Different designs estimate provide somewhat different estimates, but all are local to the sample. One measures, loosely speaking, a "local average treatment effect (LATE)" for particular firms, in particular countries, which are "treated" with particular governance changes. LATE terminology, developed for IV, applies to DiD as well. If some firms in the treatment group do not comply with the treatment, one can estimate LATE for the "compliers," using assignment to treatment as an instrument for actual treatment.

[12] For example, Black *et al.*'s (2006) study Korean reforms in 1999, which require firms with assets >2 trillion won (so $x^{forcing}$ is assets) to adopt several board structure reforms (which collectively are $\mathbf{x}^{forced}$). Other measures of firm size, such as sales on market capitalization, can also be seen as forcing variables.

The remainder of this part provides a brief overview of the principal designs, intended to highlight their similarities, and the potential gains from combined designs. Parts 4-8 provide more details on each design.

2.5 First Look at Difference-in-Differences (DiD)

We discuss here briefly how the conditions for a good shock apply to DiD. We gloss over many details, which we address in Part IV. To use DiD, one needs separate treated and control groups, with data both before and after the treatment. One can then estimate the average treatment effect for the treated (ATT) as (after-minus-before change for treated firms) minus (after-minus-before change for control group).

$$ATT_{\mathrm{DiD}} = E_{treated}[y_{1,a} - y_{1,b}] - E_{controls}[y_{0,a} - y_{0,b}]$$

This estimate can be implemented as a firm fixed effects regression. Assume two periods, one before and one after treatment, put aside covariates, and let *post* be a post-treatment dummy and f_{i} be firm dummies. We estimate:

$$\{\text{2-period DiD}\} : y_{it} = \alpha + f_i + (\beta * post) + (\delta_{DiD} * post * w_i) + \varepsilon_{it} \qquad (9)$$

Here $\hat{\delta}_{\mathrm{DiD}}$ is the empirical estimate of ATT.[13] The two-period eqn. (9) can be extended to allow for multiple pre- and post-periods, running from $-\mathrm{n_{pre}}$ to $+\mathrm{n_{post}}$. Let $t = 0$ be the last pre-treatment period, $\mathrm{g_t}$ be period dummies and add a t subscript to $\mathrm{w_i}$, which becomes $\mathrm{w_{it}}$ (= 1 for treated firms if $t > 0$, 0 otherwise):

$$\{\text{panel DiD}\} : y_{it} = \alpha + f_i + g_t + (\delta_{DiD} * w_{it}) + \varepsilon_{it} \qquad (10)$$

For eqn. (9) or (10) to provide an unbiased estimate of ATT, one needs to assume that the after-minus-before *change* in potential outcomes (if firms are not treated), is independent of assignment to treatment:

$$\text{DiD requirement 1 (parallel changes): } \mathrm{w_i} \perp \left(y_{0i,a} - y_{0i,b}\right) \qquad (11)$$

Relative to a randomized trial, this "parallel changes" assumption replaces the random assignment of potential outcomes in *levels*, provided by a randomized trial, with what one might call "random assignment of *changes*." With only one before and one after period, the parallel changes assumption is not testable. With panel data, it is partially testable in the pre-treatment period, as we discuss below.

What will make the parallel changes assumption credible? We want assignment to treatment to come from an exogenous shock. This makes it less likely that unobserved covariates drive both assignment to treatment and the after-minus-before change in outcome. We also want the treated and control groups to be

[13] To estimate the coefficient on an interaction term such as $\mathrm{w_{it}}$, one must normally include each interacted variable separately in the regression. Here, $\mathrm{w_i}$ is captured by the firm effects f_{i}.

similar prior to treatment – to be similar on outcomes and trends in outcomes (ideally highly so), and to have reasonable covariate balance on a rich set of observed covariates, other than $x^{forcing}$ and x^{forced}. This makes it more likely that the observed change in treated firms reflects the impact of the shock, rather than other differences between the treated and control firms.

Let $\mathbf{x}^{other}$ be the pre-treatment covariates, other than $x^{forcing}$ and x^{forced}. We want assignment to treatment to be nearly independent of everything except $x^{forcing}$ and $x^{forced:}$:

$$\text{DiD credibility requirement 2:} w_i \overset{near}{\perp} \left(y_{0i,a}; y_{0i,b}; y_{1i,a}; \mathbf{x}_i^{other}; \mathbf{u}_i\right) \tag{12}$$

A check for covariate balance is needed, yet is not part of standard DiD design. In principle, eqn. (12) can hold only conditional on observed pre-treatment covariates $\mathbf{x}_{i,b}$, as in any observational study. But in practice, if treated and controls differ substantially on $\mathbf{x}_i^{other}$, we will worry that they may differ on unobservables too.

One also wants to ensure that the shock is "strong" – it meaningfully changes the forced variable (shock condition 1). A strong shock makes it easier to find a significant treatment effect, and makes it more likely that the treatment, rather than some unobserved factor, is driving the observed result.[14]

The only-through conditions for DiD are best illustrated by example. Most shocks come from legal changes. For clean design, we want the rule that produces the shock to be adopted at random, but many regulators don't act that way. A regulator that adopts rule *A* might also adopt related rule *B* at roughly the same time, where rule *B* (or *A* and *B* together, or a broader set of policies *P* that includes both) causes a change in outcomes for treated firms. The "only through" claim – that shock *A* is the only relevant cause of the observed post-shock difference in outcomes – must be defended against this concern. If one posits a particular channel for how a reform affects the outcome, one must exclude other possible channels, often through choosing a control group that is similar to the treatment group on all covariates except the one of interest.

2.6 First Look at Event Studies

Event studies are a well known corporate finance research design. We do not discuss event study "basics", and instead focus on intuitions and design advice that reflect their similarities to other causal inference designs.

First, an event study can be seen as a special form of DiD.[15] A classic event study measures the impact of information on the share prices of "treated" firms

[14]For an analogous argument that a strong IV is less vulnerable to small departures from the IV-validity assumptions, see Small and Rosenbaum (2008).

[15]The similarity between DiD and event studies will be apparent to anyone familiar with both, but is rarely noted in the event study literature. Gelbach *et al.* (2013) is an exception, perhaps because the authors come from the causal inference tradition.

over an "event window" around the time the information is disclosed, as:

$$[\text{abnormal return}] = [\text{total return}] - [\text{normal return}]$$

The normal return is an estimate of the unobserved potential outcome, if the firm had not been treated. One estimates this potential outcome using observed returns to a control group (the firms in the index used to estimate the normal return).

This perspective suggests ways to improve on event study design. For DiD, one should work hard to ensure that control firms are highly similar to treated firms. In event studies, in contrast, one typically computes normal returns using a broad market index and a simple model of share returns, often the "market model":

$$r_{it} = \alpha_i + \beta_i * r_{mt} + \epsilon_{it} \tag{13}$$

Here r_{it} is the return to firm i on day t; r_{mt} is the return to the market index, α_{i} and β_{i} are parameters which are estimated (often during a pre-event period), and ϵ_{it} is the abnormal return. Sometimes, a 3- or 4-factor model is used instead (e.g., Perez-Gonzalez, 2006). But the firms in the index are typically not limited to those similar to the treated firms. Instead, researchers often use a broad index, control for one or several overall pricing factors, and assume that each firm reacts linearly, to changes in those factors.

The DiD analogy suggests that event study credibility will increase if treated and control units are more similar. Covariate balance between treated and control firms should be assessed and, where appropriate, improved through balancing methods. Balance is especially important for studies with long event windows, which allow more time for firm characteristics, not captured by event study models, to affect returns. In DiD language, there is more time for violations of the parallel trends assumption to become important.[16]

Causal inference from an event study also relies on an only through condition, involving investors' pre-shock information sets. The "event" releases new information. This information can affect outcomes both through the underlying shock (a governance reform, say) and in other ways. Consider takeover defenses. An announcement that a firm has adopted a defense can affect share price by strengthening the firm's defenses, or by changing investor expectations that the firm will receive a bid or will fight a bid if received. Unless the "revised expectations" channels can be ruled out (often, they cannot), we cannot infer that defense effectiveness caused any share price impact.

[16] A caveat: Some events will affect other similar firms also. For example, a takeover bid for firm A will change investor expectations about the likelihood of a bid for similar firm B. This is testable – one assesses whether the announcement predicts abnormal returns to potentially affected firms, relative to a suitable control group. If this is a concern, the affected firms should be removed from the control group. For takeover bids, in effect, the control firms should be similar, but not so similar as to be affected.

2.7 First Look at Instrumental Variables (IV)

The classic "econometrics textbook" response to reverse causation, omitted variables, and simultaneity issues is to find an IV for *gov*, and run a two-stage least squares (*2SLS*) regression.

In 2SLS, the instrument z substitutes for the instrumented variable; and we *assume* that the power of the *instrument* to predict the outcome (say q) reflects the true power of the *instrumented variable*, here *gov*. This assumption is reflected in the *2SLS* estimate of the coefficient on *gov*, which is, without covariates:

$$\widehat{\beta}_{2SLS} = \frac{Cov\,(z,q)}{Cov\,(z,gov)} \tag{14}$$

A classic statement of the requirements for a valid instrument z for *gov* is that:[17]

(i) z is correlated with *gov* (preferably strongly, to increase statistical power and avoid weak instrument issues); and

(ii) Cov (z, ε) = 0, where ε is the unobserved true error in the original regression.

This statement is unhelpful and has likely contributed to frequent use of invalid IVs. The first condition can be tested in the sample. The second condition replaces the untestable and often false assumption underlying OLS that Cov(*gov*, ε) = 0, with the untestable and often false assumption that Cov(z, ε) = 0. A better statement of the requirements for a valid instrument would be, following Angrist and Pischke (2009, §4.1):

(i) *instrument strength*: z is correlated with *gov* (preferably strongly, to increase statistical power and avoid weak instrument problems); and conditioned on the observed covariates **x**:

(ii) *instrument as good as randomly assigned*: z can't be influenced by the outcome variable q (thus ruling out reverse causation and simultaneity). This is sometimes loosely phrased as z being "exogenous" to the variables in the original OLS or panel regression. But we need more than this: z must be as good as randomly assigned – it must be independent of the potential outcomes, either fully ($z \perp y_1, y_0$) or conditioned on covariates ($z \perp y_1, y_0|\mathbf{x}$); and

(iii) *only-through condition* (the hardest to satisfy in practice): z predicts the outcome q *only through* the instrumented variable *gov*, not directly or through unobserved variables **u**.

Framing the conditions for a valid instrument this way highlights the similarity between IV and shock-based designs. For a *shock-based* IV, these general IV

[17] See, e.g., Wooldridge (2010, ch. 5). One also needs any covariates to satisfy Cov(**x**, ε) = 0.

conditions map directly onto the requirements for a good shock stated above. In principle, random assignment of an IV can hold only conditioned on observed covariates. But in practice, a design is likely to be credible only if we approach unconditional random assignment, except for the shock forcing variable (eqn. (6)). A check for covariate balance is needed, yet is not part of standard IV design.

For shock based IV, satisfying the only through condition involved both a well-known direct condition (z predicts *q* only through *gov*) and an implicit one – the need for an isolated shock. Both should be defended.

For IV in general (not just shock-based IV), framing the conditions for a valid instrument in this way facilitates careful thinking about when conditions (ii) and (iii) are likely to be true. For corporate governance, an external shock can sometimes plausibly satisfy these conditions, but even that is hardly certain (compare Rosenzweig and Wolpin (2000)). A standard financial variable cannot – or at least we've never seen a convincing example where it does.

IV estimates a "local average treatment effect" (LATE). A shock that requires or encourages some firms (but not others) to change *gov* can be used to instrument for *gov*. However, *2SLS* estimates a causal effect only for "compliers" who adopted the *gov* change only because of the shock (Angrist *et al.*, 1996). One can use DiD to estimate an "intent to treat" effect of the shock, and shock-based IV to estimate LATE for the compliers.

2.8 First Look at Regression Discontinuity (RD) Designs

A design that relies on an abrupt discontinuity, which separates firms into treated and control, can be credible even when one lacks "before" and "after." This design is often called "regression discontinuity," although it need not involve regression analysis. Assume that a legal rule causes *gov* to change only if a firm exceeds a threshold level for a forcing variable, such as firm size. Firms just below and just above the threshold should be similar, so the just-below-threshold firms can form a control group for the just-above-threshold firms. There are good, recent reviews of RD design (Imbens and Lemieux, 2008; Lee and Lemieux, 2010), so our discussion is summary in nature, and focuses on design features that are similar for RD and other shock-based designs.

Let $\mathrm{x}^{\mathrm{forcing}}$ be the "forcing variable" for the discontinuity in *gov* and $\mathbf{x}^{\mathrm{other}}$ be the other pre-treatment covariates. Within a bandwidth around the discontinuity, assignment to treatment should be independent of everything except $\mathrm{x}^{\mathrm{forcing}}$:

$$\mathrm{w_i} \perp \left(y_{0i}; y_{1i}; \mathbf{x}_i^{other}; \mathbf{u}_i\right) \tag{15}$$

Compare the similar condition for DiD credibility in eqn. (12).

If discontinuity-based assignment is close enough to random, one can estimate treatment effects exactly as for a randomized experiment. More often, however, the forcing variable may directly predict the outcome. One can control for the (presumably smooth) direct effect of the forcing variable on the outcome. A regression-based treatment effect estimate, with a simple linear control for $x^{forcing}$, can be implemented as:

$$y_i = \alpha + (\delta_{\mathrm{RD}} * w_i) + \beta * x_{i,b}^{forcing} + \varepsilon_i \tag{16}$$

Here $\hat{\delta}_{RD}$ is the estimated treatment effect.

As for any shock-based design, we need the shock to be exogenous. For RD, this means that firms do not manipulate which side of the threshold they fall on. We need the shock to be strong. Strength can be assessed graphically – the proportion of compliers should be visibly higher just above the threshold than just below it. And we need the shock to satisfy only through conditions – it must be isolated from other shocks that might affect the outcome, and must predict the outcome only through the forced variable.

If some above-threshold firms don't comply with the treatment, some below-threshold firms voluntarily comply, or both, one has a "fuzzy" discontinuity. An above-threshold dummy can then be used as an instrument for actual treatment. As Angrist and Pischke (2009, §6.2) put it, "Fuzzy RD is IV." One measures LATE – the treatment effect for firms who would comply if above the threshold, but not if below it. The usual conditions for a valid IV apply.

2.9 Similarities across Methods

As we discuss above, all shock-based methods rely on common requirements for a "good shock." We discuss here some additional similarities across methods, as well as the potential to use multiple methods in a single study.

First, all methods depend on random or nearly random assignment to treatment. Methods other than RT weaken fully random assignment in some way; but become more credible as they approach random assignment (other than for the forcing variable for RD and DiD). One should confirm nearly random assignment by checking for covariate balance. Testing for parallel pre-treatment trends, if one has panel data, can also be seen as a check for balance in pre-treatment changes in the outcome variable.

The need to confirm shock strength and covariate balance applies across methods. Assessing strength in a "first stage" is routine for IV and RD. It should be so for DiD. Assessing covariate balance is common for RD. It should be so for all shock-based designs.

Second, there will often be value in working to improve covariate balance (see §2.10).

Third, methods with partial compliance – whether DiD, RE, or RD – can also be analyzed as IV, with assignment to treatment used as an instrument for actual treatment. Methods with full compliance are also closely related to IV, with the shock as an instrument for treatment.

We can readily show the similarity between DiD and shock-based IV. For simplicity, assume one has data for two time periods (before and after), ignore covariates, and define an instrument w_{it} for gov_{it}, equal to 1 for treated firms after the shock, 0 otherwise (the same definition of w_{it} we used for DiD). In the first stage of *2SLS*, one predicts *gov* using the instrument w:

$$\begin{aligned} gov_{it} &= \alpha_{1S} + \beta_{1S} * w_{it} + \varepsilon_{1S,it} \\ &= \widehat{\alpha}_{1S} + \widehat{\beta}_{1S} * w_{it} + e_{1S,it} \end{aligned} \tag{17}$$

In the second stage, one estimates, substituting the instrumented variable into the firm fixed effects eqn. (4):

$$\begin{aligned} q_{it} &= \alpha_{\mathrm{2SLS}} + f_i + b_{\mathrm{2SLS}} * gov_{IV,i} + \varepsilon_{it} \\ &= (\alpha_{\mathrm{2SLS}} + \beta_{\mathrm{2SLS}} * \widehat{\alpha}_{1S}) + f_i + \left(\beta_{\mathrm{2SLS}} * \widehat{\beta}_{1S} * w_{it}\right) + e_{it} \end{aligned} \tag{18}$$

The structure of eqn. (18) is identical to DiD eqn. (10). The 2SLS coefficient β_{2SLS} is related to the DiD coefficient δ_{DiD} by:

$$\widehat{\beta}_{\mathrm{2SLS}} = \frac{\widehat{\delta}_{\mathrm{DiD}}}{\widehat{\beta}_{\mathrm{1S}}} = \frac{\text{effect of shock on } q}{\text{effect of shock on } gov} \tag{19}$$

This IV estimate is known as a Wald estimate.

If we add covariates, the DiD and *2SLS* estimators will diverge slightly, because the covariates will affect the first-stage estimate $\hat{\beta}_{1s}$, which is the partial effect of *w* on *gov*, controlling for the **x**'s. But in a credible *DiD* framework, $\mathrm{w}_{\mathrm{it}} \overset{\mathrm{near}}{\perp\!\!\!\perp} x_{it}^{other}$, so the univariate estimate of $\hat{\beta}_{1s}$ should be similar to the multivariate estimate.

Fourth, the same shock can often be exploited using different methods. For example, the 2002 adoption of the Sarbanes-Oxley Act ("SOX") has been used in DiD, ES, IV, and RD designs. When feasible, multiple approaches can be used in a single study. At a minimum, each offers a robustness check for the others.

Fifth, all shock-based designs can benefit from "placebo tests," even if the tests sometimes differ. For DiD, ES, and IV one can apply a placebo shock at different times than the actual shock; for RD, one can test for a discontinuity in the outcome at different thresholds. For all designs, one can test for the absence of an impact on placebo outcomes, that should not be affected by the shock.

Sixth, shock-based causal inference is inherently "local." Shocks affect only part of *gov*, perhaps a small part. This is the only part of *gov* for which one can estimate a causal effect. For RD, credible inference is further limited to firms near the discontinuity; for IV, inference is limited to "complier" firms, whose behavior is changed by the shock.

2.10 Balancing Methods

A core need across methods is for the treatment and control groups to be as similar as possible. Balance can often be improved through a variety of balancing methods developed for pure observational studies, including trimming the sample to common support, and using matching or inverse propensity score reweighting methods. It is beyond our scope to discuss the many balancing methods and how to choose among them.[18] Currently, few shock-based papers use them; this is often a lost opportunity.

Some notes: First, if sample size is an issue, judgment is needed on how far to go in using balancing methods to make the two groups similar, at the cost of making them smaller. Second, trimming implicates the "local" nature of all causal inference. Inference is limited to the post-trimming group, and becomes suspect as one moves away from that group. Third, results that are sensitive to use of a particular balancing method are less reliable than results that are robust on this dimension.

2.11 Outside Our Scope

We leave as outside our scope many important topics in causal inference, including: (i) the role of theory in guiding what causal questions are worth asking, how to ask them, and whether one has met the conditions for credible inference, especially the only-through condition(s); (ii) the extent to which panel data with firm fixed effects (or random effects) and extensive covariates, but no shock, can provide credible causal inference; (iii) the importance of extensive covariates for credible inference; (iv) standard errors, including the need with panel data to cluster on firm or at a higher level, two-way clustering, and handling a small number of clusters;[19] (v) the search by researchers for significant results and its implications for credibility;[20] (vi) selection and survival bias issues affecting which firms enter the data set, and which survive (and how long) during the sample period;[21] (vii) structural model estimation;[22] (viii) interrupted time series designs;[23] and (ix) we advocate combining shock-based and balancing methods, but do not address which

[18] See generally Imbens and Rubin (2015). For trimming to common support, see Crump *et al.* (2009); for matching, see Rosenbaum (2010); for inverse propensity weighting, see Busso *et al.* (2014).

[19] On clustering generally, see, e.g., Bertrand *et al.* (2004); Petersen (2009). On two-way clustering on both firm and time, see Kezdi (2004) (simulations suggest that two-way clustering can be appropriate with as few as 10 observations in the shorter dimension); Thompson (2010) (formulas); Cameron *et al.* (2011) (Stata code cgmreg.ado available on Colin Cameron's website). On the wild cluster bootstrap, ans the likely best available response if one has a small number of clusters, see Cameron *et al.* (2008) (Stata code cgmwildboot.ado available on Judson Caskey's website) MacKinnon and Webb (2016).

[20] See, e.g., Leamer (1978); Glaeser (2006); Harvey *et al.* (2016).

[21] For entry into the sample, firms choose both whether to become public and whether to operate as "companies" or another type of legal entity. On the latter choice, see Demirguc-Kunt *et al.* (2006).

[22] For a recent review, see Strebulaev and Whited (2013). Welch (2012) discusses the value of using quasi-experiments to test structural models. We are not aware of corporate finance examples, but for examples from labor economics, see, Duflo *et al.* (2012); Galiania *et al.* (2015). Coles *et al.* (2012) develop a structural model of how insider ownership affects firm value and use it to assess non-shock based instruments used by others.

[23] See, e.g., Morgan and Winship (2014, §11.1).

balancing methods to use. We do not cover several promising approaches which do not appear in our sample and are thus far rarely used in finance and accounting, including: (x) Bayesian analysis;[24] (xi) "principal strata" approaches to causal inference, which generalize "causal IV" concepts;[25] (xii) sensitivity bounds on treatment effects;[26] and (xiii) (except briefly) synthetic controls. We focus on "internal validity" within the sample and put aside external validity. In related work, we apply the methods advocated here to several, already strong shock-based IV papers in our sample, and show that these methods can lead to large changes in coefficient estimates, and sometimes to the complete disappearance of apparent results (Atanasov and Black, 2016b).

3 Research Designs in Empirical Corporate Governance

We turn here from theory to practice: what do empirical corporate governance researchers do; how often is what they do credible; and what can one learn from reviewing "good practice" papers. To explore what researchers do, we pick 22 major journals in accounting, economics, finance, law, and management, which publish some corporate governance papers. We download reference data (journal, year, volume, pages, title, authors, abstract) and the full text of the article for all academic articles published in these journals from January 2001 through June 2011. The final database consists of 13,461 papers. Table 1 lists the 22 journals and the distribution of papers across journals and years. From these, we identify 863 empirical corporate governance papers, of which 74 use shock-based designs.

3.1 Constructing the Empirical Corporate Governance Sample

We implement two text searches: 1) a search of title and abstract; and 2) a search of the full text of each article. In both searches, we search for a set of expressions related to corporate governance or empirical method, and determine how often each search term appears. We obtain output in the form:

Paper	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6
1	0	20	3	1	0	0
2	1	0	50	0	3	0
3	0	0	1	27	25	5

[24] Shanken and Tamayo (2012) is a recent exception.

[25] See, e.g., Frangakis and Rubin (2002); Frumento *et al.* (2012).

[26] For different approaches, see, e.g., Altonji *et al.* (2005); Hosman *et al.* (2010); Rosenbaum (2010). For a finance implementation, see Black *et al.* (2014).

Journal	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	1H 2011	Total
Accounting Review	29	46	42	46	47	43	43	52	69	72	36	**525**
American Economics Review	183	189	183	173	181	187	190	190	199	223	37	**1,935**
Financial Management	18	21	21	21	23	21	16	31	33	60	9	**274**
Journal of Accounting Research	35	49	28	24	23	29	31	38	37	29	20	**343**
Journal of Accounting and Economics	12	15	37	23	26	29	29	38	25	33	22	**289**
Journal of Banking and Finance	95	101	95	133	131	163	187	215	213	248	124	**1,705**
Journal of Business	22	24	26	46	81	103	–	–	–	–	–	**302**
Journal of Corporate Finance	18	21	28	32	45	35	47	48	39	47	59	**419**
Journal of Empirical Legal Studies	–	–	–	28	16	20	31	32	30	32	15	**204**
Journal of Finance	79	88	93	90	86	87	84	81	77	69	30	**864**
Journal of Financial Economics	62	59	61	76	79	88	103	98	93	101	91	**911**
Journal of Financial Intermediation	11	16	15	18	18	20	20	20	20	20	20	**198**
Journal of Financial and Quantitative Analysis	25	28	37	38	36	37	41	37	54	57	10	**400**
Journal of Law and Economics	32	26	24	22	26	25	28	28	32	24	16	**283**
Journal of Law, Economics and Organization	19	20	20	21	22	21	31	20	22	22	7	**225**
Journal of Political Economy	45	48	45	58	41	38	31	31	31	30	10	**408**
Management Science	116	105	111	149	136	142	134	153	142	141	60	**1,389**
Quarterly Journal of Economics	42	40	40	40	40	40	44	41	43	44	11	**425**
Review Econ and Statistics	67	57	91	77	66	58	58	61	59	77	49	**720**
Review of Finance	11	25	21	17	17	20	21	21	21	23	14	**211**
Review of Financial Studies	38	47	38	37	39	41	59	81	147	119	59	**705**
Strategic Management Journal	52	71	78	68	70	63	73	76	71	74	30	**726**
Total	**1,011**	**1,096**	**1,134**	**1,237**	**1,249**	**1,310**	**1,301**	**1,392**	**1,457**	**1,545**	**729**	**13,461**

Table 1: Surveyed Journals: Total Articles in Our Database

Description: Distribution, across journals and years, of academic articles in 22 selected journals, from January 2001 through June 2011.

Interpretation: Summary statistics for all articles tin our dataset, before narrowing to empirical corporate governance articles.

We identify as potential corporate governance articles all articles that meet one or more of the following conditions:

1. "corporate governance" included in title or abstract;
2. "governance" included 3 or more times in the abstract and text;
3. Two or more of the following groups of terms are mentioned 5 or more times each in the abstract and text:
   - *Board of directors group:* "board" near "director";[27]
   - *Ownership group:* "ownership", "controlling", "blockholder", "minority shareholder";
   - *Shareholder rights group:* "voting", "shareholder right", "activism";
   - *Agency costs group:* "agency cost", "entrench", "private benefit";
   - *Tunneling group:* "tunneling", "self-dealing", "related-party transact", "asset stripping", "expropriat", "freeze-out";
   - *Miscellaneous group:* "investor protection", "anti-takeover", "cross-list", "disclosure"; "compensation".

We generally use flexible searches. For example, we search for two word terms with or without a hyphen between the two words; for longer terms, we search for shorter roots (for example, "expropriat" rather than "expropriation"). We began with a longer list of search terms, but dropped terms that produced a high rate of false positives, such as "board" (without "director" nearby) or "dilut". Our search identifies 1691 potential corporate governance articles, but surely misses some that would be caught by using a looser screen. A full list of search terms and methodology is available from the authors on request.[28]

We review the 1,691 potential corporate governance papers and drop, in order: 190 theoretical and survey papers; 412 papers that are not about corporate governance; 8 case studies; 13 experimental papers; and 55 papers which make incidental use of corporate governance variables as covariates in regressions. This leaves 1,013 empirical corporate governance papers. We likely have some false negatives (corporate governance papers that we wrongly judged not to be

[27] Here and in other searches, to implement "near", we generally check whether term A is separated from term B by 20 or fewer characters, with terms appearing in either order. We experimented with using a larger number of characters, but recovered relatively few additional papers that fit the search concept we were looking for. For some searches using common words, e.g., "difference" near "difference", we limited the separation to 10 characters.

[28] We assessed a range of screens, and progressively loosened them until the hit rate fell below our tolerance for finding needles in haystacks. For example, a narrow filter returned 1,227 papers, of which 902 were "empirical corporate governance" papers (a 74% hit rate). The last loosening increased the number of potential corporate governance papers by 119, of which only 6 were in fact corporate governance papers (a 5% hit rate).

empirical corporate governance papers), but few false positives (papers that we wrongly treated as involving empirical corporate governance). For our principal goals, which are to assess the state of the art of corporate governance research and provide research guidance, most false negatives are unlikely to be centrally about corporate governance, and even more unlikely to reflect good practice.

We then read the 1,013 empirical corporate governance papers and classify their explicit or implicit causal concepts into five categories (some papers involve more than one):

- Purely descriptive (e.g., distribution of ownership of a sample of firms)
- Related to causation:
  - Corporate governance predicts something
  - Corporate governance predicts the relation between *something1* and *something2*
  - One component of corporate governance predicts another component
  - Something predicts corporate governance

Many authors avoid words such as "cause" or "identification", and instead use terms such as "association", or "predict," but the underlying research question involves causation. For example, a paper that uses board structure to predict Tobin's q is typically motivated by the question, "Does a change in corporate governance *cause* a change in firm market value?" Some authors recognize both that the underlying research question involves causation and that their evidence is only indirect, and describe their results as "consistent with" a causal relationship. Some expressly discuss endogeneity, but many do not. Some authors avoid the word "cause" but use near synonyms such as "determine", influence", "affect", or "effect of".

We report the number of papers making each type of causal inquiry in Table 2. We drop 40 purely descriptive papers and 110 papers that only study what predicts corporate governance.[29] This leaves 863 empirical corporate governance papers which involve what one might call "potential causation," from governance to an outcome. Table 3 summarizes the distribution of these papers across journals and years.

[29]We exclude these papers not because the question of what causally predicts governance is uninteresting, but because it is very hard to find papers with meaningful potential for credible causal inference. One likely needs an external shock to a variable that predicts corporate governance. We can imagine such shocks, and the literature may move in this direction, but as yet, it has not. We also neither search for, not include in our final sample, papers that study what might be called "debt governance."

Category	Causal Inquiry (one article can make more than one)	Obs.
1	None (purely descriptive)	40
2	Something causes governance	200
3	Governance causes some outcome variable	730
4	Governance modifies another causal relation	106
5	One aspect of governance causes another aspect	90

Table 2: Corporate Governance Articles: Types of Causal Inquiries

Description: For sample of 1,013 empirical corporate governance articles, types of implicit or explicit causal inquiries in sample of empirical corporate governance papers.

Interpretation: Nature of causal inquiries in empirical corporate governance articles in our sample, before limiting sample to 863 papers with one or more claims in categories 3-5

3.2 Identifying Causal Inference Strategies

We search the 863 "causal inquiry" papers for key research design terms. We search for the following terms and variations (compare our terms to Bowen *et al.* (2016)):

- *General regression terms*: "regress", "least squares", "OLS", "logit", "logistic" "probit", "Tobit", "Poisson", "negative binomial", "multivar", "glm", "hierarch" near "model", "standard (and abbreviations) error", "measurement error", "specification error"; "R^2", "R-squared";
- *Panel data terms:* ["fixed" or "random"] near ["effects" or "panel"]; "panel" near "data" or "regress"; "random coeff"; "longitudinal"; "Fama-MacBeth"; "Breusch-Pagan", "Hausman test";
- *Matching and propensity score terms*: "match" near "sample" or "method" or covar or "character", "propensity score", "peer-adjust", "mahalanobis", "synthetic control", "covariate" near "balance" "propensity" near "weight" or "reweight"; "ignorab", "on observables";
- *Instrumental variable terms:* "instrumental variable" "two-stage" "three-stage"; "method" near "moments"; "simult" near "equation"; "arrelano"; "overidentif"; "valid" near "instrument"; "exclusion restriction"; "Hansen test"; "Sargan"; "2SLS" "3SLS"; "GMM";[30]
- *Heckman selection model terms:* "Heckman" near ["model" or "selection" or "two-stage"];
- *DiD terms:* "triple difference"; "difference" near "difference"; DD; DiD; DID; DiDiD;

[30] We could not use "IV" as a search term because it appears too often as "Section IV" or "Table IV".

Journal	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	1H2011	Total
Accounting Review	0	1	2	4	5	4	6	5	10	8	5	**50**
American Economics Review	1	0	0	0	0	2	0	0	0	1	0	**4**
Financial Management	1	4	2	0	5	1	1	3	6	8	5	**36**
Journal of Accounting Research	0	2	1	5	4	4	1	5	4	2	3	**31**
Journal of Accounting and Economics	0	2	6	3	1	5	7	5	5	9	6	**49**
Journal of Banking and Finance	1	2	12	5	12	7	7	17	32	18	11	**124**
Journal of Business	1	1	1	1	4	5	0	0	0	0	0	**13**
Journal of Corporate Finance	6	8	3	9	13	18	19	21	14	20	23	**154**
Journal of Empirical Legal Studies	0	0	0	0	0	0	1	0	1	1	1	**4**
Journal of Finance	1	7	4	3	7	9	8	5	12	6	3	**65**
Journal of Financial Economics	5	5	10	12	13	12	14	16	14	24	13	**138**
Journal of Financial Intermediation	0	1	0	1	1	2	2	1	1	2	2	**13**
Journal of Financial and Quantitative Analysis	1	0	8	2	2	1	5	2	7	5	1	**34**
Journal of Law and Economics	0	2	2	1	2	1	0	4	0	0	0	**12**
Journal of Law, Economics and Organization	1	0	0	1	0	2	0	0	1	0	0	**5**
Journal of Political Economy	0	0	1	0	0	1	0	0	0	0	0	**3**
Management Science	0	1	0	0	0	0	0	0	1	1	0	**3**
Quarterly Journal of Economics	2	0	1	1	1	1	2	0	1	0	0	**9**
Review Econ and Statistics	0	0	0	0	0	0	0	0	1	0	1	**2**
Review of Finance	0	3	0	0	0	0	0	2	2	2	0	**9**
Review of Financial Studies	3	0	0	2	3	0	3	5	19	12	3	**50**
Strategic Management Journal	4	5	6	5	5	6	6	4	6	7	2	**56**
Total	**27**	**44**	**59**	**55**	**78**	**81**	**82**	**95**	**137**	**126**	**79**	**863**

Table 3: Final Sample: Articles in Which Corporate Governance Potentially Causes Something

Description: Distribution, across journals and years, of 863 empirical corporate governance articles which implicitly or explicitly provide evidence on whether corporate governance causes a change in an outcome variable or in another aspect of governance, or modifies a relation between two other variables.

Interpretation: Summary statistics for empirical corporate governance articles in our dataset, before separating research designs into shock-based and non-shock-based.

- *Event study terms:* "event study", "event window", "normal return", "abnormal return", "CAR"
- *RD terms:* "discontinuity"; "RD"
- *Bias and sensitivity analyses terms:* "endogeneity", "reverse causation", "omitted variable" "hidden bias"; "covariate balance"; "Rosenbaum" or "Manski" near "bounds"; "sensitivity analysis";
- *Bayesian inference terms:* "Bayes"; "Markov chain"; "Gibbs"; "MCMC";
- *General causal inference terms:* "treat! group"; "control group"; "causa"; "potential outcome"; "counterfactual"; "Rubin" near "model"; "selection" near "bias"; or "observable", "endogen"; "exogen"; "SUTVA", "stable unit value"; "ignorab"; "assign" near "treatment"; "construct validity"; "internal validity"; "external validity".

We verify manually that the papers that satisfy a search term actually used the indicated design, rather than, say, citing other papers that used this design or explaining that the design cannot be used with their data.[31]

3.3 Identifying Shocks and Shock-Based Papers

We next searched the 863 potential causation papers for evidence that they relied on a shock, using the following terms and variations:

– *Terms focusing on legal or regulatory change*: ["legal" or "law" or "rule" or "regul" or legis] near ["shock" or "change" or "cutoff" or "threshold" or "new" or "adopt" or "reform"]

– *Terms focusing on specific laws or rules*: ["corporat" or "securities" or "disclos" or "accounting" or "tax" or "bankruptcy" or "insolvency" or "takeover" or "blue sky"] near ["legal" or "law" or "rule" or "regul" or "legis"]

– *Terms focusing on court decisions*: "Delaware" near "chancery" or "court"; ["supreme or "district" or "appeal" or "appell"] near "court"; "court near "decision" or "ruling"' "legal or "court" near "case"

– *More general terms*: "Natural" or "quasi" near "experiment"; "exogen" near ["shock" or "change" or "variation" or "cutoff" or "threshold"]; "crisis"; "collapse";

[31] For each set of search terms, we identify all papers containing these terms three or more times. If the search identifies 20 or fewer papers, we review them all. If the search identifies more than 20 papers, we randomly pick 20 papers for manual review. If all of the sampled papers use the strategy we assume other papers with three or more mentions do as well. We manually review all papers that mention a term one or two times and code whether the paper uses the strategy. We follow a similar approach in identifying "shock" papers.

– *Specific types of shocks:*

  ◦ SOX (also "Sarbanes-Oxley", "Sarbox");
  ◦ Cadbury Committee;
  ◦ Regulation FD (search for "reg" near "FD" or "Fair Disclosure");
  ◦ IFRS
  ◦ Privatization
  ◦ Financial crisis

We then reviewed each paper that satisfied one or more of these searches. This produced 142 papers, which use 50 distinct shocks, in more than 20 countries.

We next identify manually the empirical methods used in the 142 papers that use shocks. A majority of these papers use standard shock-based methods, but far from all. For example, Choi *et al.* (2007) exploit 1999 Korean reforms to the board structure of large firms, which were phased in over 2000 and 2001.[32] They use pooled OLS and find a positive association over 1999-2002 between proportion of outside directors and Tobin's q. But their method is not "shock-based." We also exclude 10 papers that use legal origin as an instrument for *gov*; few scholars today would consider legal origin to be a valid instrument.[33]

We are left with 74 shock-based papers, of which 63 use legal shocks.[34] Table 4 summarizes the principal causal inference strategies used in our sample of empirical corporate governance papers, and how often these papers use shock-based designs. Table 5 provides information on the relative impact of the shock-based versus non-shock-based papers in our sample. The shock-based papers are more recent, on average. We use several metrics of impact: Web-of-Science citations, Social Science Research Network (SSRN) citations, and SSRN downloads. Of the 74 papers, 73 are on Web of Science, and 47 are on SSRN. Per year since publication, the shock-based papers have roughly twice as many downloads per paper as the non-shock papers and 50% more citations from other papers on SSRN; both differences are statistically significant. On Web of Science, the shock-based papers have about 1/3 more citations per year (difference not statistically significant).

Table 6 summarizes the shocks used in these papers; we also plan to publicly post a "shocks database" which includes other shocks useful in corporate finance and accounting research (Atanasov and Black, 2016a). We hope that providing a list of shocks will encourage researchers to search for more shocks. There are

[32]We discuss this shock above; it is also used in Black *et al.* (2006) and Black and Kim (2012).

[33]Legal origin cannot satisfy the only through condition because it predicts many aspects of a country's culture and legal rules, observed and unobserved, which may correlate with *gov* and the outcome variable. See La Porta *et al.* (2008).

[34]We define "legal shocks" broadly to include law-like shocks from sources other than governments: stock exchange rules; accounting rules; and voluntary corporate governance codes. Some shock-based papers use DiD without using the term; we treat these as DiD papers.

Strategy (one article can use more than one)	Total	Shock-based	Legal Shock
Difference-in-Differences	54	36	34
Event study (including long-term)	185	35	27
Instrumental variables	248	8	6
Heckman selection	69	0	0
Regression Discontinuity	2	2	2
Total distinct papers		**74**	**63**

Table 4: Distribution of Causal Inference Strategies

Description: Summary of causal inference strategies used in sample of 863 empirical corporate governance articles over 2001-2011. Shock-based IV papers exclude 10 papers using legal origin as an instrument for *gov*. See appendix for details on the shock-based papers.

Interpretation: This panel shows the relative frequency of the causal inference strategies used by the articles in our dataset, and how often these strategies are shock-based.

	Mean			Median		
Article type	**Shock**	**Non-Shock**	**P-value**	**Shock**	**Non-Shock**	**P-value**
Total through Dec. 2014						
Web of Science Citations	38	45	0.162	25	21	0.777
SSRN Downloads	1,572	1,272	0.328	932	**724**	**0.049*****
SSRN Citation	49	63	0.107	31	27	0.300
Per year since publication						
Web of Science Citations	5.76	5.85	0.887	4.75	3.33	0.144
SSRN Downloads	283.08	203.06	0.099*	198.21	**118.23**	**0.004*****
SSRN Citations	8.21	8.55	0.709	6.76	**4.61**	**0.044*****
Papers in Sample	74	789		74	789	
Papers on Web of Science	73	786		73	786	
Papers on SSRN	47	391		47	391	
Years since publication	5.96	7.04		5	7	

Table 5: Impact of Shock vs. Non-Shock Empirical Corporate Governance Papers

Description: Table shows Web of Science citations, Social Science Research Network (SSRN) downloads, and SSRN citations for 74 shock-based versus 789 non-shock-based papers. Web of Science (SSRN) citations are by other papers on Web of Science (SSRN). All amounts are per-paper. Per year amounts are based on years since publication. p-values are from 2-sample t-test with unequal variances for difference in means, and $\chi 2$ test for difference in medians. *, **, *** indicate significance at the 10%, 5%, and 1% levels. Significant results (at 5% or better) are in **boldface**.

Interpretation: Shock-based papers tend to be newer than non-shock papers. Measured per year since publication, the median shock-based paper posted to SSRN is downloaded and cited more heavily.

Shock	**Papers**	**Distinct Shock Types**
Legal shock	63	32
Sarbanes-Oxley Act and related rules	20	1
Antitakeover laws; judicial decisions on takeover defenses	9	2
Korean governance reform	2	1
Adoption of IFRS	2	1
Cadbury Committee recommendation	2	1
2003 US dividend tax cut	2	1
SEC rule eases delisting by foreign firms	2	1
Financial or economic crisis	3	2
Other:		
Class-action suit (effect on other firms with overlapping directors)	1	1
Election outcome	1	1
Election year	1	1
Gender of CEO's first-born child	1	1
Korea Corporate Governance Fund (effect on non-targeted firms)	1	1
Sudden death of director or CEO	3	1
Total	**74**	**40**

Table 6: Shocks Used for Causal Inference

Description: Summary of shock-based empirical corporate governance papers included in sample of 863 empirical corporate governance articles.

Interpretation: Illustrates the range of available shocks, and the potential for one shock, or type of shock, to be used in more than one study.

likely many good ones yet to be found, with new ones arriving as laws change. We also expect that many of the shocks we list can be used in follow-up projects, either extending the original work or exploring different outcomes.[35] If you find shocks not on our list, please let us know so we can update the database.

Table 7 shows the distribution across journals and years of papers using shock-based research designs. There is a marked increase in the second-half of our period. During 2001-2006, only 4% of papers use shock-based designs; this rises to 11% during 2007-2011. There is also large heterogeneity across journals with percentage of shock-based design papers ranging from 0% to 50%; economics journals have higher percentages.

[35]An example from our own work. We exploit the 1999 shock to governance of large Korea firms in several "finance" papers (Black *et al.*, 2006; Black and Kim, 2012; Black *et al.*, 2013). We were then approached by an accounting scholar who wanted to apply that shock to address the impact of corporate governance on firm financial reporting. See Nasev *et al.* (2016).

Journal	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	1H2011	Shock-based	Total	% shock-based
Accounting Review	0	0	0	0	1	1	0	0	0	1	0	3	50	6.0%
American Economics Review	0	0	0	0	0	0	0	0	0	0	0	0	4	0.0%
Financial Management	0	0	0	0	0	0	0	0	1	1	1	3	36	8.3%
Journal of Accounting Research	0	0	0	0	0	0	0	2	2	1	2	7	31	22.6%
Journal of Accounting and Economics	0	0	2	0	0	1	2	0	0	3	1	9	49	18.4%
Journal of Banking and Finance	0	0	0	0	0	1	0	1	0	1	0	3	124	2.4%
Journal of Business	0	0	0	0	0	0	0	0	0	0	0	0	13	0.0%
Journal of Corporate Finance	0	0	0	0	1	0	4	0	1	1	1	8	154	5.2%
Journal of Empirical Legal Studies	0	0	0	0	0	0	1	0	0	1	0	2	4	50.0%
Journal of Finance	0	1	1	0	0	0	2	0	2	2	1	9	65	13.8%
Journal of Financial Economics	0	0	0	2	1	1	1	1	2	7	1	16	138	11.6%
Journal of Financial Intermediation	0	0	0	0	0	0	0	0	0	0	0	0	13	0.0%
Journal of Financial and Quantitative Analysis	0	0	0	0	0	0	1	0	1	1	0	3	34	8.8%
Journal of Law and Economics	0	0	0	0	0	0	0	1	0	0	0	1	12	8.3%
Journal of Law, Economics and Organization	0	0	0	0	0	1	0	0	1	0	0	2	5	40.0%
Journal of Political Economy	0	0	1	0	0	0	0	0	0	0	0	1	2	50.0%
Management Science	0	0	0	0	0	0	0	0	0	0	0	0	3	0.0%
Quarterly Journal of Economics	0	0	0	0	1	1	1	0	0	0	0	3	9	33.3%
Review Econ and Statistics	0	0	0	0	0	0	0	0	1	0	0	1	2	50.0%
Review of Finance	0	0	0	0	0	0	0	0	0	0	0	0	9	0.0%
Review of Financial Studies	0	0	0	0	0	0	0	0	3	0	0	3	50	6.0%
Strategic Management Journal	0	0	0	0	0	0	0	0	0	0	0	0	56	0.0%
Total	0	1	4	2	4	6	12	5	14	19	7	74	863	
Shock-based as Percent of All Papers	0.0%	2.3%	6.8%	3.6%	5.1%	6.2%	14.6%	5.3%	10.2%	15.8%	8.9%			8.6%

Table 7: Distribution of Papers Using Shock-Based Research Designs

Description: Distribution across journals and years of empirical corporate governance papers which use shock-based research designs. Cells show number of shock-based papers. Last column and bottom row show percent of papers which use shock-based designs.

Interpretation: Table shows relative rarity of shock-based designs, and tendency for these designs to be less uncommon later in our sample period. Table also shows which journals are more likely to publish shock-based papers.

Attention by corporate finance researchers to endogeneity has improved markedly over our sample period (see also Bowen *et al.* (2016)). Early on, endogeneity was often ignored. In the middle of our period, the most common response was an unconvincing IV or Heckman selection approach. Recently, we see more careful causal inference papers; usually shock-based. These papers tend to appear in more highly ranked journals. But many papers in top journals still use unsatisfying methods, and some promising methods are rarely used. Even our "good practice" papers often fall short of what one might have done.

3.4 Use of Multiple and Combined Designs

We stress above the value of using different research designs to exploit a single shock, as well as using combined designs (such as DiD plus matching or DiD plus RD). Only a few papers in our sample do so.[36]

4 Shock Based DiD Designs

We begin our tour of research designs with DiD. This design was rare early in our sample period, but its use has been increasing over time.[37] DiD, together with its near-cousin, the event study, is the workhorse of shock-based design in finance and accounting research. We provide more complete details than for other designs, both because we discuss DiD first and because there is no good survey in the methods literature.[38]

4.1 DiD Designs in Our Sample

Table 8 provides summary information on the DiD papers in our sample. We find 52 papers that say they use a DiD design, or do so without using this term. Of these, only 36 have a plausibly exogenous shock. Some of the remaining studies have firm (and time) fixed effects and no more, or matching and no more. In our view, these should not be called DiD designs. Some base a DiD analysis on a firm-chosen

[36] Iliev (2010) uses combined IV/RD and event study/RD designs, and separately uses RD and a bit of DiD. Black *et al.* (2006) use a combined IV/RD design. And four papers use DiD and an event study separately.

[37] The earliest DiD paper in corporate finance we know of is French and Roll's (1986) study of the effect of stock exchanges being open on share price volatility. The authors do not use DiD terminology, but have a DiD design, where the exogenous shock is the New York and American Stock Exchanges being closed on Wednesdays for the second half of 1968 (due to a paperwork backlog); treated days are these Wednesdays, and control days are other non-weekend, non-holiday days. We thank the editor for this reference.

[38] The best discussions we know of are Angrist and Pischke (2009, §5.2) and Imbens and Wooldridge (2009, §6.5). Both are on the thin side.

event, such as cross-listing on a foreign stock exchange, or replacement of the CEO. In our view, these are not "shock-based" designs.[39]

A number of papers use a shock which affects all firms, so there is no true control group, but affects some firms more than others. The research design involves looking for after-minus-before differences in strongly-affected versus mildly affected firms, either by "binning" the sample based on sensitivity to the shock or using a continuous measure of sensitivity. These designs have no standard name, we call them "DiD-continuous."

Of the 36 shock-based DiD papers, 34 rely on legal shocks. There are 27 "true DiD" and nine DiD-continuous papers. A significant percentage of the true DiD papers (11/27) also assess how the outcome varies based on the sensitivity of treated firms to the shock – an approach that one can call "DiDiD-continuous" or "DiD plus sensitivity." Twelve have multiple shocks (of these, seven study antitakeover laws, adopted by different states at different times). Two are triple-difference designs. Table 9 provides details on the designs used in the DiD papers.

4.2 Elements of Shock-Based DiD Design

Ideally, a legal shock can approximate a randomized experiment, by applying a legal rule as-if-at-random to some firms, but not to similar firms in the same or other jurisdiction. However, most legislatures and regulators don't regulate at random. Thus, a core challenge is to justify the as-if-random nature of the shock (Rosenzweig and Wolpin, 2000). We discuss below the main credibility concerns and how to address them.

4.2.1 Is the Shock Truly Exogenous?

An initial concern is whether the shock is sufficiently exogenous – is the only apparent difference between treated and control firms that some were treated and others were not? This concern can be illustrated with a counterexample. Suppose that some firms lobby for and receive a favorable legal rule, while other apparently similar firms do not. Treated and control firms may well differ on unobservables – including unobservables that directly relate to the benefit they receive from the rule. That need not invalidate the DiD design, which estimates only ATT, not ATE. But the lobbying efforts would provide reason for caution – both in inferring whether control firms would have realized similar gains, if treated, and on whether

[39] We exercised judgment as to when authors claimed to use DiD. An example is Malmendier and Tate (2009), who compare CEOs who receive external awards to similar CEOs who don't, and study how the award affects firm behavior. We classified this as a pure observational study even though they use the term "difference in differences" once, to refer to a single regression estimate. We focus here on true panel datasets. Similar strategies can be used for "repeated cross section" data, with different units observed in each time period. There are no repeated cross section papers in our dataset.

Year	Uses DiD or claims to	Shock-based	Legal shock	DiD-continuous	Citations to Shock-based DiD papers
2001	0	0	0	0	
2002	1	1	1	0	Dahya *et al.* (2002)
2003	3	3	3	2	**Bertrand and Mullainathan (2003)**; Lo (2003)*; Nagar *et al.* (2003)*
2004	1	1	1	1	Ryan Jr. and Wiggins III (2004)*
2005	4	4	3	1	**Dinc (2005)**; Altamuro *et al.* (2005); Chetty and Saez (2005)*; Aivazian *et al.* (2005)
2006	4	1	1	0	Rauh (2006)
2007	4	3	3	1	Gomes *et al.* (2007)*; Dahya and McConnell (2007); Carvalhal da Silva and Subrahmanyam (2007)
2008	3	2	2	0	Daske *et al.* (2008); Hope and Thomas (2008)
2009	12	8	7	1	Low (2009); Chhaochharia and Grinstein (2009); Cicero (2009); Faccio and Parsley (2009)*, Gao *et al.* (2009); Qui and Yu (2009); Kamar *et al.* (2008); Yun (2009)
2010	13	9	9	2	Altamuro and Beatty (2010); Atanasov *et al.* (2010)*; Bargeron *et al.* (2010); Chava and Purnanandam (2010)*; Engel *et al.* (2010); Giroud and Mueller (2010); Iliev (2010); John and Litov (2010); Wang (2010)
2011	7	4	4	1	Byard, Li, and Yu (2011); DeFond *et al.* (2011)*; Francis *et al.* (2011); Kinney and Shepardson (2011);
Total	**52**	**36**	**34**	**9**	

Table 8: Difference-in-Differences Designs over Time

Description: Summary of papers that use or claim to use DiD designs, including "DiD-continuous" papers where all firms are exposed to the shock, but have differing sensitivities to the shock. * indicates DiD continuous design.

Interpretation: Table shows frequency over time of DiD designs. **Boldface** indicates our "good practice" papers. Most DiD studies are based on legal shocks.

Research Design Elements	**Total Papers**
DiD is main research design	26
True DiD (true control group exists)	24
DiD-continuous design	9
Also uses "binned" groups	6
DiDiD	
True DiDiD	2
DiDiD-continuous (DiD plus sensitivity to shock)	10
DiDiD-double continuous	1
Aspects of DiD Design	
Control group is pre-rule compliers	4
Multiple shocks at different times	12
Summary statistics separately for treated and control	29
Formal test for covariate balance	6
Combined Designs	
DiD plus (crude) matching	3
DiD plus careful balancing methods	0
DiD plus RD	0
Uses panel data (not just pre-post)	31
Assess whether pre-shock trends are parallel	4
Assess robustness to non-parallel pre-treatment trends	0
Distributed lag model for treatment effects	2
Placebo tests	8
Different control group	5
Placebo shock outside treatment period	3
Placebo outcome variable	1

Table 9: Details on Shock-Based DiD Papers

Description: Table summarizes selected aspects of research design for the 36 shock-based DiD papers in our sample.

Interpretation: Table shows which aspects of DiD design are commonly or rarely used. Several aspects that should be standard, such as testing for parallel pre-shock trends, are rare.

those unobservables might have produced non-parallel trends in the post-shock period, even without the new rule.

Thus, an aspect of shock credibility is assessing whether the treated firms favored or opposed the rule. An ideal rule would be one adopted for other purposes, which affects some firms as an unintended byproduct. The Desai and Dharmapala (2009) shock-based IV study, discussed below, of how check-the box tax rules, adopted for small private firms, affected tax planning by multinational firms, offers a good example. Conversely, doubt about whether lawmaking was as-if-random increases the need, already strong, to ensure covariate balance and confirm parallel pre-treatment trends.

Even if firms cannot choose whether to be shocked, if they know the shock is coming, they may modify their behavior in advance of the shock. This will affect – and often attenuate – the observed effect (Hennessy and Strebulaev, 2015; Malani and Reif, 2015). But anticipation can affect estimated treatment effects in more pernicious ways, including sign reversals. For example, if firms anticipate a large shock (a tax change, say) and the actual shock is smaller than expected, the estimated treatment effect can have the opposite sign from the true effect (Hennessy and Strebulaev, 2015). Attenuation can also occur if firms expect that the shock may be temporary.

4.2.2 *Checking for Covariate Balance (Including Common Support)*

If a shock was truly as-if random, the treated and control groups should be similar on a broad range of covariates, measured pre-shock. The papers in our sample vary greatly in the care with which they assess the risk of non-random assignment to treatment. On the positive side, 29/36 provide separate summary statistics for the treated and control groups. Five papers address balance indirectly by using multiple control groups. However, only three papers use matching to improve balance, and in all the matching is crude. None uses either careful balancing methods or a combined DiD/RD design.[40] None of the DiD papers – indeed, none of the shock-based papers – uses the term "covariate balance" or confirms common support.[41]

4.2.3 *Checking for Pre-Treatment Trends*

Even if observed covariates are well-balanced, one still needs to worry about pre-treatment trends. If treated and control firms have different pre-treatment trends for the outcome variable, then without the treatment, those trends might have continued, stopped, or reversed (regression to the mean). Pre-treatment trends can also influence why some jurisdictions adopted rules, while others did not, a circumstance that Besley and Case (2000) call an "unnatural experiment."

With only two periods, one before and one after the shock, we can't assess whether pre-treatment trends are parallel. This is an important weakness in the two-period DiD design, and puts great stress on similarity between treated and controls. If multiple pre-shock periods are available, we still can't test for parallel *changes* from just before to just after the shock, but we can test for parallel *trends* prior to the shock:

$$\text{DiD credibility req. 3 (parallel trends): } \mathrm{w_i} \perp\!\!\!\perp (y_{0i,t} - y_{0i,(t-1)}) \forall t \leq 0 \qquad (20)$$

With panel data, a good way to assess whether pre-treatment trends are parallel is with a "leads and lags model" (our term).[42] In DiD eqn. (10), replace

[40]We discuss combining DiD with balancing methods in §4.6, and combined DiD/RD designs in §7.

[41]A few non-shock papers address covariate balance, and provide useful examples. See, e.g., Armstrong *et al.* (2010); Lin and Su (2008), Murphy and Sandino (2010), and Stuart and Yim (2010).

[42]We have heard this called an "Autor" model, following Autor (2003).

the treatment dummy w_{it} with a family of year-specific variables w_i^k, each = 1 for treated firms in period *k* (including both pre-and post-treatment periods), and 0 otherwise. Without covariates:[43]

$$y_{it} = \alpha + f_i + g_t + \sum_{k=-n_{pre}}^{n_{post}} \left(\delta_{DiD}^k * w_i^k\right) + \varepsilon_{it} \tag{21}$$

The w_i^k should be small and insignificant during the pre-treatment period, with no apparent trend. During the post-treatment period, they will map out the treatment effect over time.

This model lends itself to graphical interpretation. Figure 1, drawn from a study by one of us, provides an example in which it is visually clear that there were no differences in pretreatment trends between treated and control units. Data permitting, the leads-and-lags graph should cover an extended pre-treatment period. In our experience with annual data, what appears to be random noise with, say, three pre-treatment years, can look like a trend, if one adds more pre-treatment years.

The conditions for a good shock, especially untestable condition 3 (as-if-random assignment to treatment) and the related but testable condition 4 (co-variate balance and common support) imply that the parallel trends assumption should be satisfied. Conversely, parallel pre-treatment provide evidence supporting as-if-random assignment to treatment. However, covariate *imbalance* increases the risk that the parallel trends assumption will be violated, even if pre-treatment trends appear reasonably parallel.

Only four shock-based papers, of which two are DiD papers (Dahya and Mc-Connell, 2007; Rauh, 2006), address differing pre-treatment trends by showing a multi-period graph of trends for treated versus controls, before and after the shock. Two additional papers show pre-treatment results, but only for one period ((t-2) to (t-1)).

One can also apply placebo shocks. This involves using only pre-treatment data, and applying fake shocks at different times during the pre-treatment period.[44] The period-specific outcomes should be similar (and close to zero) before and after the fake shock. In Figure 1, imagine dropping the post-shock data points and moving the vertical line separating the pre- and post-periods to an arbitrary point in the pre-treatment period. Given a “clean” leads-and-lags graph such as Figure 1, with no apparent pre-treatment trends, once can predict that fake shocks will be insignificant. Three DiD papers in our sample don’t show pre-treatment trends, but do use a placebo shock.

[43] One period must be omitted and becomes a reference period; a good choice is often several period before the treatment, say $k = -3$.

[44] One can also apply a placebo shock in the post-treatment period, after the shock effect has been fully felt. Dinc (2005), discussed below as a good practice DiD paper, provides an example.

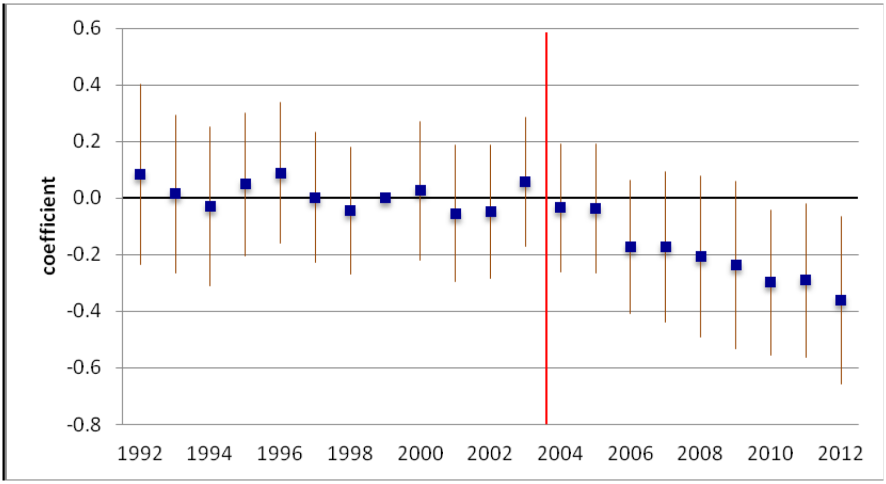

Figure 1: Leads and Lags Model Example (from Paik *et al.* (2013))

Description: Coefficients and 95% confidence intervals (based on heteroskedasticity-robust standard errors) from regression of *ln* (large paid medical malpractice claims per 1,000 physicians) on year and state fixed effects, interactions between tort-reform dummy (=1 for states that adopt damage caps, 0 for no-cap) and year dummies (with 1999 as the omitted year), and constant term, for 197,979 large paid claims reported to NPDB over 1992-2012 in 12 reform states and 20 no-cap states. Regressions are weighted by the average number of physicians in each state over 1992-2012. Vertical line separates pre-reform period from principal reform period, which begins in 2003.

The leads-and-lags approach can be usefully applied to covariates $\mathbf{x}_{it}$. Simply replace y_{it} in Eqn. (13) with the x_{it}, one at a time, perhaps controlling for the remaining covariates. A good shock should produce reasonably parallel pre-treatment trends for all covariates. For covariates that should not be affected by the treatment, this approach also provides a set of placebo outcomes – for these covariates the lagged coefficients on the interaction term, in the post-treatment period, should remain close to zero.

4.2.4 Assessing Robustness with Firm-Specific Trends

In addition to assessing whether there appear to be non-parallel pre-treatment trends, one can include firm-specific trends in the DiD model, and see whether the results survive. One can also use trends for groups of firms. For example, one might use industry trends or, in a multicountry study, country trends. None of our sample papers do this. Yet given a reasonable number of pre-periods, this is often a sensible robustness test.

To allow for pre-treatment trends, start with the DiD panel data eqn. (10) and interact the firm dummies with a linear time trend for each firm:[45]

$$y_{it} = \alpha + f_i + g_t + (\gamma_i * f_i * t) + (\delta_{\mathrm{DiD}} * w_{it}) + \varepsilon_{it} \tag{22}$$

Experience in other areas suggests that including unit-specific trends will kill a fair number of DiD results.[46] Yet unit or group trends should be used with care. Suppose that one has random fluctuations in y_{it} over time in the pre-treatment period, with no true trend. Including unit-specific trends will turn those random fluctuations into an estimated trend. In our experience, this inflates the standard error for δ_{DiD} and often pushes the δ_{DiD} coefficient around a fair bit. The shorter the pre-treatment period, relative to the post-period, the more likely these effects are. In effect, if the γ_i tend to have the same sign as δ_{DiD}, the f_i*t terms compete with w_{it} as explanations for the post- minus pre-treatment differences in y_{it}.

One generally should not include unit time trends in a main DiD model. They can be a useful robustness check but only that. The DiD design assumes that trends were parallel for the treated and control groups pre-treatment, and would have remained so without the treatment. By adding unit trends, one assumes instead that any *non-parallel* pre-treatment trends would have continued without the treatment. But the pre-treatment trends could be noise, or reflect pre-treatment forces which would not have continued post-treatment, and might even have reversed (a form of regression to the mean). Unless one understands the substantive reason for the trends, there is little basis to believe that any pre-treatment trends would have continued in the post period, but for the treatment.

If one finds non-parallel pre-treatment trends, more careful balancing of treated and control firms can help, as can finding additional covariates that absorb (in effect, explain) the non-parallel trends. Short of that, non-parallel trends are a severe blow to the credibility of a causal claim, with no obvious solution.

4.2.5 Is the Shock Isolated?

A further concern with non-random lawmaking is that a legislature that adopts rule A might also adopt related rule *B* at roughly the same time, and rule *B* (or *A* and *B* together, or a broader set of policies *P* that includes both) actually causes any change in outcomes. The influence of a set of policies, adopted at different times, might be reflected in differing pre-treatment trends, but there is no guarantee of this. In effect, even if A, B, and P are exogenous to the firm, they are not

[45] This equation uses both pre- and post-shock data to estimate the trends. A variation, given a long enough pre-shock period, is to estimate the trend coefficients using only pre-shock data. This approach is more consistent with the general dictum of causal inference that one should not control for post-treatment variables that might be affected by the treatment.

[46] Angrist and Pischke (2009, §5.2.1), discuss both the leads-and-lags model and including unit trends in a DiD model, and provide an example where unit-specific trends kill an effect that seems strong without them.

exogenous to each other. To conclude that rule A caused a change in outcome y, we must rule out the B and P channels.

If a discrete alternative can be identified, one can sometimes run a horse-race between the two explanations.[47] But often, one can only search for confounding shocks or policies and discuss whether the only through condition appears to be valid.[48]

4.2.6 Broad versus Narrow Shocks and the Only Through Condition

Some shocks are narrow and affect firms in a single, identifiable way. Others can affect firms in multiple ways. For example, 20 papers in our sample use the adoption of SOX as an exogenous shock. SOX affected firms in many ways. To assess whether a particular aspect of SOX affected firms, one must rule out other channels. Sometimes this will be possible, by ensuring that treated and control firms are similar in all aspects of SOX-compliance other than the one being studied. For quantitative criteria (majority of independent directors, say), one can combine DiD with an RD or similar design, in which one compares firms that are close to but below the compliance threshold to similar firms close to but above the threshold.

Most of the SOX-based studies in our sample did not take either of these steps. This leaves them vulnerable to the concern that another aspect of SOX, or another difference between treated and control firms, explains the post-SOX differences that they find.[49]

4.2.7 Pre-Treatment Compliers as the Control Group

A number of DiD papers involve legal shocks that apply to all firms. Here, the control group is drawn from firms which already complied with the new rule (pre-

[47] Consider Black and Kim (2012), who study a 1999 legal shock to the governance of large Korean firms. They find no other contemporaneous (or nearly so) laws that might explain the outcome (higher Tobin's q for large firms which were subject to the governance reform). Still, most large firms belong to major Korean business groups, called *chaebol*. Perhaps the large firm reforms signaled that the government would crack down, in unspecified ways, on *chaebol* firms. Black and Kim address this risk by running a horse-race between a large-firm dummy and a *chaebol* dummy, to see which better predicts their results; the large-firm dummy wins when it should.

[48] The shock-IV paper by Desai and Dharmapala (2009) provides a good example of such a discussion. Desai and Dharmapala use a setting where an as-if-random effect of reform is likely: a tax law change that was intended to affect private firms, which had the unintended consequence of facilitating tax avoidance by multinationals. The strategy of studying unintended consequences of legal shocks is generalizable.

[49] Compare the SOX study by Donelson *et al.* (2012), outside our sample, who study the effect of the SOX requirement that public firms have majority-independent boards on financial fraud, by comparing non-compliant firms to already-compliant firms. The authors perform several checks for alternate explanations, including assessing whether non-compliant firms which were exempt from SOX experienced similar changes in fraud rates.

rule compliers), and the treatment group is pre-rule noncompliers. An example is Dahya and McConnell (2007), who study the impact on UK firms of the early 1990s Cadbury Committee recommendation that all public firms have at least three non-executive directors (NEDs). The treated firms are those which previously had 0-2 NEDs, the control firms already had 3+ NEDs.

A core concern with this design is that the compliers may differ from noncompliers in various ways, both observed and unobserved, that could lead to violation of parallel trends. Parallel trends in the pre-treatment period help, but are not dispositive. If the compliance variable has more than two values, it can help to narrow the treatment and control groups, to study firms that are as similar as possible. For example, in the Dahya and McConnell study, one might compare treated firms with 2 NEDs pre-Cadbury versus control firms with 3 NEDs.[50] This is similar to a combined DiD/RD design, with a discrete forcing variable (number of NEDs). It will approach a true combined DiD/RD design if there are many available values of the compliance variable.

4.2.8 Shock Strength

As we discuss in Part 2, one wants to ensure that the shock is "strong" – it meaningfully changes the forced variable. A weak shock makes it more likely that an unobserved factor could be driving the observed results. With panel data and an observed forced variable, one can test for shock strength by regressing the forced variable on the treatment indicator and covariates:

DiD credibility requirement 4 (shock strength):

$$x_{it}^{forced} = \alpha + f_i + g_t + (\eta * w_{it}) + (\mathbf{b}_{it} * \mathbf{x}_{it}) + \varepsilon_{it} \tag{23}$$

The coefficient η on the shock should be statistically strong and economically meaningful. The shock strength condition in eqn. (23) is similar to testing for instrument strength in IV.

4.2.9 Shock Strength with a Latent Forced Variable

Some DiD studies involve a latent forced variable. These studies face a joint "shock strength" and causal channel challenge – did the shock really change firm behavior, in the expected direction? They rest on a claim that shock *A* changed an unobserved intermediate outcome *u* for treated firms, which then changed an observed outcome. For example, Rauh (2006) and Low (2009) assume that Delaware judicial decisions in 1995, principally *Unitrin v. American General,* broadened permissible takeover defenses and led target managers to feel more

[50] The discussion in text oversimplifies the Dahya and McConnell (2007) design, in which they also compare firms with less than 3 NEDs which choose to comply with the Cadbury recommendation to firms which choose not to.

secure (unobserved intermediate outcome). This is akin to an IV design where the instrumented variable (strength of takeover defenses) is unobserved. The strength of the first stage is assumed rather than shown.

For these two papers, first-stage strength is possible, but not self-evident and not well-defended. Both rely on Subramanian (2004), a law professor who asserts that *Unitrin* (and perhaps other 1995 cases) strengthened takeover defenses. But other authors are more equivocal (e.g., Gilson and Black, 1995, p. 894-895). What might provide supporting evidence? A post-1995 drop in hostile takeover bids might do the trick. So let's look. The number of hostile tender offers by year over 1991-1995 was 2-2-3-10-11. The number over 1996-2000 was 8-14-22-16-14. The assumption that *Unitrin* strengthened takeover defenses and thus discouraged hostile bids is not supported.[51] In contrast, Giroud and Mueller (2010) do examine shock strength. They find that state adoption of an antitakeover law leads to lower takeover probabilities, but only in competitive industries.

To generalize: DiD analyses based on a weak shock are suspect, much like IV analyses with a weak instrument. Often, the effect of the shock on firm behavior will be clear, but not always, as the Delaware example shows. One can assess credibility by checking for other effects that one would expect to find (or not find), if the shock operated as posited.[52] Use of IV forces the researcher to assess instrument strength in the first stage. A similar check for shock strength should be part of DiD analysis, both when the shocked variable is observed and when it is not (but one can test for the presence of follow-on effects).

A second lesson from the Rauh (2006) and Low (2009) example involves the value of research design preceding analysis, when feasible.[53] Subramanian (2004) found a "disappearing Delaware effect" (higher Tobin's q's for Delaware firms) around 1995, sought an explanation, and developed a plausible story about stronger takeover defenses after the *Unitrin* decision. Ironically, Litvak (2014) shows that there was never a Delaware effect in the first place. In our view,

[51] Source: *Mergerstat Review* (various years). We did not separately study Delaware firms but the trend toward *more* hostile bids in the second half of the 1990s is strong enough so that further analysis would be unlikely to change the conclusion that there is no evidence that 1995 Delaware takeover decisions suppressed takeovers, as Rauh and Low assumed. Heron and Lie (2015) also find little effect of 1995 Delaware decisions on takeovers.

[52] An example from our own work: Atanasov *et al.* (2010) study the effect of 2002 Bulgarian legal reforms on tunneling through dilutive equity offerings and freezeouts. An initial step in the analysis is to show that the reforms affected offerings and freezeouts. We report that: (i) highly dilutive offerings are the norm pre-reform; offerings are used to raise capital post-reform; (ii) the mean freezeout price/sales ratio is 0.15 pre-reform, and jumps to 0.65 post-reform. Thus, the reforms had a strong impact on firm behavior. Catan and Kahan (2016) criticize studies of antitakeover statutes for having a weak implicit first stage.

[53] As Rosenbaum (2010, p. 7) recommends, "Before examining outcomes that will form the basis for the study's conclusions, a written protocol [should describe] the design, exclusion criteria, primary and secondary outcomes, and proposed analyses." See also Cochran's (1965, p. 236) advice that "The planner of an observational study should always ask himself the question, 'How would the study be conducted if it were possible to do it by controlled experimentation?'"

Subramanian looked too hard for an explanation for his results, and Rauh and Low accepted his explanation too uncritically. Before relying on subtle law stories, non-experts would do well to vet them with experts. The finance coauthor of this paper (Atanasov) assessed Low (2009) as a likely "good practice" paper – she does many things well.[54] The law author (Black) questioned shock strength, looked for data on takeover rates, and the data was not consistent with the assumed channel.

4.2.10 SUTVA Independence

All causal inference designs assume SUTVA independence. Yet, none of our shock-based papers discuss this assumption. We offer here some examples of why one should worry, but have no good solutions to offer.

Suppose we want to study whether a change in *gov* increases firm efficiency and thus profitability. If *gov* increases efficiency, and the treated firms are a small subset of all firms in an industry, we might observe an effect on profitability. But if the treated firms are a large fraction of all firms, they will compete away the efficiency gains. Consumers will benefit, but profits at treated firms may not rise.

Suppose we want to understand the impact on firm value of a change in disclosure rules.[55] Disclosure rules can have externalities. Investors may trust disclosure for all firms if most improve their disclosure, yet distrust a change in disclosure by a single firm, because they worry about adverse selection – firms will tend to disclose what makes them look good. Disclosure by some firms in an industry can also help investors monitor other firms in the industry. This is a positive externality that may affect the value of the control firms, and thus reduce an estimated treatment effect. But the externality could go the other way. The decisions by some firms to cross-list in the U.S., thus committing to improved disclosure and perhaps signaling low intent to engage in self-dealing, will send a negative signal about firms that do not cross-list. A comparison of Tobin's q for cross-listed versus non-cross-listed firms will then overstate the effect of cross-listing.

4.2.11 Attrition

A concern for any DiD design is differential attrition for the treated and control groups, especially, the risk that the treatment could cause attrition to differ systematically between treated and control units. The parallel trends assumption includes an assumption that attrition will be similar in both groups but for the treatment. This can be tested in the pre-treatment period. If the treatment (or,

[54] Rauh (2006) and Low (2009) are the only "true DiDiD" papers in our sample and use a sensible third difference (existence of a staggered board). Low is one of only three DiD studies that use matching to improve covariate balance. She also assesses (though without reporting results) leads and lags of the shock and reports finding significant effects only in the first two post-shock years.

[55] See Greenstone *et al.* (2006), discussed below as a good-practice event study.

the prospect of future treatment) induces differential attrition, this can introduce bias. Only 7 DiD papers in our sample assess attrition.[56]

4.2.12 Placebo Tests

A number of "placebo tests" can be used to assess the credibility of the DiD design. First, as discussed in §4.2.2, one can place a placebo shock at an arbitrary time (or times) during this period and see if this artificial shock predicts outcomes. One can scramble which firms are treated and which are control.[57] If one has two different control groups, one can use one as a pseudo-treatment group and the other as pseudo-controls; one can also divide a single control group into pseudo-treated and pseudo-controls. And one can study outcomes that should not be affected by the shock. All of these placebo tests should produce null results.

In our DiD sample, only a few papers use placebo tests. Three papers apply a placebo shock during the pre- or post-treatment period; five apply placebo shocks to an alternate control group; one uses an outcome variable that should not be affected by the treatment.

4.3 Controlling for Covariates

Thus far, we have presented DiD estimates of treatment effects without controlling for covariates $\mathbf{x}_{it}$. Which covariates to include in a DiD design is a nuanced question, especially if one has panel data with multiple pre- and post-treatment periods. We consider here selected aspects of this issue.

4.3.1 Pre-treatment Covariates

Pre-treatment covariates, measured at or just before the time of the shock, are safe to include; they will rarely cause bias. With a panel data structure and firm fixed effects, they are useless, because they will be absorbed by the firm-fixed effects. But they can be worth including if one uses first differences (FD). In Atanasov and Black (2016b), we re-examine Duchin *et al.* (2010), who use an FD design, and show that their shock-based IV results disappear if one controls for an important pre-treatment covariate.

[56]Three papers show that treated and control firms have similar exit rates (Iliev, 2010; Altamuro and Beatty, 2010; Greenstone *et al.*, 2006). Four reestimate their DiD models with a balanced panel that excludes firms which enter or exit the sample (Altamuro and Beatty, 2010; Chetty and Saez, 2005; Giroud and Mueller, 2010; Lo, 2003). Two papers impute observations missing due to attrition using data from similar surviving firms (Dinc, 2005; Greenstone *et al.*, 2006).

[57]If one repeatedly randomizes treated and controls, the distribution of pseudo-treatment effects provides a way to estimate standard errors. See Conley and Taber (2011), Ho and Imai (2005), and Rosenbaum (2010, ch. 2) Randomization inference can also provide a way to estimate standard errors for DiD (or ES) with a small number of treated firms (Conley and Taber, 2011; Gelbach *et al.*, 2013).

The ability to easily add pre-treatment covariates can provide a practical reason for using FD instead of FE. If one uses FE and lacks balance on pre-treatment covariates, one can often attain reasonable balance through balancing methods. We discuss some nuances of what to balance on below.

4.3.2 Time-varying Covariates with Panel Data

Assume that the case for as-if random assignment is strong. Pre-treatment and time-invariant covariates will be captured by the firm fixed effects. Should one include post-treatment, time-varying covariates in the DiD equation (9) or (10)? Suppose we can divide the time-varying covariates into "affected" covariates (potentially affected by the treatment) and "unaffected" covariates (unlikely to be unaffected by the treatment). Including unaffected covariates can increase precision, will not introduce bias, and can increase confidence that the treatment, not some other difference between the treatment and control groups, caused the observed difference in outcomes.

Including affected covariates, in contrast, can bias the estimated treatment effect. One might call this "included variable bias." Suppose, for example, that the treatment dummy covaries with a covariate x, and x covaries with the outcome. Including x in the regression will bias the coefficient on the treatment dummy, often toward zero (Gerber and Green, 2012). Unfortunately, it is often unclear which covariates are potentially affected by the treatment.

Which covariates to include becomes a harder question if there is a real risk of non-random assignment to treatment. Consider a panel data design. One can include covariates in levels, changes, or both. With both, DiD eqn. (10) becomes:[58]

$$y_{it} = \alpha + f_i + g_t + (\delta_{\text{DiD}} * w_{it}) + \left(\eta_{level} * \mathbf{x}_{i,pre}\right) + \left(\eta_{change} * \mathbf{\Delta x}_i\right) + \varepsilon_{it} \qquad (24)$$

Including time-varying covariates can reduce the importance of non-parallel trends, one cares only about remaining trends, conditioned on the covariates. If these covariates might be affected by the treatment, one must trade off bias from non-parallel trends versus bias from controlling for affected covariates. Consider, for example, the Black and Kim (2012) study of a legal shock to the board structure of large Korean public firms. The treated firms must have 50% outside directors, an audit committee, and an outside director nominating committee. In a combined DiD/RD design, mid-sized firms, just below the size threshold in the law, provide a control group; they can have only 25% outside directors, and do not face the committee requirements. Suppose that (i) board structure affects Tobin's q directly,

[58]None of the DiD papers in our sample includes both levels and trends in covariates. Desai and Dharmapala, an IV paper, does so. See also Black and Kim (2012) (including the forcing variable in DiD regressions in both levels and changes). Whether to include both involves the substantive issue one is studying, including: (i) is there reason to think that both levels and changes in a particular covariate could predict the outcome; and (ii) is the sample large enough so that the loss in degrees of freedom is a small cost?

(ii) disclosure also affects Tobin's q, and (iii) the board structure shock causes treated firms to improve their disclosure. Controlling for disclosure will lead to underestimating the effect of the board structure shock on q. But if large firms would have improved their disclosure (relative to controls) around the time of the shock, even without the shock, then *not* controlling for disclosure would overestimate the causal effect.

There is no ideal solution; one needs context-specific judgment, informed by theory, plus robustness checks for "suspect" covariates – ones for which the case for inclusion is unclear. One robustness check for the suspect covariates is to report results with and without them. Another is to run leads-and-lags graphs with the suspect covariates substituted for the outcome variable. If both pre- and post-treatment trends are parallel, this strengthens the case for including the covariate.

Our judgment is that when non-parallel trends are a real risk, one should often include time-varying covariates, other than those that are likely to be directly affected by the treatment, or explain the choice not to do so. A result without these covariates, which weakens with them, is suspect.[59]

The DiD studies in our sample generally do not say much about reasons for choosing covariates; none assesses robustness to different choices of covariates. Taking both steps would increase credibility. In particular, a result which is robust to choice of covariates is more likely to be robust to unobservables as well. This is an opinion, not a theorem. Conversely, if a result is sensitive to choice of covariates, this is evidence of covariate imbalance and indicates need to use extensive covariates to limit omitted variable bias, and perhaps to use formal sensitivity bounds.[60]

4.3.3 Improving Covariate Balance with Panel Data

Suppose one has panel data with a number of pre-treatment periods, and wants to use balancing methods to improve covariate balance, as we recommend above. What should you balance on? The possibilities include: balancing on covariates for the most recent pre-treatment year, balancing on covariates across multiple pre-treatment years; and balancing on the most recent year plus trends. We are aware of no guidance in the methods literature, nor of papers that directly address this issue.

[59]We depart here from the standard advice in the causal inference literature, which is to control only for covariates that are clearly *not* affected by the treatment. This is the right advice if assignment to treatment is unconfounded, as this literature assumes. If non-parallel trends exist in the data, limiting them through balancing methods or capturing them through covariates can improve credibility.

[60]In our sample, only two non-shock-based studies use these bounds. Armstrong *et al.* (2010) use Rosenbaum (2010) bounds; Broughman and Fried (2010) use Altonji *et al.* (2005) bounds. Outside our sample, Black *et al.* (2014) use Hosman *et al.* (2010) bounds and Oster's (2013) extension of Altonji *et al.* (2005).

Using only the most recent pre-treatment period throws away information from prior periods. Using all available periods gives equal weight to each, yet one might care more about balance in year -1 than in year -5. One could give greater weight to more recent years, but the weights would be ad-hoc. Given a reasonably long pre-period, one could balance on (level in year -1) and (trend over full pre-treatment period).

4.4 DiD with Effects that Appear over Time

Some shocks produce an immediate effect on the outcome; some may change post-treatment trends but not levels; some may affect both levels and trends; for some, an impact on outcomes may emerge over time in a pattern that cannot be neatly captured as a change in level, a change in trend, or both. To allow for a change in both level and trend, start with the panel DiD eqn. (10) and add a post-reform trend:[61]

$$y_{it} = \alpha + f_i + g_t + \left(\delta_{\text{DiD,level}} * w_{it}\right) + \left(\delta_{\text{DiD,trend}} * w_{it} * t\right) + \varepsilon_{it} \tag{25}$$

Often, treatment effects will emerge over time, but cannot be captured through an immediate and permanent change in level; an immediate and permanent change in trend, or both. One example is gradual phase-in, which might take a rough S-curve shape. A "distributed lag" model, which lets the treatment effect vary with the time since reform can allow for phase-in, and does not require the researcher to impose a time structure on the treatment effect. The distributed lag model is similar to the leads and lags model in eqn. (21), except instead of year-specific treatment dummies w_i^k, which turn on in a single period k and then off, one uses a set of lagged treatment variables w_s^{k-lag}. The first lag w_i^{1-lag} turns on for treated firms in the first post-reform period and stays on thereafter; the second lag w_i^{2-lag} turns on in the second post-reform period and stays on, and so on. The model is:

$$y_{it} = \alpha + f_i + g_t + \sum_{k=1}^{n} \left(\delta_{DiD}^k * w_i^{k-lag}\right) + \varepsilon_{it} \tag{26}$$

The coefficient on w_i^{1-lag} estimates the treatment effect in the first post-treatment year; the coefficient on w_i^{2-lag} estimates the *additional* effect in year two, and so on. One can stop the series once the treatment effect is expected from theory, or observed in the data, to be nearly complete. The last lag uses fewer years of data and thus tends to be noisier; thus, it can be useful to let the last lag cover two or three periods.

[61] The change in trend is an interaction between the level and a time trend. To interpret the coefficient on an interacted variable, one must generally include its non-interacted components. In eqn. (25), the time trend is absorbed by the year dummies.

One can sum the lagged effects to obtain an overall treatment effect and accompanying t-statistic.[62] One can also assess whether *when* a treatment effect appears is consistent with theory – if the impact is expected to be gradual (or sudden), is it? No paper in our sample uses a distributed lag model; this is a missed opportunity.[63]

4.5 DiD-Continuous Designs

In many finance contexts, all firms are subject to a shock but have different sensitivity to the shock based on their (possibly endogenous) background characteristics. One can then seek to assess whether the shock affects high-sensitivity firms differently than low-sensitivity firms. An example from our own research (Atanasov *et al.*, 2010): In 2002, Bulgarian legal reforms limit "equity tunneling" – dilutive share offerings and freezeouts of minority shareholders. We use pre-reform data to estimate each firm's propensity for equity tunneling, and study whether firms with high tunneling propensity react more strongly to the shock than low propensity firms.[64] Let $propens_i$ be equity tunneling propensity. The DiD-continuous estimation equation is:

$$\{\text{panel DiD-continuous}\}:\ y_{it} = \alpha + f_i + g_t + \left(\delta_{DiD} * post_{it} * propens_i\right) + \varepsilon_{it} \quad (27)$$

By comparison with the panel DiD equation (10), the term in w_{it} is replaced by an interaction of a post-reform dummy with sensitivity-to-shock.

DiD-continuous designs have not, to our knowledge, been studied in the methods literature.[65] Yet they can be credible, given a good shock and a reasonable way of estimating sensitivity to the shock. One must impose a parametric form on the sensitivity, and posit a particular channel for the effect of the shock on the outcome. This is an only through assumption (see §4.2.6), which must be defended. For example, in Atanasov *et al.* (2010), the reform shock should affect the outcome only through its impact on tunneling, not through some other difference between

---

[62] In Stata, one runs the regression in eqn. (26), followed by the lincom (linear combination) command: lincom $w^{1-lag} + w^{2-lag} + w^{3-lag} + w^{4-lag} + \ldots + w^{n-lag}$, where n is the last lag. The sum of coefficients will be similar to the point estimate in period n-lag from a leads and lags model; differences will be due to the reference period (a particular pre-treatment period for the leads and lags model; an average over the pre-treatment period for the distributed lag model) and whether the last lag in the distributed lag model covers more than one period.

[63] The closest is Bertrand and Mullainathan (2003), who estimate a related mixed model: they examine a regression with on-and-off treatment dummies in years t-1 and t (to check for pre-treatment trends), year t+1, and years (t+2 and after); Qiu and Yu (2009) similarly estimate separate effects in year t, t+1, and (t+2 and after).

[64] Atanasov *et al.* (2010) estimate two tunneling propensities, one for share dilution and one for equity freezeouts. For simplicity, the example in text uses a single propensity.

[65] A few methods papers generalize the propensity score to allow for continuous treatment levels in pure observational studies, but do not discuss DiD-continuous designs (e.g., Imai and Van Dyk, 2004; Hirano and Imbens, 2004).

high- and low-sensitivity firms. Only two DiD-continuous papers in our sample recognize and defend this assumption.[66]

A common variant on a DiD-continuous design is to divide the sample into high-sensitivity and low-sensitivity subsamples, and run classic DiD, using the low-sensitivity firms as a control group. Of the DiD-continuous papers in our sample, five use bins (sometimes two, sometimes more) for the sensitivity variable, either as the main specification or as a robustness check. One can also drop firms with middle sensitivities and compares, say, top third (or fourth) to bottom third (fourth). One paper in our sample drops middle-sensitivity firms.

4.6 Enhanced DiD Designs

We discuss here several ways that DiD designs can be strengthened, in appropriate situations.

4.6.1 Triple Differences

Some DiD designs are vulnerable because an identifiable difference exists between treated and controls. One can sometimes use a third difference, and thus a "triple difference" (DiDiD) design to address this difference. Only 2 papers in our sample use DiDiD; but event studies are a form of DiD, and our two good practice event studies both use DiDiD designs. Below, we discuss one of them as a motivating example and then present the regression algebra.

Litvak (2007) studies the impact of the adoption of SOX on foreign firms cross-listed in the U.S. on cross-listing levels 2 and 3, which were made subject to most SOX rules. She finds that SOX adoption events predict lower share prices of cross-listed firms, relative to matched non-cross-listed firms from the same country.[67] A natural objection is that the cross-listed firms may be exposed to U.S. markets in ways other than SOX, which would violate the only through assumption that SOX compliance is the only reason for the observed share price drops. Litvak addresses this concern by observing that foreign firms, cross-listed on levels 1 and 4, are not subject to SOX, but are otherwise likely to be exposed to U.S. markets in ways similar to the level 2-3 firms. The three differences in the DiDiD design are: (i) after minus before a SOX adoption event; (ii) cross-listed firm versus matched non-cross-listed firm; and (iii) level 2-3 matched pair (level 2-3 cross-listed firm versus its match) versus level 1-4 matched pair. The overall return to level 2-3 pairs over the principal adoption events is -11%, but the return to level 1-4 pairs is -5%. The difference provides a -6% DiDiD estimate for the impact of SOX adoption on the share prices of level 2-3 firms.

---

[66] Atanasov *et al.* (2010); Gomes *et al.* (2007).

[67] Litvak (2007, 2008) is a followup DiDiD study using the same third difference. It is outside our sample because it was published in *European Financial Management*, which is not a journal we survey.

As this example suggests, a good third difference is *not* just a second control group which may have different unobserved differences from the treatment group. Such a control group is useful for a placebo check (§4.2.12). Instead, the third difference should respond to a discrete known defect with the base DiD comparison.

A DiDiD design requires additional notation. With panel data, one treatment group, and two control groups, let $c_{it} = 1$ for the first control group in the past-shock period, and replace eqn. (10) with:

$$y_{it} = \alpha + f_i + g_t + \left(\delta^t_{DiD} * w_{it}\right) + \left(\delta^c_{DiD} * c_{it}\right) + \left(\delta_{DiDiD} * w_{it} * c_{it}\right) + \varepsilon_{it} \qquad (28)$$

The coefficient of interest is on the triple interaction term δ_{DiDiD}.

4.6.2 Testing Parallel Trends for DiDiD Designs

In an appropriate case, DiDiD can strengthen a DiD design, but the DiD design will usually remain primary. A principal reason is that DiDiD makes stronger parallel trends assumptions. Thus, DiDiD-only results, which are not present in DiD are suspect.

Consider Litvak's (2007) design. She compares level 2-3 firms (exposed to SOX) to matched local firms; level 1-4 firms to matched local firms, and then level 2-3 pairs to level 1-4 pairs. Each of these comparisons comes with a parallel trends assumption, sometimes more than one. For the first DiD comparison of (level 2-3 firm to local match), one assumes that the extra return to the level 2-3 firm is due to U.S. market exposure during non-event days and (U.S. market exposure plus SOX) during the event days. If U.S. market exposure doesn't affect returns, she doesn't need the third difference. If U.S. exposure does affect returns, one cannot test for parallel trends, because the effect is present at all times. For the second DiD comparison of (level 1-4 firm to local match), one assumes that the extra return to the level 1-4 firm is due to U.S. market exposure – again not testable. For the triple difference, one assumes that the extra return during the event days *to level 2-3 pairs*, versus *level 1-4 pairs*, is due to SOX. This is testable during non-event days.

The DiDiD design replaces a single parallel trends assumption for DiD with multiple assumptions. This makes the DiDiD design inherently more fragile than DiD, even if all of the parallel trends assumptions are testable. In theory, both double-difference parallel trends assumptions could fail but offset each other, leaving a valid triple difference design. In practice, if the testable double-difference trends are non-parallel, we should worry greatly about triple-difference validity. Thus, all relevant parallel trends should be tested (if testable). No study in our sample, either DiDiD or the event study analogue, does so. Figure 2 provides an example, from a study by one of us, of how badly non-parallel the third difference can be.[68]

An alternative to DiDiD, which will sometimes be available, is to limit the sample in a way that makes the third difference unnecessary. The advice might be: Good covariate balance trumps a third difference.

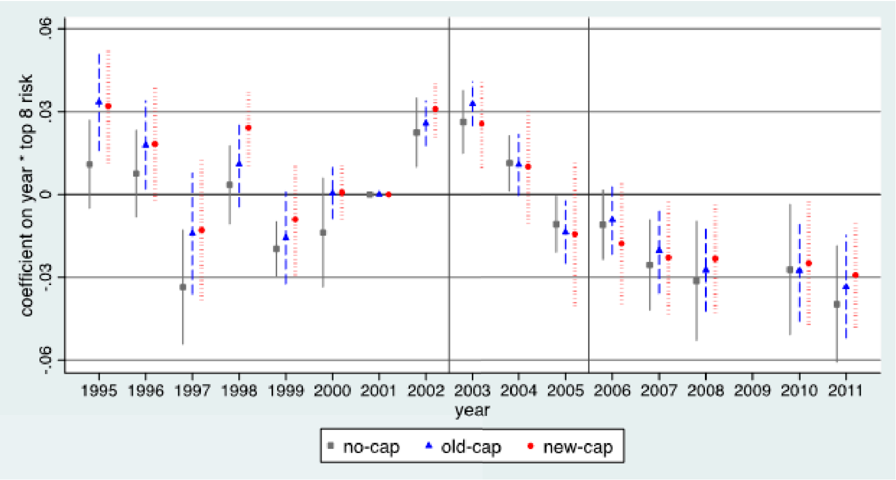

Figure 2: Non-Parallel Third Difference Example (from Paik *et al.* (2016))

Description: Figure is based on regressions of *ln* (physicians/100k population) on interactions of year dummies with high-risk specialty dummy, with state*specialty and year FE, and constant term, for 8 high-risk specialties versus 7 low-risk specialties, for 9 treated (new-cap) states, 20 no-cap states (control group 1), and 22 old-cap states (control group 2) over 1995-2011. Vertical lines indicate cap adoption period. Vertical bars show 95% confidence intervals for each point estimate.

4.6.3 DiDiD-Continuous and -Double Continuous Designs

A design that is related to both DiD-continuous and DiDiD designs can be called DiDiD-continuous, where the third difference involves sensitivity to the shock. One runs conventional DiD, then identifies firms that are more or less affected by the shock, and assesses whether the treatment effect is larger for more-affected firms. For example, Qiu and Yu (2009) study the impact of antitakeover laws on the cost of debt; and Giroud and Mueller (2010) do the same for profitability. Both sets of authors posit, and confirm, that the impact should be stronger in less-competitive industries, measured using the Herfindahl-Hirschman Index.

[68] Paik *et al.* (2016). Two prior papers used a DiDiD design in this setting, saying the third difference was needed to control for possible non-parallel double-difference trends, but did not check whether the third difference was indeed parallel during the pre-shock period. The cure was worse than the (possible) disease.

Let $sens_{it}$ be the continuous sensitivity variable. The DiDiD-continuous design, with panel data, is similar to DiDiD, except that the sensitivity replaces the third difference:

$$y_{it} = \alpha + f_i + g_t + \left(\delta^t_{DiD} * w_{it}\right) + (\gamma * sens_{it}) + (\lambda * w_{it} * sens_{it}) + \varepsilon_{it} \qquad (29)$$

As with DiD-continuous designs, one can also divide the sensitivity variable into bins. If the sensitivity variable has only two levels, eqn. (20) is formally identical to the DiDiD eqn. (19). The difference is in interpretation. In DiDiD, the coefficient on the triple interaction is the core treatment effect of interest. In a DiD-plus-sensitivity design, one is interested first in the DiD coefficient, and secondarily, as further analysis or a robustness check, in how the treatment effect varies with the sensitivity variable.

If the main DiD design is DiD-continuous, then adding a sensitivity-to-shock variable, as in eqn. (20), leads to what we can call a DiDiD-double continuous design. As with other DiD-continuous and DiDiD-continuous designs, one can divide the continuous variables into bins. Finally, Low (2009), has a "pure DiDiD" design and also examines whether the DiDiD effect varies with a firm characteristic. In our taxonomy, one would term this DiDiDiD-continuous.

4.6.4 Use of DiDiD and its Related Continuous Designs to Study Interactions

In addition to addressing concerns that the treated and control groups in a DiD design differ along a third dimension, DiDiD and DiDiD-continuous designs can be used to examine the causal effect of an interaction between two variables. We don't have DiD examples in our sample, but several shock-based IV papers that adopt a similar approach. For example, Duchin *et al.* (2010) study whether information costs modify the causal effect of board independence on firm performance, and Desai and Dharmapala (2009) test whether good governance increases the valuation effect of tax shields. As we discuss in Atanasov and Black (2016b), their IV designs can be recast as DiDiD-continuous or -double continuous designs.

4.6.5 Similar Shocks at Different Times

A core threat to DiD validity is non-parallel trends. This threat becomes less likely if one can study a number of similar shocks, at different times in different places, for which the timing appears to be random. A good example is Dinc (2005), discussed below as a good-practice paper. He studies lending by government-controlled banks, and finds that their lending rises in election years, relative to other banks from the same country, and falls the year after the election. Across different countries, which hold national elections at different times, often fixed times unrelated to economic cycles, it is hard to see how non-parallel trends can explain his results.

Some multiple shock designs remain vulnerable to the risk of non-parallel trends. Consider studies of state antitakeover laws. These are often adopted at similar times, in response to takeover waves, which in turn correlate strongly with business cycles. Thus, a non-parallel trends story is natural, even though one has multiple legal shocks.

4.6.6 Reversals

An important special case of repeated shocks involves legislative reversals. If a legislature reverses course, one gets two events, with opposite predicted signs. If both events have the predicted sign, competing stories are weaker. When the opposing events are close together in time, event study methods are often used. But some reversals are slower and lend themselves to DiD analysis. Dinc's study of lending by government-controlled banks in election and non-election years again provides an example. A second example: In 1997, Brazil weakened the "takeout rights" provided to minority shareholders on a change in control. In 2001, it strengthened them again. Nenova (2005) and Carvalhal da Silva and Subrahmanyam (2007) exploit this rule adoption and reversal (without using DiD terminology).[69]

4.7 Combining DiD with Balancing Methods

4.7.1 Shock-Based DiD Plus Balancing

Researchers rarely combine DiD (or, for that matter, event studies or IV) with balancing methods. Only three DiD papers in our sample match treated to control firms, and all use only crude matching methods. Yet a combined DiD/balancing design is often feasible and, we believe, would often enhance DiD credibility.[70] An example can illustrate how matching can matter. Both Litvak (2008) and Doidge *et al.* (2009), study the impact of adoption of the Sarbanes-Oxley Act on foreign firms cross-listed in the U.S. Litvak uses crude matching – she matches each cross-listed firm to a non-cross-listed firm from the same home country and industry, and similar on size. This should produce reasonable covariate balance between her treatment and control groups (balance is not assessed in her paper). In contrast, the Doidge, Karolyi and Stulz control group is all non-U.S. firms with financial data on Worldscope. They use regression, with only a few covariates, to address the differences between cross-listed and other groups. The regression

[69] If a legal minimum is relaxed (the opposite of the Brazil pattern), relaxation does not mean that firms will abandon the practice. See Hope and Thomas (2008) (studying relaxation of accounting rules on segment reporting). One can analyze the relaxation as an "intent to treat" design, with the rule change as an instrument for actual change. We discuss these designs in Part 6.

[70] Compare Smith and Todd (2005), who report that a DiD-plus-matching estimator out-performs pure balancing methods, when applied to the National Supported Work dataset. To our knowledge, the first paper to combine DiD with balancing is Heckman *et al.* (1997).

coefficients will be determined mostly by the control firms, which numerically dominate the sample, and might not fit well the cross-listed firms, which are on average much larger.[71]

With one pre-treatment period, one can use standard balancing approaches. But suppose one has data for multiple pre-treatment periods. How can one use this data to improve balance, and reduce the core threat to DiD validity posed by non-parallel trends? There is almost no guidance in the balancing literature, but one approach that makes sense to us would be to balance on both the level of covariates in the last pre-treatment period, plus the trend in covariates during the full period.

4.7.2 Non-Shock DiD Plus Balancing

A number of DiD studies in our dataset do not begin with an exogenous shock. For example, Cheng and Farber (2008) compare the change in CEO option grants for firms that restate earnings with the change for non-restating control firms; Hail and Leuz (2009) compare the change in cost of capital for firms which cross-list in the US to the change in the median cost of capital for all non-cross-listed firms from the same country. Using balancing methods to make control firms more closely similar to treated firms is often valuable for shock-based DiD, but is even more important for these and other non-shock DiD papers.

4.7.3 Synthetic Controls

Studies at the state or national level sometimes have only a few (in the extreme, one) usable shocks. This increases the risk that a post-shock divergence between treated and control units is due to unobserved characteristics of the treated units, rather than the shock. In this situation, the synthetic control method of Abadie *et al.* (2010) can be useful.[72] We discuss this method only briefly, because to date it has been rarely used in finance and accounting research (Berger *et al.* (2015), is a recent exception).

One begins with a "donor pool" of control units, and uses pre-treatment data to construct a synthetic control for each treated unit that best matches the treated unit on the outcome variable during the pre-shock period, and ideally on covariates as well. The synthetic control is a linear combination of donor states, with positive weights that sum to 1. One needs a long pre-treatment period for the synthetic control to be reliable. The synthetic control approach lends itself to graphical analysis. It does not produce standard errors. One should always report the

[71] These two papers largely address different questions; we comment here only on how they form a control group, not on the overall credibility of their results. Conflict disclosure: Litvak is married to one of us (Black).

[72] See also Abadie and Gardeazabal (2003); Abadie *et al.* (2015).

weights on each donor state, and assess whether the donor units with high weights are sensible matches.

In our experience, in non-finance applications with states that undergo legal reform as treated units, and non-reformed states as donors, the donor weights are sometimes sensible but sometimes not.[73] We are agnostic on when synthetic controls are a significant advance over classic DiD, once one has, say, five or more treated units. But the method offers, at a minimum, a useful robustness check.

4.8 Good Practice DiD Papers

We discuss here two "good practice" DiD papers that illustrate the approaches discussed above. We chose these papers in part for their rich variety of approaches and robustness checks, but provide only an overview of main lines. For each, we also discuss what else one might do, data permitting.

Bertrand and Mullainathan (2003) examine whether state antitakeover "business combination" laws affect plant-level profitability and productivity. The treated group is plants owned by firms incorporated in states that adopt these laws; the control group is plants in the same state owned by firms incorporated in other states. Law adoptions are staggered through time, so they have multiple treatment events.

The authors are aware that states can pass a number of antitakeover laws, at the same or similar times. They argue that business combination laws are the important ones. In effect, they argue for a particular only-through condition – that adoption of a business combination law predicts their outcomes, but adoption of other antitakeover laws, at around the same times, do not. They address the concern that the state's past economic performance generates political pressure to adopt antitakeover laws. First, they argue that these laws were often pushed by one or a few firms in each state, but affected all firms incorporated in that state. Second, they study outcomes at the plant rather than the firm level – they compare plants in the same state owned by firms incorporated in different states, and thus affected by different antitakeover laws. Third, they look back in time to $t = -1$ [with the reform in year 0] and find no differences in outcomes between treated and controls over $t = (-1, 0)$. They also look for a delayed impact of reform, and find an increase over both (0,+1) and (+1, +2).

The authors examine covariate balance between treated and control plants, and find that treated firms are larger and own larger plants. They address imbalance on firm and plant size in two ways. First, they allow plants of different size to have different time trends. Second, they confirm robustness if they limit the control group to states which later pass antitakeover laws; this improves covariate balance.

[73] See, e.g., Paik *et al.* (2017).

What else might the authors have done? Current best practice, data permitting, would include fuller analysis of the evidence for parallel pre-treatment trends, and would likely include a longer post-treatment period, testing placebo shocks (at different times) and placebo outcomes.

The authors might also have run a horse-race between different antitakeover laws, and between these laws and firms' ability to adopt poison pills.[74] Their failure to do so looms large given the criticism of studies of antitakeover laws by legal scholars, notably Coates (2000) and Catan and Kahan (2016). These scholars argue that business combination laws were a sideshow and the main event was creation and judicial or legislative validation of the poison pill defense around the same time, mostly during the late 1980s.

Our second good practice DiD paper is Dinc (2005) who studies the effect of government ownership on bank lending – one of only three non-legal shock DiD papers in our sample. Dinc uses national election years as exogenous shocks to lending by state-controlled banks; privately owned banks in the same country are the control group. Lending by state-controlled banks rises in election years in emerging markets, but not in developed markets. Like Bertrand and Mullainathan, Dinc has multiple shocks, because different countries have elections at different times. This reduces concern that the results reflect elections coinciding with macroeconomic shocks. Dinc also controls for macroeconomic trends by interacting election years with macroeconomic indicators.

Dinc assesses covariate balance and finds significant differences between state-controlled and private banks in size, average lending behavior, profitability, and leverage; his design assumes that a combination of bank fixed effects and time-varying covariates will address this imbalance. He uses alternate shock years – the years before and after a national election – and reports no significant differences in lending behavior during pre-election years and a reversal of the election-year rise in lending in the year after the election. He surmises that political influence on state-controlled banks is stronger in less developed countries, splits the sample into emerging and developed markets, and finds that state control predicts lending increases in election years only in emerging markets.

What else might Dinc have done? We are not pursuaded by his analysis of lending in the year before and after-elections. State-controlled banks lend less than their private counterparts in both the year before and the year after elections. As we read his tables, the combined effect is likely significant, and total lending growth for state-controlled banks over $(-1,+1)$ relative to election year 0 is close to zero. A leads-and-lags graph would show this. One could use balancing methods to provide better balance between state-controlled and private banks. He also does not distinguish between elections called by the government, and elections on

[74]They could have included separate dummy variables for each type of law, and assess whether their favored law has predictive power, while others do not. Compare Black and Kim (2012), who run a "horse race" between their large-firm dummy and a *chaebol* dummy, which is strongly correlated with the large-firm dummy.

a fixed schedule; the case for confounding by macroeconomic factors is weaker for the latter.

5 Shock-Based Event Studies

Event studies have a long history in finance, starting with Fama *et al.* (1969). For reviews, see MacKinlay (1997); Bhagat and Romano (2002a, 2002b). Event studies can be used both to study firm-specific events and – of principal interest here – the effect of exogenous shocks on the value of publicly traded securities. We assume the reader is familiar with the event study design and discuss here only the insight into event study design that flows from viewing an event study as a type of DiD design. We discuss event studies of share prices, using the simple "market model" in eqn. (13). Similar comments would apply to studies that use a 3- or 4-factor model, or study other securities.

5.1 Event Studies as DiD

Event studies can be seen as a type of DiD. The outcome variable is share returns; the treatment group is the firms one wants to study, and the control group is the other firms included in the market index.

To illustrate the link between DiD and event studies, we first strip the event study to its essentials. Convert the simple DiD case with one pre-treatment and one post-treatment period, from equation (9), to first-difference form:

$$\{\text{first-difference DiD}\} : \delta y_i = \beta + (\delta_{DiD} * w_i) + \epsilon_i \tag{30}$$

Now consider a simple event study, with: (i) one event date; (ii) a common intercept for all treated firms, instead of firm-specific intercepts; (iii) using market-adjusted returns (MARs), with no adjustment for the β's of the treated firms; and (iii) an equally-weighted market index, from which (iv) the treated firms are excluded. Index treated and control firms by i and "stack" them so the $\mathrm{n_t}$ treated firms have values $i \in [1, \mathrm{n_t}]$ and the control firms have values $\mathrm{i} \in [\mathrm{n_{t+1}}, \mathrm{n}]$. For the treated firms, the MAR model is:

$$\{\text{for treated firms}\}\ r_i = \alpha + r_m + \epsilon_i$$

For the control firms, we can similarly write:

$$\{\text{for control firms}\}\ r_i = \alpha + r_m + \epsilon_i$$

Now put the two groups together, letting $\mathrm{w_i}$ be a treatment dummy and defining a revised constant term $\alpha' = (\alpha + \mathrm{r_m})$

$$\{\text{for all firms}\}\ r_i = \alpha' + \gamma * w_i + \epsilon_i \tag{31}$$

This form is identical to eqn. (30). We can recognize α' as the average event-period return to all firms (the market return) and γ as the average *extra* event-period return to treated firms (the ATT).

It is easy to generalize eqn. (31) to allow for a multi-day event window, and for stock price data that includes the period before and after the event window. For a k-day event window, one can add a time subscript, and let w_{it} =1 during the event window and zero otherwise.

$$r_{it} = \alpha_t' + \gamma * w_{it} + \epsilon_{it} \tag{32}$$

A number of event studies in our sample use the "regression" form in eqn. (32), either instead or in addition to the "classic" form in eqn. (13). Papers that use both approaches include Litvak (2007) (discussed below as a good practice paper) and Black and Khanna (2007).

Event-studies, as a form of DiD, face many of the same concerns as other DiD studies, and should often adopt the DiD design strategies discussed in Section 4. From that perspective, a number of standard event study practices – both what researchers do, and what they don't consider doing – appear peculiar. We offer specific examples below.

5.1.1 Shock-Based Event Studies

Our sample includes 185 event study papers. Most involve takeovers or firm-initiated changes in corporate governance (e.g., adding outside directors; adopting anti-takeover provisions). Some papers study the effects of actions by outside investors (e.g., hedge funds, sovereign wealth funds). These papers as not shock-based, because these actions are not exogenous to firm characteristics. Of the event studies, 35 are based on an exogenous shock; of these 27 rely on legal shocks.

Table 10 shows the distribution over time of the 35 shock-based event study papers. The spikes in 2007 (8 papers) and 2010 (9 papers) are driven by SOX studies. Table 11 provides details on these papers. An event study is the main method in 27 of the papers. Only seven papers have a clean control group; the remaining 28 rely on differences in firm sensitivity to the shock. This design is similar to a DiD-continuous design; we will call it "event study continuous."

Among the eight papers that use non-legal shocks, three rely on sudden death (Faccio and Parsley, 2009; Nguyen and Nielsen, 2010; Salas, 2010); two use financial crises (Baek *et al.*, 2004; Lemmon and Lins, 2003), one uses the outcome of a close presidential election (Goldman *et al.*, 2009), one uses class-action lawsuits as a shock to non-sued firms that have interlocking directors with the sued firm (Fich and Shivdasani, 2007), and one uses actions by an Korean activist corporate governance fund as a shock for non-targeted companies with similar governance (Lee and Park, 2009).

Year	Event Studies	Shock-based	Legal shock	Event study continuous	Citations to Shock-based Event-Study papers
2001	3	0	0	0	
2002	7	0	0	0	
2003	4	2	1	2	Lo (2003)*; Lemmon and Lins (2003)*
2004	8	1	0	1	Baek *et al.* (2004)*
2005	19	0	0	0	
2006	19	3	3	1	**Greenstone *et al.* (2006);** Akhigbe and Martin (2005)*; Chang *et al.* (2006)
2007	25	9	8	7	**Litvak (2007)** Black and Khanna (2007)*; Brown *et al.* (2007)*; Carvalhal da Silva and Subrahmanyam (2007); Chhaochharia and Grinstein (2007)*; Fich and Shivdasani (2007)*; Engel *et al.* (2007)*; Zhang (2007)*; Wintoki (2007)*
2008	23	2	2	2	Ghosh *et al.* (2008)*; Li *et al.* (2008)*
2009	29	5	2	5	Chen *et al.* (2009)*; Faccio and Parsley (2009)*; Goldman *et al.* (2009)*; Hochberg *et al.* (2009)*; Lee and Suh Park (2009)*
2010	33	11	9	8	Akhigbe *et al.* (2010)*; Bae and Goyal (2010)*; Berkman *et al.* (2010); Brochet (2010)*; Doidge *et al.* (2010)*; Fernandes *et al.* (2010)*; Francis *et al.* (2010); Giroud and Mueller (2010)*; Iliev (2010); Nguyen and Nielsen (2010); Salas (2010)*
2011	15	2	2	2	Huang *et al.* (2011)*; Larcker *et al.* (2011)*
Total	**185**	**35**	**27**	**28**	

Table 10: Event Study Designs over Time

Description: Summary of papers in our sample that use event study designs. Shock-based designs include "event study – continuous" papers where all firms are exposed to the shock, but have differing sensitivities to the shock. * indicates event study continuous design. **Boldface** indicates our "good practice" papers.

Interpretation: Table shows that use of shocks as a basis for event studies increases over our sample period, and that event study-continuous designs are a common way to exploit shocks. Most shock-based event studies are based on legal shocks.

	Total Papers
Event study is main research design	27
Control group	
True control group (not subject to shock)	7
True DiDiD design	2
DiDiD-continuous (DiD plus sensitivity to shock)	2
DiDiD-double continuous	1
Event-study continuous	28
Limit control group to similar firms	4
Assess covariate balance	3
Use placebo shock on different date	3
Common event date	31
Form portfolios	20
Cluster errors on event date (multiple event dates)	1
No method to address common date or invalid method used	11
Multiple dates	21
Reversal	8

Table 11: Details on Shock-Based Event Study Papers

Description: Table summarizes selected aspects of research design for the 35 papers in our sample that use a shock-based event study.

Interpretation: Table shows which aspects of DiD design are commonly or rarely used in shock-based event studies. Most aspects of DiD design are rarely used. Multiple event dates and reversals are often used.

5.2 Elements of Event Study Design

5.2.1 Choice of Control Group

In DiD analysis, a core concern is similarity between the treatment and control groups. The more similar the groups, the more plausible the core assumption that the two groups would have followed parallel paths during the period of study, but for the shock one is studying. Good design includes a careful check for common support and other aspects of covariate balance.

Event studies, in contrast, often use a broad market index as the control group. A typical study uses a simple parametric control (each firm's β relative to the market index) to address differences between the treated and control firms.[75] For example, a typical market index includes firms from a wide variety of industries, and a wide variety of β's and other firm characteristics that can predict returns.

Once one views event studies as a form of DiD, using a broad market index as the control group seems odd. In a classic DiD study, one would never construct a DiD control group comprised of "every control unit I could find." Yet this is what

[75] Most event studies do not remove the treated firms from the market index. One should, at least if the treated firms have enough weight in the market index to meaningfully affect the results. We put aside that detail here.

many event studies do. There is extra work in constructing a custom index, and defending the choice of firms that are included in the index. But the payoff is greater credibility of results.

Only four event-study papers in our sample use matching to improve balance between treated and control firms; these are also the only papers that assess covariate balance. We offer two examples of how DiD ideas can inform the choice of control firms. Within our sample, Litvak (2007), discussed below as a good practice paper, studies the impact of SOX on foreign cross-listed firms; she matches cross-listed firms to similar non-cross-listed firms. Outside our sample, Cremers and Ferrell (2014) study the *Moran v. Household International* decision in 1985, which permitted the "poison pill" takeover defense, by comparing returns to Delaware firms, which were affected by the decision, to an index of firms with dual-class common shares, which should not be affected, because they are not vulnerable to hostile takeovers.

5.2.2 Parallel Trends

Confirming parallel pre-treatment trends will often be less of a concern for event studies. In a semi-strong efficient market, one would normally not expect past trends, even if they exist, to continue into the future. But confirming parallel pre-treatment trends can still be important, especially for a long event window. Evidence that returns to treated and control firms are similar outside the event window will increase comfort that divergence during the event window is due to the event being studied, rather than some other factor. Care in building a control group makes it more likely that pre-treatment trends will be parallel.

No event study paper in our sample directly tests whether the treated and control groups have similar returns before the shock period. Black and Kim (2012) (not in our sample) provide a graphical example of such a check. Greenstone *et al.* (2006), discussed below as a good practice paper, confirm that treated and control firms have similar returns *after* the event period. Two other papers confirm that there are no abnormal returns at a different (placebo) event date or dates.

5.2.3 The Only Through Condition

An event study is also related to IV. The "event" is a shock to investor information. That shock can affect outcomes though the underlying substantive event (a governance reform, say), and can be understood as an instrument for that event. But new information can also affect share price through other channels, as in the example in Section 2 of adoption of a takeover defense. For any event study that relies on voluntary firm actions, the "revised expectations" channel will be hard – sometimes impossible – to exclude. An external shock that is outside the control of the treated firms helps, but may not offer a complete solution, if some other difference between treated and control firms could explain the observed results. Similar to IV and DiD, often the best one can do is to identify plausible alternative stories, and look for ways to rule them out.

Consider, for example, the 1999 Korean reforms to corporate governance of large firms, studied by Black *et al.* (2006) and Black and Kim (2012). These reforms may directly improve firm value, but could also signal that the government will tighten its oversight of large firms in other ways. Or, since most large Korean firms belong to *chaebol* groups, the reforms could signal increased oversight of these groups. The first story is less plausible if smaller firms, which voluntarily adopt similar reforms, experience similar price increases; the second can be tested by running a "horse-race" between a large firm dummy and a *chaebol* dummy.

5.2.4 Cross-Sectional Correlation

In an event study which relies on a legal shock, most treated firms will have a common event period, which could lead to cross-sectional correlation in returns. Of the 32 papers with this issue, 13 deal with it, usually by running an event studies on portfolios of treated firms. An alternate approach is to run regressions with returns as the dependent variable, identify firm groups for which correlation is likely (by industry, say) and cluster on group (Black and Khanna, 2007). One can also use bootstrapped standard errors, generated by comparing treated to control firms outside the event period (Larcker *et al.*, 2011).

5.2.5 Delayed Reforms

Some legal reforms are applied to one group of firms earlier than another. The group with deferred compliance can then become a control group for firms subject to faster compliance. The estimated treatment effect will be biased downward, since investors will often expect that the control group will have to eventually comply with the reform, and some reforms explicitly provide for this. But the bias may be modest if investors expect that the deferral period might be extended, or even become permanent. If the threshold is "sharp" (based on firm size, say), a combined event study/RD design is likely to be available, and can help to ensure that the treated and control firms are similar.

Iliev (2010) study of SOX §404, discussed below as a good practice RD paper, provides an example. He conducts an RD study of the impact of SOX §404 on firms just above the \$75M public float threshold for initial compliance in 2004, relative to smaller firms who could delay compliance, plus an event study that compares firms just above the threshold to those just below the threshold, during the SOX adoption period in 2002. The temporary exemption for firms with public float below \$75M was eventually made permanent. Similarly, Black and Khanna (2007) conduct an event study of India's "Clause 49" reforms, which applied first to large

Indian firms, then mid-sized firms, then small firms. The small firms provide a control group for the large firms; the mid-sized firms are an intermediate group. Dharmapala and Khanna (2013) use an event study/RD design of the value of enforcement of Clause 49. They compare returns to "treated" small firms, just above the compliance threshold, to control firms just below the threshold, when the Indian securities commission launches its first enforcement action.

5.2.6 *Defining the Event Period(s)*

Defining the appropriate event period for a legal shock can pose challenges. Often, laws are adopted over a period of time, with a number of discrete legislative events of varying significance. Some events may predict higher likelihood of adoption, others may predict lower likelihood, still others will change the expected substance of the law. It can be valuable to conduct "short window" event studies of key events, and also to measure cumulative returns over the period from (first important event, last important event). Consistent results across multiple short windows can greatly boost credibility.

Larcker *et al.* (2011) illustrate how one can handle this complexity. They study two issues – say-on-pay and shareholder access to the company proxy statement for director nominees – and identify 18 events. Eight are legislative and increase the likelihood of say-on-pay regulation; three events also increase the likelihood of proxy access regulation. The remaining 10 events are from the SEC and concern proxy access; five increase and five decrease the likelihood of regulation.

It can be dangerous to use event window conventions, developed for corporate news, to study legislative events. For example, many takeover studies use an event period that starts before the announcement date, to capture news leakage. Gagnon and Karolyi (2012) apply this approach in assessing investor reaction to the U.S. Supreme Court decision in *Morrison v. Australia National Bank* (2010), where the Supreme Court ruled that persons who trade American Depository Receipts (ADRs) of firms cross-listed in the U.S. can sue for violation of U.S. securities laws, but persons who trade equivalent shares in the firm's home country cannot. If the right to sue is valuable, ADRs should rise in value, relative to home country shares, and Gagnon and Karolyi so find, using a (-1, +1) event window. Gagnon and Karolyi have a clean shock and an excellent control group. But their results are driven by returns on day -1. It is unheard of for news on the substance of a Supreme Court decision to leak. Thus, a better event period would be $(0, +1)$. This would produce a positive but insignificant return to ADRs.

5.2.7 *Reversals*

Some event studies benefit from unexpected legislative change of direction. These can greatly boost credibility. Litvak (2007) discussed below as a good practice paper, provides an example. She studies the effect of SOX on cross-listed firms, and studies both legislative and SEC events. After Congress adopted SOX, which applies to some cross-listed firms, the firms it applies to mounted a lobbying campaign for exemptions by the SEC. An initial SEC rule applied SOX §302 (CEO and CFO certification of financial statements) fully to cross-listed firms. Two months later, SEC Chairman Harvey Pitt gave a well-publicized speech in which he promised regulatory flexibility. He was soon fired for unrelated reasons, and the next SEC rule applied SOX §404 (certification of financial statements) fully to cross-listed firms. Litvak finds negative reactions to the first and third of these announcements, and a positive reaction to the second.[76]

5.2.8 Non-Exogenous "Shocks"

Can a legal shock be sufficiently exogenous to support credible causal estimates, even if the affected firms lobbied for the change? Perhaps. There are no examples in our dataset, but consider the Acemoglu *et al.* (2016) event study of gains to "connected" banks from the 2008 appointment of Timothy Geithner as U.S. Treasury in November 2008, during a financial crisis. The connected banks likely lobbied for this appointment. One might still believe that the after-minus-before change in share prices for less-connected, control banks is a good proxy for the unobserved change in treated banks, had they not been treated.[77]

For event studies without an exogenous shock, much like DiD without an exogenous shock, careful balancing of the treated and control groups can do much to enhance credibility.

5.2.9 Anticipation

Shock exogeneity includes firms' inability to anticipate the shock (or a future reversal). The analogue for event studies is that investors cannot anticipate the news event. This is often problematic for event studies, because in an efficient market, investors are doing their best to anticipate future news. For example,

[76] Another example with multiple reversals is Muravyev (2013). He studies the effect of Russian reforms which give preferred shares veto rights for charter amendments that reduce their rights. These are part of corporate law reforms that were adopted by the Russian Duma in 1999, unexpectedly rejected by the other legislative house, the Federation Council, adopted again by the Duma in 2000, again unexpectedly rejected by the Federation Council, and then finally adopted in 2001. Sometimes, reversal-based results may be credible even with only an interrupted time series design. An example is Mitchell and Netter (1989) who provide evidence that a proposed tax on corporate acquisitions contributed to the 1987 stock market crash: prices fall when the tax is proposed, and rebound when the proposal is dropped.

[77] One would also need to be convinced that the only through condition is satisfied. For Acemoglu *et al.* (2016), the Geithner appointment would need to affect connected banks only through his future actions, not because it signals the banks' lobbying prowess, or the U.S. government's likely response to the financial crisis.

a takeover bid sharply raises the probability of a takeover of the target, but the probability was often not zero beforehand.

5.3 Good Practice Event Study Papers

Our first good practice event study paper is Greenstone *et al.* (2006). The authors study a 1964 reform which increased disclosure requirements for US firms traded over-the-counter (OTC). The reforms affect four separate disclosure areas: registration statements for public offerings, ongoing financial disclosure once public, proxy statements, and trades by insiders. Different firms are affected differently by the shock. The authors construct two treatment groups: (1) "large" OTC firms with no "recent" public offering (in the last three years)[78] go from disclosure in none of these areas to disclosure in all four; (2) large OTC companies with public offerings in the last three years go from two disclosure areas to four. They have three control groups: (3) small OTC companies without recent public offerings (who go from 0 to 0 disclosures); (4) small OTC companies with recent public offerings (go from 2 to 2 disclosures); and (4) exchange-listed firms (already subject to all four disclosures).

The authors use these groups in a rich DiD-like event-study setup with three alternate control groups (0-0, 2-2, and 4-4) and sensitivity to shock as an additional comparison (0-4 vs. 2-4 treated firms). They classify firms into a 5×5 grid on market capitalization and book/market ratio, and confirm covariate balance on other variables within each cell of the grid. The authors find that the 0-4 treated firms have a larger announcement effect than the 0-2 firms. They apply placebo shocks in the 1965-1966 period, after firms are complying with the new disclosure rules, and find no differences in returns between the treated and control groups.[79]

Our second good practice event-study paper is Litvak (2007), who studies the effects of SOX on the market values of foreign companies cross-listed in the US. Foreign firms cross-listed with "level 2" or "level 3" ADRs became subject to SOX; firms cross-listed on level 1 or 4 did not. Litvak matches each cross-listed firm to a similar home country firm on industry and size. In an informal check for covariate balance, Litvak notes that matching produces reasonable balance on one governance metric – the S&P disclosure score.

In a DiD analysis, she measures the "pair return" around key SOX events as (return to level-2 or 3 cross-listed firm minus return to its match). But these pair returns could reflect the general exposure of cross-listed firms to U.S. securities markets, rather than the effect of SOX. Litvak addresses this potential violation of the only through condition through a triple difference design, in which she

[78] Large firms are firms with assets in 1962 > \$5 million or assets > \$1 million and 500 or more shareholders.

[79] But see Battalio *et al.* (2011) and Mulherin (2007) who question the validity of the long-event-window analysis of Greenstone *et al.* (2006).

compares pair returns for level 2-3 firms around SOX events to pair returns for level 1-4 firms.

Other attractive features of the research design include: assessing which legislative and regulatory events should predict pair returns for cross-listed firms (a different question than which events predict a reaction by US firms); studying multiple events, including reversals; and assessing whether home country governance, firm disclosure policy, and growth prospects predict sensitivity to the shock – in effect, a DiDiD-plus-sensitivity design. Some limitations: Litvak's matching is crude. It would also be useful to explicitly check covariate balance for a range of covariates.

6 Instrumental Variable Strategies

IV is a standard econometrics technique, often used in empirical finance research, but rarely with an exogenous shock as the basis for the IV. Of our 863 papers, 285 papers use either IV or a Heckman selection model (which is basically IV under another name); some use both.[80] Of these, only eight IV papers and no Heckman selection papers use shocks.[81] Table 12 shows the distribution of IV, Heckman selection, and shock-based IV papers over time, and lists the eight shock-based IV papers. Of note – the first shock-based IV papers appeared only in 2006.

It was not feasible to manually review all IV and Heckman papers and verify whether we missed any shock-based papers in our general search for shocks. We did review the 71 non-shock IV and Heckman papers published in our last two sample years (2010-2011) and found no misclassified shock papers. Of the 71 papers, 46 use IV but not Heckman, 7 use Heckman but not IV, and 18 use both. We classified the instruments, and assessed whether they were credible. As we discuss below, we did not judge any of the non-shock instruments in these papers to be credible. Earlier papers are even less likely to involve credible non-shock instruments.

The predominance of suspect, often poorly defended instruments is consistent with other studies. Larcker and Rusticus (2010) survey accounting papers over

[80]Heckman selection models require either an instrument for which firms are selected into the sample, which is used in the first-stage selection equation, but does not otherwise influence the outcome and thus can be omitted in the second stage; or else strong functional form assumptions about how the selection process occurs. The second approach has been all but abandoned, because the functional form assumptions are neither plausible nor verifiable in the sample, and results are sensitive to violation of the assumptions.

[81]Adams *et al.* (2005) illustrates the challenges in finding a truly exogenous shock. They use a shock-like instrument (founder death *before* the sample period, to instrument for whether the CEO is a founder). This instrument is creative but not exogenous – it is likely correlated with other firm characteristics, including firm age, CEO age, and CEO tenure. Founder or CEO death *during* the sample period, used as an IV, even sudden death, would raise similar concerns – the risk of death rises with CEO age. In contrast, sudden death can be a good basis for an event study, and is used in several ES papers in our sample (Faccio and Parsley, 2009; Nguyen and Nielsen, 2010; Salas, 2010).

	Number of papers			
Year	**IV**	**Heckman selection**	**Shock-based IV**	**Citation**
2001	8	1		
2002	7	2		
2003	15	0		
2004	12	4		
2005	17	6		
2006	26	7	1	Adams and Santos (2006)
2007	23	8	1	**Bennedsen *et al.* (2007)**
2008	28	13	1	Guner *et al.* (2008)
2009	39	6	2	Desai and Dharmapala (2009); Giannetti and Laeven (2009)
2010	42	11	2	Duchin *et al.* (2010); Iliev (2010)
2011	31	11	1	Dharmapala *et al.* (2011)
Total	**248**	**69**	**8**	

Table 12: Instrumental Variable and Heckman Selection Strategies over Time

Description: Summary of 285 papers using IV, Heckman selection or (for 32 papers) both methods, included in sample of 863 empirical corporate governance articles. List excludes one fuzzy RD paper (Black *et al*, 2006). The good practice IV paper is in **boldface**.

Interpretation: Table shows rarity of shock-based IV papers, including complete absence in first half of sample period.

1995-2005 and find 42 IV papers, but none with instruments that they consider credible; Lennox *et al.* (2012) study Heckman selection papers and find 75 papers published in top-5 accounting journals over 2000-2009, but none that they consider credible.

6.1 *Trends in the Use of IV and Heckman Selection*

We did not systematically code the non-shock IV and Heckman papers in our sample, but observed strong changes over time. Early in our sample period, IV analysis was rarely the principal method. Instead, it was often an afterthought, included in robustness checks, in "unreported results", sometimes in a footnote, sometimes without even specifying the instruments, often without reporting the first stage of *2SLS*. Careful discussion of instrument validity was rare. In the middle of our period, researchers begin to take endogeneity more seriously, but their instruments were rarely convincing or carefully defended.

In the last few years, the requirements for a valid instrument, including the need to satisfy exogeneity and the only through condition, are more often discussed seriously. The first shock-based instrument paper appeared only in 2006 (Adams and Santos, 2006). More researchers also acknowledged that "we have possible endogeneity but no good instruments"; where a few years earlier, a common approach was "better a bad instrument than no instrument." Yet unconvincing instruments still predominate. Occasionally, authors acknowledge that instrument validity is suspect, but reports results anyway – perhaps to satisfy a referee's request.[82] We discuss in the next section the principal non-shock instruments used in our last two sample years (2010-2011), and why they are problematic.

6.2 *Credibility of Non-Shock Based Instruments*

Table 13 summarizes the instruments used in the 71 non-shock IV and Heckman selection papers. The criteria for valid "regular" and Heckman instruments are similar, so we discuss both together. Most don't satisfy exogeneity; the ones that do don't satisfy the only through condition.

6.2.1 Types of Instruments

The most common category of instruments for *gov* (37 IV papers; 21 Heckman papers) are contemporaneous firm financial variables. These are not exogenous, and are unlikely to satisfy the only through condition. Lagged variables (independent, dependent, or both) are also popular instruments, often using the Arellano-Bond "system" or "difference GMM" approaches (21 IV papers; 5 Heckman papers). These are only slightly less implausible than contemporaneous firm variables. If the lagged variable is time-persistent (as most financial variables are), the lagged version is not reliably exogenous; if not, it won't predict the non-lagged value well enough to be usable. There is no obvious middle ground (Roberts and Whited, 2013).

Geographic averages are also often used (9 IV and 5 Heckman papers). Geography can sometimes generate a plausible instrument, but we found no convincing instances in our sample.[83] Exogeneity is often unclear, because firms choose where to locate, and because other firms in a region may be responding endogenously to the same forces as the "subject" firm. Moreover, location can predict outcomes through channels other than *gov*. Consider, for example, Hochberg and Lindsey (2010), who use what we saw as the *best* geography-based instrument in our sample. To instrument for stock option grants, they use grants by other firms in the same region but different industries. Using firms in other industries

[82] See, for example, Morck *et al.* (2011) ("If we use the popular instrumental variables legal origin, latitude, and major religion to estimate exogenous components of our bank control measures, and use these to re-estimate the tables in second stage regressions, we obtain qualitatively similar results. Although these instruments pass standard weak instruments tests, they plausibly affect economy outcomes through many channels, and therefore cannot be regarded as valid instruments").

[83] A classic example is Acemoglu *et al.* (2000), who use settler mortality as an IV for whether a country develops institutions conducive to local economic development. Glaeser *et al.* (2010) question the validity of this instrument.

	Regular IV	**Heckman**
Total Papers	64	25
Instrumented (or Selection) Variable(s) (one paper can use > 1)		
Governance	48	15
Firm outcome	15	
Other "independent" variable	21	10
Importance of strategy		
Main approach	19	1
Robustness check; reported	37	13
Robustness check, unreported	8	11
Single or multiple instruments		
One	11	3
Multiple (one instrumented variable)	49	22
Multiple (one for each instrumented variable)	4	
Type of instruments (one paper can use > 1)		
Contemporaneous firm variable	37	21
Lagged independent variable	11	4
Lagged dependent variable	3	1
lags of both dependent and independent vars. (Arellano-Bond)	8	
Geographic averages	8	5
Industry-level values	10	6
Country-level values	18	10
Other	2	
Nature of IV equation(s) (one paper can use > 1)		
Two-stage (2SLS or Heckman)	42	24
GMM models (Arellano-Bond and similar)	8	
3SLS	14	
Heckman functional form		1
IV Validity		
First-stage reported (for 2SLS or Heckman)	20	13
Test of IV strength (for 2SLS with multiple instruments)	13	
Pass/fail rule of thumb for instrument strength	8/5	
Test of overidentifying restrictions	20	

Table 13: Details on Non-shock IV and Heckman Selection Papers from 2010-2011

Description: The table summarizes 71 papers, published in 2010-2011, which use non-shock-based IV and/or the Heckman selection procedure. Of these, 46 papers use only IV, 7 use only Heckman selection, and 18 use both.

Interpretation: The table shows nature of research designs used in non-shock IV and Heckman selection papers.

strengthens the claim for exogeneity. Still, their instrument may not satisfy the only through condition, because geography can predict the outcome (firm performance) through channels other than stock options. Or consider John and Litov (2010), who find that firms with higher scores on the Gompers-Ishii-Metrick "G" index of takeover defenses have higher leverage. Their geography-based instrument for G is the average G score for other firms with headquarters in the same state. But the only through condition could easily fail. For example, causation could run from headquarters state → growth rate → [stronger takeover defenses and higher leverage].[84]

Industry averages are also popular (10 IV and 6 Heckman papers), but face similar problems with both exogeneity and only through. For exogeneity, many firms in an industry may respond endogenously to the same forces. For only through, industry-level factors can affect many firm characteristics, some unobserved. Country averages of the endogenous variable (7 IV papers, 3 Heckman papers) have a stronger claim to being exogenous to the firm, but the only through remains suspect, because country-level factors can affect both observed and unobserved firm characteristics. Other state- or country-level variables (18 IV papers, 10 Heckman papers), such as unemployment rates or GDP/capita, raise similar concerns.[85]

6.2.2 *Multiple Instruments*

Most of the non-shock IV papers (49/64) use multiple instruments (usually for a single instrumented variable). Another four papers use different instruments for different instrumented variables. In 20 of the multiple-instrument papers, the authors report that their instruments satisfied a Hansen test of overidentifying restrictions. This can test the validity of a second instrument, *assuming* the first one is valid and (less often recognized) homogeneous treatment effects. The Hansen test won't help if either condition is not met. It also won't help much if the valid IV is weak (see Atanasov and Black (2016b), for further discussion).

[84]Perhaps the authors were not fully convinced themselves; they report similar results using only their second instrument, the average G index for other firms with the same legal counsel. This instrument is also suspect. Legal counsel is a firm choice, hence not exogenous; and is influenced by location, hence may not satisfy "only through". We are more persuaded by their DiD analysis, which relies on state adoption of antitakeover laws.

[85]Only two non-shock instruments do not fall within the categories discussed in text. One is the "law firm" instrument used by John and Litov (2010), discussed above. The second is geographic distance between an outside director's home and a firm's headquarters, used by Fahlenbrach *et al.* (2010) to instrument for whether the outside director is a CEO of another company. Here, exogeneity requires that the firm does not choose directors taking into account travel difficulty; the only through condition requires that the distance between director and firm affects the outcome only through whether the director is a CEO of another firm. Neither seems plausible to us. Instead, we can imagine an imperial CEO, who wants to limit board oversight, choosing directors who are both far away and not CEOs of other firms. Causation would then run from imperial CEO to distant, non-CEO outside directors. In fairness to the authors, IV is a robustness check; their principal analyses use matching and DiD.

It is usually hard to find a single valid instrument for a particular *gov* measure. If a researcher is lucky enough to find two instruments, a good way to exploit this rare opportunity is likely to use each instrument separately and then assess whether the two estimated causal effects are similar, rather than use both together in the same IV analysis. Two valid instruments may identify different "local" treatment effects (see Angrist and Evans (1998), and Section 6.5 below).

6.2.3 Types of IV Analyses and First Stage Results

Of the non-shock IV papers in our 2010-2011 subsample, 47 use 2SLS; the remaining 17 use three-stage least squares or GMM (Arellano-Bond or similar). In our view, non-IV results should always be reported together with any IV analysis, to allow comparison of results. Of the 64 non-shock IV papers, 45 use OLS as their primary method. Of the remaining 19 IV-primary papers, only 6 report OLS results.

We also believe that first-stage results should be reported in any IV analysis. Doing so can serve several purposes. First, it is part of assessing instrument strength. For multiple instruments, it lets the reader judge whether there is a weak instruments issue and which instrument(s) are driving the overall results. Yet this is not the norm in our sample – only 20 of the 64 non-shock IV papers report the first stage. None of the papers which use only 3SLS (or 3SLS and GMM) do so; for GMM, there is no obvious first stage to report.

6.2.4 When To Use Heckman Selection for Causal Inference: Never!

Many authors use Heckman selection models to address biased selection into treatment, albeit with flawed instruments. Would these models have value, if a good instrument could be found? We think not. If one has a valid instrument for *gov*, one can use standard IV methods. It is not clear why the Heckman method would offer any advantages. If not, the Heckman methodology is not appropriate. There is no middle category.

6.3 Warning Signs for Flawed Instruments

6.3.1 Weak Instruments

A known warning sign for a flawed instrument is low explanatory power in the first-stage regression. The weak instrument problem applies primarily to studies with multiple instruments (Angrist and Pischke, 2009, ch. 4.6.4). A rule of thumb is that an F-test for multiple instruments should exceed 10 to avoid "weak instrument bias" (Stock *et al.*, 2002). Bias can arise even if an instrument is exogenous and meets the only through condition.[86]

[86] Imbens and Rosenbaum (2005) propose the use of randomization inference to develop confidence intervals for a weak instrument. No paper in our sample uses this approach.

Both shock-based and non-shock based instruments can be weak. The 9 shock-based IV papers report information on instrument strength, in various ways. But two report F-values that are well below 10. Even otherwise excellent papers can use what we would today recognize as weak instruments. For example, Desai and Dharmapala (2009) have a first stage F-value of around 3.[87] Among the 49 non-shock papers with multiple instruments, only 13 report evidence of instrument strength; of these, eight satisfy the $F > 10$ rule of thumb; the other five do not.

6.3.2 Coefficient "Blowup" with a Flawed Instrument

A less well known warning sign for an instrument that violates the only through condition is when the IV coefficient estimate is much larger than the OLS estimate. This is common for IV papers in our sample. Consider Table 4 in John and Litov (2010). They report 2SLS estimates of the effect of the G-Index on leverage around five times larger than OLS estimates. One explanation for this large difference is that their instruments likely affect leverage through many channels. IV attributes the instrument's entire effect on leverage to the instrumented variable.

Blowup is especially likely when an instrument: (i) is a weak or moderate predictor of the instrumented variable; and (ii) has some other direct or indirect effect on the outcome. One can see the source of the problem in the Wald estimate for the IV coefficient estimate in eqn. (19). that estimate is (coefficient on instrument in predicting outcome), *divided by* the coefficient on the instrument in predicting the instrumented variable. If the instrument weakly predicts the instrumented variable, the IV estimate will be far larger than one would expect, based on the instrument's direct power to predict the outcome.

Sometimes, the argument for the exclusion restriction is strong but not airtight. If the instrument is strong, a small violation of the exclusion restriction will lead to only moderate bias in the IV estimate. But if the instrument is weak, even a small violation of the exclusion restriction can produce a severely inflated IV coefficient (e.g., Conley *et al.*, 2012), and estimates that are highly sensitive to omitted variables (Small and Rosenbaum, 2008).

6.4 Shock-Based IV

We next turn to the eight shock-based IV papers in our sample. These papers use a variety of shocks. Six use legal shocks – SOX (Duchin *et al.*, 2010; Iliev, 2010); US state laws (Adams and Santos, 2006); Swedish pension reform (Giannetti and Laeven, 2009); US tax rules (Desai and Dharmapala, 2009), and the U.S. Homeland Investment Act (Dharmapala *et al.*, 2011). Of the two remaining papers, Guner *et al.* (2008) use a U.S. banking crisis in the 1970s and Bennedsen *et al.*

[87] Adams and Santos (2006) also report $F < 10$.

(2007) use biological chance that determines the gender of the first-born children of CEOs of family-run firms.

The shock-based IV papers are more likely than non-shock IV papers to address IV basics. For example, all report non-IV results; all report evidence on IV strength, and seven of the eight report first stage results.

An exogenous shock makes it more likely that an IV will be credible, but offers no guarantees. A shock-based IV must still meet the usual IV conditions, including shock strength, exogeneity *of the IV*, and the only through condition. Of the eight shock-based IV papers, only Bennedsen *et al.* (2007) and Iliev (2010) have instruments that appear clean to us. For the others, we have doubts, ranging from mild to strong, about exogeneity, only through, or both. We explain those doubts below.[88]

6.4.1 Direct Use of Shock as IV

There are two principal ways to exploit a shock in an IV analysis. The simpler approach is direct use of the shock as the IV. If a rule applies to some firms but not to other similar firms, one can use a dummy for the rule as an instrument. For example, Black *et al.* (2006) use a discontinuity in Korean corporate governance rules at assets of 2 trillion won as an instrument for *gov*. Of our eight shock-based IV papers, five use the shock directly as an instrument. If the shock is exogenous, the IV will be as well.[89]

As we discuss in §2.9, if one has panel data covering both before and after the shock, direct IV is similar to "intent-to-treat" DiD and should lead to similar results. There are subtle differences in inference, which are not explored in any of the eight shock-based IV papers, which we treat as beyond our scope.[90] If only post-shock data is available shock-based IV is still an available strategy, but DiD is not.

[88]We made these assessments before undertaking, in Atanasov and Black (2016b), a close re-examination of Iliev (2010), Desai and Dharmapala (2009), and Duchin *et al.* (2010). In that re-examination, we conclude that all three of these papers have invalid instruments, for varying reasons.

[89]In addition to Black *et al.* (2006), see Adams and Santos (2006); Bennedsen *et al.* (2007); Guner *et al.* (2008); and Iliev (2010).

[90]In brief, IV provides a LATE estimate, limited to compliers (see §6.6). DiD provides an average treatment effect for all treated firms relative to all control firms. The DiD control group will include "always-takers" who arrange to receive the treatment even though not "encouraged" to do so by the shock. If compliance is not mandatory, the treated group will include some "never takers" who refuse the treatment if it is offered. In contrast, the IV estimate is limited to compliers. Because the IV estimate is based only on compliers, it will tend to be numerically larger than the DiD estimate, but not statistically stronger. See Atanasov and Black (2016b) for further discussion of the difference between DiD and shock-based IV estimates.

6.4.2 Shock Interacted with Pre-Shock Firm Characteristics

A more subtle approach, used by three papers in our sample, involves interacting the shock with a pre-shock covariate (potentially more than one) that correlates with the firm's sensitivity to the shock. Dharmapala *et al.* (2011) use a tax holiday for U.S. firms that repatriate foreign income to construct IVs for repatriation. The IVs interact a post-shock dummy with two measures of the firm's sensitivity to this shock. Duchin *et al*. (2010) interact a legal shock with a measure of the firm's information environment and Gianetti and Laeven (2009) interact their shock (reform that increases pension fund purchases of Swedish equities) with several pre-shock measures of which firms these funds are likely to favor.

Using a pre-shock firm covariate as an IV is similar to using lagged firm characteristics to instrument for current characteristics. That is usually suspect, so why is it plausible here? The interaction-based IV can be credible because the shock is exogenous. In effect, the shock breaks the endogeneity between lagged and current firm characteristics. Just as direct shock-based IV with an interaction-based IV is similar to DiD, shock-based IV with panel data is similar to a DIDID on DIDID-continuous design.

6.4.3 Shock Interacted with Post-Shock Firm Characteristics: Loss of Exogeneity

A shock-based IV should involve interacting the shock only with *pre-shock* firm characteristics, presumably measured just before the shock. If one interacts the shock dummy with a time varying covariate, endogenous variation in the covariate during the post-shock period will cause endogenous variation in the IV, because in the post-shock period the IV is simply (1 * covariate).

Two of the eight shock-based IV papers in our sample fall into this trap. Desai and Dharmapala (2009) interact a nicely exogenous shock (1996 tax reform) with three time-varying covariates, NOLs, long-term debt, and short-term debt over 1997-2001. And Guner *et al.* (2008) interact their shock (the 1980-1982 financial crisis) with post-crisis cash flow.

6.5 Elements of Shock-Based IV Design

The elements of good shock-based design, discussed in §4 for DiD, largely apply to shock-based IV. They often have special force for IV because of blowup risk. We discuss here selected aspects of shock-based IV design. Our discussion stresses aspects that are common accross shock-based designs, not just IV. See Angrist and Pischke (2009, ch. 4) for extended discussion of "causal IV."

6.5.1 Covariate Balance: Is the Shock as Good as Random?

To satisfy the exogeneity and only through conditions, a shock needs to be as good as randomly assigned – shocked firms should be similar to non-shocked firms on all relevant attributes. One should check for covariate balance between treated firms and control firms, rather than assuming that balance exists. Poor balance can provide a warning for possible instrument validity. Conversely, one may be able to strengthen instrument validity by combining IV with balancing methods (Keele and Morgan, 2013).

With panel data, the check for balance should include checking for divergent pre-treatment trends. For IV, even more than for DiD, there is no good solution for non-parallel trends; instead the exogeneity and only through conditions are suspect.

Among the shock-based IV papers in our sample, only two check for covariate balance. Bennedsen *et al.* (2007), confirm that firms whose CEOs have first-born sons are similar to firms whose CEOs have first-born daughters. And Duchin *et al.* (2010) show that pre-SOX compliers and non-compliers with the SOX and NYSE/NASDAQ requirement for board composition have similar pre-SOX trends for one of their outcome variables.

6.5.2 Indirect Channels and the Only Through Condition

Even if a shock is exogenous appears to be as good as randomly assigned, it must still satisfy the only through condition(s). The more random the rule appears to be, with regard to the studied effect, the more plausible this condition will be. Of the eight shock-based IV papers, only five address this core issue. For example, Adams and Santos (2006) argue that the rules for voting by corporate trustees can plausibly affect bank performance only through the bank's managers voting the bank's own shares, held in trust. We worry that home state could predict performance in other ways, but the authors address the only through condition, and their claim is plausible.

In contrast, Giannetti and Laeven (2009) do not discuss the only through condition, and their IV likely violates this condition. They find that share prices of larger Swedish firms, favored by large institutions, rise relative to smaller firms over a 5-year period following reforms that increase pension fund investment in equities. One must believe that the *only* reason for the relative rise in share prices is the increase in pension fund cash flow rights from 13.6% in 2000 to 19.1% in 2005 (with a smaller rise in voting rights). To us, this seems implausible.[91]

For a broad shock, such as SOX, one must ensure that the treated and control firms are similar on other dimensions affected by the shock. One could try to control for the other effects of a broad reform like SOX, but this will rarely be convincing.

[91] In Atanasov and Black (2016b), we discuss all eight shock-based IV papers, and our views on whether each is likely to satisfy the only through condition.

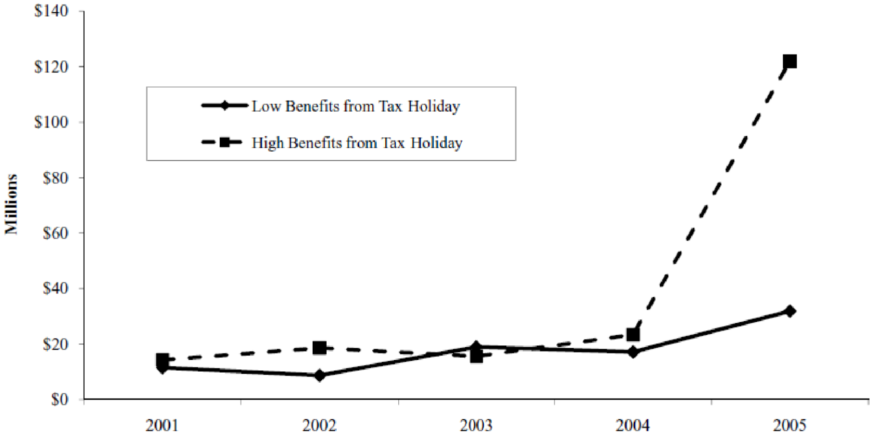

Figure 3: IV Strength Example (from Dharmapala *et al.* (2011))

Description: Mean repatriations for different types of U.S. multinationals. Dashed line displays mean repatriations for firms that are expected to have high benefits from a tax holiday. Firms are expected to have high benefits from the tax holiday if, in 2004, they face lower corporate tax rates abroad and have an affiliate that is a holding company or in a tax haven. The solid line displays mean repatriations for firms that do not meet either of these criteria.

6.5.3 IV Strength: Graphing the First Stage

In addition to the usual tests of IV strength, shock-based IV lends itself to graphing first stage strength. This can provide powerful evidence that the shock is doing what the design assumes. Dharmapala *et al.* (2011) offer an example. They have two IVs, constructed by interacting a tax holiday shock with measures of firms' expected benefits from the holiday. They graph the pre-shock to post-shock change in the instrumented variable (repatriation of foreign income) for "treated" firms that (i) have lower tax rates outside the U.S. and (ii) have an affiliate that is a holding company or incorporated in a tax have; versus "control" firms that have neither of these attributes. Figure 3 shows their results.

6.5.4 How to Handle Post-Treatment Covariates

Suppose one has panel data both before and after the shock. Much as for DiD, it is not obvious how to handle post-shock covariates, which could be affected by the shock (see §4.3). It is customary in IV analysis to include covariates. The 2SLS analysis assumes they are exogenous, but this only assumes away the problem.

In our shock-IV sample, only Bennedsen *et al.* (2007) discuss this issue. They provide evidence that their shock does not affect firm characteristics other than the one they study (family succession). All papers except Bennedsen *et al.* (2007) and Dharmapala *et al.* (2011) use post-shock covariates.

6.5.5 Multiple Instruments

Six of our eight shock-based IV papers use multiple instruments. Given the threats to IV validity, especially the only-though condition, we'd prefer fewer instruments, ideally only one where feasible (this is not feasible if one is studying an interaction). This facilitates careful assessment of instrument validity. In particular, one should not mix shock-based and non-shock IVs. The non-shock IV is unlikely satisfy the exogeneity and only through conditions.

If a paper has good reason to use multiple instruments, we'd like to see results with each instrument separately, as a robustness check. If results are similar, this supports instrument validity; if not, then not. For example, Dharmapala *et al.* (2011) use two instruments, created by interacting their tax holiday shock with two firm-level measures of the value of repatriating foreign income. We'd prefer to see results for each separately.

6.5.6 Instrumenting for What?

Sometimes, one has a reasonably clean instrument for *some* aspect of *gov*, but it's less clear what aspect. For example, Black *et al.* (2006) (which we classify as a fuzzy RD paper rather than a shock-IV paper) use a 1999 Korean legal reform as a shock to the overall governance of the large firms (assets $>$ 2 trillion won) subject to the reform. These reforms directly hit only board structure, so one could instead treat this shock as affecting only board structure. In effect, there is uncertainty about the channel through which the shock operates – board structure versus all of governance. The authors address this issue by reporting results both ways.

Shock-based IV, unlike DiD using the same shock, requires a first stage. It thus forces one to be explicit about the channel through which the shock affects the outcome. Compare our discussion in §4.2.9 of the need with a DiD design to verify shock strength with a latent forced variable.

6.5.7 Placebo Tests

Shock-based IV, much like DiD, allows the researcher to conduct a "placebo" test: the instrument should predict the outcome in *2SLS* only after the shock, not before. In our sample, Guner *et al.* (2008) perform such a placebo test. They use the number of directors appointed during a banking crisis as an instrument for the number of *bankers* serving on company boards after the crisis. They report that a placebo instrument (the number of directors appointed in a non-crisis period) does not predict the instrumented variable in the first stage, nor the outcome in the second stage.

6.5.8 Is an IV Needed?

Suppose that you have an instrument that appears to satisfy the exogeneity and only-through conditions. What can you usefully do, besides running *2SLS*? One logical step is to run a Hausman (or equivalent, such as Durbin-Wu-Hausman) test for endogeneity. These tests assume a valid instrument. In Durbin-Wu-Hausman, the first stage is the same as the first stage of *2SLS*; in the second stage one adds the residual from the first stage to the usual *2SLS* second stage. A significant coefficient on the first-stage residual implies rejection of the null of no endogeneity.[92] If the test does not reject the null, one gains comfort in non-IV methods, which do not have blowup risk and usually have smaller standard errors.[93]

6.5.9 Choosing Between DiD and Shock-Based IV

As we discuss in Section 2, DiD and shock-based IV are close cousins. In our view, whenever researchers use shock-based IV, they should also run DiD (or one of its continuous variants). One advantage of doing so: IV assumes that the instrument affects the outcome only through the instrumented variable. DiD does not – instead, the coefficient on the shock dummy estimates the total effect of the shock on the outcome. In Atanasov and Black (2016b), we discuss how one can understand DiD as akin to an intent-to-treat estimate.

6.6 LATE: What a Valid Instrument "Identifies"

If there are heterogeneous treatment effects, IV estimates a local average treatment effect (LATE), for the subsample of firms which change *gov* as a result of the instrument (whom Imbens and Angrist (1994), call "compliers"). It will not estimate an effect for "always takers" who would have adopted the governance reform anyway, or for "never takers" who would not adopt the reform, whether the rule exists or not.

One must also assume no "defiers" – firms who would adopt a reform without the rule, but won't do so with the rule. This assumption will often be reasonable for

[92] See Wooldridge (2010), §6.3.1. In our dataset, Black *et al.* (2006) use this test.

[93] A warning. Suppose, as in Black *et al.* (2006), that the instrument for *gov* affects only part of an overall governance measure. The Durbin-Wu-Hausman test will fail if the treatment effect for the part of *gov* that you can instrument for differs from that for the part you can't instrument for, which is not testable. In effect, the coefficient on the first-stage residual in the Durbin-Wu-Hausman regression is affected by both endogeneity and the relative strength of the instrumented and uninstrumented portions of governance in predicting the outcome. Thus, Durbin-Wu-Hausman may wrongly reject the null of no endogeneity, or wrongly fail to reject the null, if instrumented *gov* is significantly stronger or weaker than the uninstrumented *gov* in predicting the outcome.

legal shocks. For example, it is unlikely that many firms will adopt a governance reform voluntarily, yet will refuse to do so if a law requires this.

The LATE estimate will equal the population average effect only if treatment effects are the same for all firms in the population. This is a strong assumption, and for corporate governance studies, usually an unjustified one. For example, Bennedsen *et al.* (2007) estimate the causal effect of having a second generation child as CEO versus hiring an external CEO only for: (i) family-owned firms; who (ii) would choose an external CEO if the former CEO's first-born was a girl, but would have chosen a family CEO if the first-born was a boy. They cannot estimate the effect for firms which would choose an external CEO regardless of the sex of the first-born child ("always-takers"), nor for firms which would choose a family CEO in either case ("never-takers").

Similarly Iliev (2010) study of the cost of SOX §404 compliance for firms near the regulatory threshold (public float of \$75M in *2004*) estimates this effect for firms whose behavior would be changed by his instrument (public float > \$75M in *2002*). The observed compliers with his instrument are shrinkers – firms whose float shrank from above \$75M in 2002 to below \$75M in 2004, which were forced to comply with SOX §404 because they were above the \$75M threshold in 2002. Compliance costs for these firms could be different than for non-shrinking firms.

For an indirect shock-based IV design, there is no clean line between compliers and other firms, but the concepts behind causal IV are similar. One is again estimating a local treatment effect, to which firms that are more affected by the shock contribute more strongly than less affected firms.

An IV estimate is also local in a second sense. One estimates a causal effect only for the aspect of *gov* affected by the rule. For example, the Korean reforms studied by Black *et al.* (2006) involve board structure (principally independent directors and audit committees). Thus, their IV is an instrument only for board structure (plus other aspects of governance which are affected by board structure).

6.7 Good Practice IV Paper

Bennedsen *et al.* (2007) is our good practice paper for shock-based IV. They study the effect of family succession on firm performance in Denmark, in both public and private firms. They find that family CEO succession causes 4% lower ROA. The key causal inference challenge is that CEO succession is a firm choice, which will be influenced by many factors, including the firm's future prospects.

The authors instrument the choice of an internal (family member) CEO with the gender of the prior CEO's first-born child. The idea is that internal succession is more likely if the first-born child is male. The instrument is exogenous – the child's gender was randomly determined (this might be less true today). It is relevant, with a strong first stage: 39% of firms with a male-first-born child appoint a family CEO, versus 29% of firms with a female-first born. The authors carefully defend the only-through assumption. They check covariate balance and verify that firm and family characteristics of treated (male-first) and control (female first) firms are similar at the time of CEO succession. Treated and control firms have similar age, size, and profitability; families are similar on size, divorce rate, and the CEO's number of spouses.

The authors note that their IV estimates are valid only for "complier" firms, whose succession decision is affected by first-born gender. They show, however, that the gap between the professional skills of family and external CEOs is the same for male-first-born and female-first-born firms, which suggests that the estimates may be reasonable for all Danish firms.

For our other good-practice papers, we note what else the authors might have done, to improve an already strong paper. For Bennedsen et al., we have no meaningful suggestions. They have, in effect, a randomized experiment, with an encouragement design. This is a beautiful paper.

7 (Regression) Discontinuity Designs

7.1 Overview of RD

RD designs, especially if combined with DiD, can approach the gold standard of a randomized experiment. The core idea behind of RD is that if firms on one side of an arbitrary threshold for the forcing variable are treated, while firms on the other side are not, assignment to treatment may be as good as random for firms close to the threshold.

These designs are becoming more popular in finance, but there are only two RD papers in our sample. One is Black *et al.* (2006), who study the Korean corporate governance reforms in 2001, which applied only to public companies with assets > 2 trillion won in assets. In a followup study beyond our sample period, Black and Kim (2012) use a combined DiD/RD design. The second is Iliev (2010), discussed below as a good practice RD paper. Our discussion of RD design is summary in nature. For more details, see the reviews by Imbens and Lemieux (2008) and Lee and Lemieux (2010).

7.2 Elements of RD Design

A number of elements of RD design are similar to DiD and IV design, so little more need be said. Covariate balance should be checked for all covariates except the forcing variable. In our sample, Iliev does this; Black *et al.* (2006) do not, but a second study (Black and Kim, 2012) does so. With panel data that covers both before and after the shock, one should also confirm pre-shock balance on the outcome variable, and parallel pre-treatment trends between the treated and control groups. For RD, as for all shock based designs: SUTVA Independence is a concern; one should check for differential attrition; and one must assess whether the shock, rather than some other rule associated with the shock, could explain the observed treatment effect.

We discuss below some RD-specific design elements. We first discuss "sharp" RD designs, in which the probability of observing the forced variable is 0 below the threshold, and 1 above the threshold.

7.2.1 Shock Strength

RD designs lend themselves to graphical depiction. The forced variable should jump at the threshold – for a sharp design, from 0 to 1. Ideally, the outcome variable (q, say) should also visibly jump. However, noise may sometimes obscure the jump in the outcome variable, so that the jump emerges from regression analysis but is not visually apparent.

7.2.2 Random Nature of the Threshold

An ideal RD design would use a forcing variable that does not directly predict the outcome. Most rules are not quite that random. But sometimes one gets lucky. For example, Dharmapala and Khanna (2013) exploit a discontinuity in the application of Indian corporate governance rules based on firms' charter capital – a bookkeeping measure that has little relevance to actual book value (which they confirm for firms within their bandwidth).

7.2.3 Bandwidth

Most RD designs study only firms within a relatively narrow "bandwidth" around the discontinuity. The further away one goes from the discontinuity, the weaker the claim that assignment to treatment is random. Yet the narrower the bandwidth, the smaller the sample. Thus, the choice of bandwidth around the discontinuity, and assessing robustness to bandwidth choice, are central aspects of research design.

The broader the bandwidth, the more important it can be to control for other pre-treatment covariates, and to control flexibly for a direct effect of the forcing variable on the outcome. Options for the forcing variable include a polynomial in the forcing variable, as opposed to the simple linear control in eqn. (16), "nonparametric" local linear regression, or separate regressions on either side of the bandwidth. Theories of optimal bandwidth exist, but we favor a sensitivity approach, in which one shows how the treatment effect changes as one varies the bandwidth (see Atanasov and Black (2016b), for an example). If the effect is real, then as one narrows the bandwidth, the coefficient on the treatment dummy should be reasonably stable. If the coefficient shrinks, this is a trouble sign. One can also assess robustness to different ways of controlling for the forcing variable.

As the bandwidth gets narrower, one needs to use caution in controlling for the forcing variable, lest a flexible control absorb the jump at the discontinuity. As one narrows the bandwidth, collinearity between the treatment dummy and the forcing variable will rise at the same time that sample size falls. One may face a choice between using a narrow bandwidth and not controlling for the forcing variable, versus using a broader bandwidth with this control; or a choice between a narrower bandwidth with a simple control for the forcing variable, and a broader bandwidth with a more flexible control.

7.2.4 Bin Width

To graph the discontinuities in the forced variable and the outcome, one may need to "bin" observations, and thus choose a bin width. Theories of optimal bin width exist, but we again favor a sensitivity approach, in which one tries different widths. If the effect is real, a graphical impression of the size of the jumps should be similar across a range of bin widths. If not, this is a trouble sign.

7.2.5 Threshold Manipulation

An important threat to RD validity is the risk that, the forcing variable aside, assignment is not truly random – that the firms on each side of the threshold differ in important ways, perhaps unobserved. A check for covariate balance helps, but is not sufficient. Suppose, in particular, that firms can manipulate the forcing variable to fall on their preferred side of the discontinuity. Firms with higher compliance costs or lower benefits might be more likely to avoid a rule than firms with lower costs – this would lead to a biased estimate of the rule's effect.

Researchers can address the risk of manipulation in several ways. One approach involves arguing that the forcing variable is non-manipulable. This is feasible if a rule uses pre-rule values of the forcing variable to determine compliance. A second approach is to show that even though the threshold could be manipulated, it is not manipulated in practice. Evidence for lack of manipulation includes similar densities of firms for values of the forcing variable just below vs. just above the threshold (McCrary, 2008). Black *et al.* (2006) assess manipulation of their threshold (2 trillion won in assets) and find no evidence of manipulation.[94] Iliev assesses manipulation of his threshold ($75M in public float in 2004), and finds evidence that some firms manipulate their float to remain

[94]Black and Kim (2012) verify similar density above and below the regulatory threshold. They also examine each firm which shrinks to below the threshold and assess whether business reversal is a likely cause of shrinking. Dharmapala and Khanna (2013) compare firms that are subject to India's Clause 49 governance rules because they have charter capital (a nearly arbitrary number, only loosely related to book value of equity) above a regulatory threshold to firms below the threshold; they confirm that the post-shock "switchers", who reduce their charter capital, are few in number and that their results hold if they drop these firms from the sample.

below the threshold (see also Gao *et al.* (2009)). He uses an IV design to address the manipulation.

7.2.6 Discrete Forcing Variable: DiD/RD-Discrete Design

An ideal forcing variable is continuous above and below the threshold, or nearly so. But even if not, one can move toward an RD design by narrowing the sample to include only treated and control firms that are similar on the forcing variable. Thus, for example, in the Dahya and McConnell (2007) study of the Cadbury Committee recommendation that UK public firms have at least three non-executive directors, one could limit the treated group to firms with two non-executive directors and the control group to firms with three non-executives, just before the Cadbury report came out. This design has no name; one might call it an "RD-discrete" design.

An RD-discrete design, without more, might be only moderately credible. But if combined with DiD, it can improve on DiD alone, in which all firms below the compliance threshold are treated, and all firms above it are controls. Coming closer to balance on the forcing variable will likely also improve balance on other covariates.

7.2.7 Fuzzy RD as IV

We discuss here additional design elements that apply to "fuzzy" designs, in which the probability of observing the forced variable jumps at the threshold, but not all the way from 0 to 1. One then has, in effect, an encouragement design. One can ignore the partial compliance and develop an intent-to-treat estimate. Or, more commonly, one can use the discontinuity as an instrument for actual treatment. The latter approach implicates the usual concerns with any IV.

Usually, even with a fuzzy RD design, one can graphically see the first-stage discontinuity in the forced variable. If not, one worries about shock strength.

As with any encouragement design, one estimates a treatment effect only for compliers – firms whose behavior depends on which side of the threshold they are on.

With panel data, the degree of fuzziness can change during the post-shock period. Consider, for example, a requirement to adopt an audit committee, for firms above a size threshold (as in the Black *et al.* (2006) study of Korea). If investors react favorably to audit committees, then over time, below-threshold firms may adopt these committees voluntarily. As they do, the design becomes increasingly fuzzy.

7.2.8 Fuzzy RD: Who are the Compliers?

To continue with the audit committee example, suppose the rule requires above-threshold firms to adopt these committees. At the same time, some below-threshold firms do so voluntarily. Who then are the "compliers"? The non-intuitive answer: *Not* the firms who would voluntarily adopt audit committees. They are the always-takers. Instead, the compliers are the firms who adopt audit committees only if forced to do so. Within the control group, these are the firms that do not adopt audit committees. One might call them "instrument-compliers," as distinguished from "rule-compliers" – firms that simply obey the rule.

An example of how this distinction matters: Suppose that adopting an audit committee will add value for some firms but not others, firms know perfectly which group they are in, and firms that benefit from an audit committee will adopt one voluntarily. Then the audit committee requirement could add value on average, and yet the fuzzy-RD (as IV) estimate of the value of audit committees will be zero, which is the right answer for firms that are instrument compliers. More generally, if there are heterogeneous treatment effects, and firms with higher treatment effects tend to comply voluntarily, the fuzzy-RD estimate will understate the average effect for all firms near the threshold.

7.2.9 Local Nature of the RD Estimate

An RD study can provide a credible treatment effect estimate only for firms close to the discontinuity, in three senses. First, the further one gets from the discontinuity, the weaker the claim that firms on either side are similar. Second, if treatment effects are heterogenous, credible inference is also limited to a reasonable bandwidth around the discontinuity. Third, with a fuzzy RD design, even within the bandwidth, inference is limited to the "instrument-compliers."

7.3 Combined DiD/RD and DiD/RD-discrete Designs

7.3.1 Value of Combined Designs

If one has a shock to firms above a threshold, and data both before and after the shock, it seems natural to combine DiD and RD designs. One limits the sample to a band around the threshold, and runs DiD on this limited sample. The limited bandwidth strengthens DiD by addressing a central DiD challenge – ensuring that treated and control firms are sufficiently similar. And the DiD design strengthens RD by controlling for any pre-shock differences between the treated and control groups. Similarly, for a discrete forcing variable, it feels natural to use a combined DiD/RD-discrete design. One can similarly combine an event study or

IV design with RD. Iliev (2010) provides an example – he uses a combined ES/IV design.[95]

7.3.2 Elements of Combined DiD/RD Design

For a combined DiD/RD design, the need to choose a bandwidth, and to assess sensitivity of results to this choice, is similar to any RD analysis. However, a combined DiD/RD analysis may tilt toward a larger bandwidth (and thus larger sample size) than would be optimal for pure RD, because the DiD approach controls for pre-shock differences in levels between treated and control firms, leaving only differences in after-minus-before changes to worry about.

For similar reasons, a combined DiD/RD design may tilt toward a less flexible control for the forcing variable, and thus less collinearity between this control and the above-threshold dummy. One can see the combined DiD/RD design as replacing the RD assumption that the flexible control for the forcing variable captures the forcing variable's direct effect on the outcome with the DiD parallel trends assumption – the distribution of outcomes for the treated group (if not treated) would move in parallel through time with the distribution of outcomes for the control group). This can be checked pre-shock.

7.4 Good practice RD Paper

Iliev (2010) is our good-practice RD paper and, for overall care with research design, our second favorite shock-based paper (after Bennedsen *et al.* (2007)). He studies the effect of SOX §404, which requires auditors to confirm the quality of firm internal controls for fiscal year 2004 and later. In 2003, the SEC exempted firms with public float (shares not held by insiders) <$75M in each of 2002, 2003, and 2004. The SEC gave below-threshold firms until 2007 to comply; the exemption was later made permanent.

Iliev limits his sample to firms with public float in 2004 between $50 and $100M. He assesses robustness with broader and narrower bandwidths, and also uses placebo thresholds at $125M and $150M. He verifies that the outcome variable (audit expense) is balanced near the threshold in 2002, during the pre-shock period. He uses a cubic in public float as his control for the forcing variable, along with other firm size controls. He first estimates a standard RD specification for audit fees in 2004 using a dummy for SOX §404 compliance.

A concern with this design is that firms may manipulate their public float to stay below the $75M threshold. The SEC rule was adopted in 2003. Firms that were below the $75M threshold in 2002 could take actions to stay below the

[95] Outside our sample period, Black and Kim (2012) use combined DiD/RD, event study/RD, and IV/RD designs to study the impact of corporate governance reforms. Lemieux and Milligan (2008) use a combined first differences/RD design.

threshold in 2003 and 2004. Iliev finds that the density of firms is smooth around the \$75M threshold in 2002 and 2003, but in 2004 and 2005, there is more mass below the threshold than above it. This is evidence that some firms manipulate their public float to avoid complying with SOX §404. To address this manipulation, he uses (public float > \$75M in *2002*) as an instrument for SOX §404 compliance in 2004. This instrument is credible because the SEC specified the \$75M threshold only in June, 2003, when it was too late for firms to change their 2002 float. It is relevant because it strongly predicts compliance in 2004. If one carefully controls for firm size (which Iliev does), this instrument plausibly satisfies the only-through condition.[96]

Iliev (2010) also performs an event study around the dates relevant to the SEC's adoption of the below-\$75M exemption. His event study design includes a placebo date test, and multiple event dates with a reversal. He limits the sample to firms close to the threshold during the relevant events in 2003 – thus employing a combined event study/RD design.

Turning to potential improvements: Iliev could have usefully recast his study as a combined DiD/RD design, with firms above \$75M in 2002 float as the treated group and firms below this threshold in 2002 as the control group. The IV and DiD designs will be similar. The only through condition for IV would be effectively replaced by the DiD parallel trends assumption.[97] He defines his bandwidth using float in 2004 – in our view, float in 2002, before the SEC adopted its rule, would have been a better choice. And he does not consider the potential for heterogenous treatment effects.

8 Conclusion

Research designs based on exogenous shocks, often law-based, are becoming increasingly common in corporate finance and accounting research. When they are available, shock-based designs can often form a stronger basis for credible causal inference than the best available non-shock designs. This article has discussed how to use and improve shock-based designs. A central perspective is that one should treat the *shock* as a central object in research design. Each major method for exploiting shocks – DiD, ED, RD, and IV – has some special features. But these methods share many common aspects. All seek to approach the ideal of random assignment of treatment. All therefore require that the shock be as good as randomly assigned – with only minor variations in the meaning of as-if-random assignment. Other core requirements for credible causal inference are also common across methods; we capture these as requirements for a "good shock."

[96] We re-examine Iliev's instrument in Atanasov and Black (2016b) and conclude that it does not satisfy the only-through condition, for subtle reasons which are best explained there.

[97] Iliev (2010) runs an unreported DiD regression as an alternative to his main RD design, but the treated and control groups are defined using 2004 free float, which firms can manipulate.

Often, several methods can be used to exploit the same shock, and inference can be strengthened by use of combined methods.

In our experience, the common aspects of shock-based design are not widely understood. They have been obscured by the practice, in both the methods literature and the empirical literature, of focusing on one method at a time.

Our study of what corporate governance researchers actually do finds large potential for improved research design. Among the 863 empirical corporate governance papers in our study, only 74 (9%) use a shock-based design. The percentage of papers with shock-based designs is only modestly higher (11%) in the second half of our study period. Even the shock-based papers often fall well short of best practice. Only a few address our five conditions for a "good shock" – shock strength; exogenous shock; as-if-random assignment; covariate balance; and only-through condition(s) – in a satisfactory manner. For those that do not, some results would surely survive – that a condition is not explored does not mean it would not be met. But many results likely would not survive.

We, as researchers, can do better. This article is a how-to guide for doing so – for making shock-based research design all that it can be.

Acknowledgements

Our paper benefited from comments from the editor, two anonymous referees, Renée Adams, John Boschen, Dhammika Dharmapala, Mara Faccio, Merritt Fox, Vladimir Gatchev, Scott Gibson, Peter Iliev, Inessa Love, W. Bentley MacLeod, Ron Masulis, John Matsusaka, John McConnell, Mat McCubbins, Randall Morck, Marina Niessner, Chuck Schnitzlein, Kim Smith, Yishay Yafeh, and participants in the Brazil Finance Association 2007 annual meeting, Third International Conference on Corporate Governance in Emerging Markets (Seoul, Korea, 2011), Canadian Law and Economics Association 2011 annual meeting; 2014 Columbia University Conference on Legal Innovation: Law, Economics and Governance, 2014 Strategy Research Initiative conference, and seminars at the College of William and Mary, George Washington University, Purdue University, Texas A&M University, University of Central Florida, University of Exeter, Virginia Commonwealth University, and York University. We thank Elise Nelson, Sumin Kim, and Sarah Reibstein for excellent research assistance; and the Searle Center on Law, Regulation and Economic Growth at Northwestern Law School for financial support.

References

Abadie, A., A. Diamond, and J. Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association*. 105: 493–505.

Abadie, A., A. Diamond, and J. Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method". *American Journal of Political Science*. 59(2): 495–510.

Abadie, A. and J. Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country". *American Economic Review*. 93: 113–132.

Acemoglu, D., S. Johnson, and R. J. A. 2000. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*. 91: 1369–1401.

Acemoglu, D., S. Johnson, A. Kermani, K. Kwak, and T. Mitton. 2016. "The Value of Connections in the Turbulent Times: Evidence from the United States". *Journal of Financial Economics*. 121: 368–391.

Adams, R., H. Almeida, and D. Ferreira. 2005. "Powerful CEOs and Their Impact on Corporate Performance". *Review of Financial Studies*. 18: 1403–1432.

Adams, R. and J. Santos. 2006. "Identifying the Effect of Managerial Control on Firm Performance". *Journal of Accounting and Economics*. 41: 55–85.

Aivazian, V., Y. Ge, and J. Qiu. 2005. "Can Corporatization Improve the Performance of State-owned Enterprises even without Privatization?" *Journal of Corporate Finance*. 11: 791–808.

Akhigbe, A. and A. Martin. 2005. "Valuation impact of Sarbanes–Oxley: Evidence from Disclosure and Governance Within the Financial Services Industry". *Journal of Banking and Finance*. 30: 989–1006.

Akhigbe, A., A. Martin, and M. Newman. 2010. "Information Asymmetry Determinants of Sarbanes-Oxley Wealth Effects". *Finance Management*. 31: 1253–1272.

Altamuro, J. and A. Beatty. 2010. "How Does Internal Control Regulation Affect Financial Reporting?" *Journal of Accounting and Economics*. 49: 58–74.

Altamuro, J., A. Beatty, and J. Weber. 2005. "The Effects of Accelerated Revenue Recognition on Earnings Management and Earnings Informativeness: Evidence from SEC Staff Accounting Bulletin No. 101". *Accounting Review*. 80: 373–401.

Altonji, J., T. Elder, and C. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools". *Journal of Political Economy*. 113: 151–184.

Angrist, J. and W. Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size". *American Economic Review*. 88: 450–477.

Angrist, J., G. Imbens, and D. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables". *Journal of the American Statistical Association*. 91: 444–455.

Angrist, J. and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Armstrong, C., A. Jagolinzer, and D. Larcker. 2010. "Chief Executive Officer Equity Incentives and Accounting Irregularities". *Journal of Accounting Research*. 48: 225–271.

Atanasov, V. and B. Black. 2016a. “Database of Shocks Used in Corporate Finance and Accounting Research”. Working paper. URL: http://ssrn.com/%20abstract=2227717.

Atanasov, V. and B. Black. 2016b. “The Trouble with Instruments: Re-examining Shock-Based IV Designs”. Working paper. URL: http://ssrn.com/%20abstract=2417689.

Atanasov, V., B. Black, C. Ciccotello, and S. Gyoshev. 2010. “How Does Law Affect Finance? An Examination of Equity Tunneling in Bulgaria”. *Journal of Financial Economics*. 96: 155–173.

Atanasov, V., V. Ivanov, and K. Litvak. 2012. “Does Reputation Limit Opportunistic Behavior in the VC Industry: Evidence from Litigation against VDs”. *Journal of Finance*. 67: 2215–2246.

Autor, D. 2003. “Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing”. *Journal of Labor Economics*. 21: 1–42.

Bae, K. and V. Goyal. 2010. “Equity Market Liberalization and Corporate Governance”. *Journal of Corporate Finance*: 609–621.

Baek, J., J. Kang, and K. Park. 2004. “Corporate governance and firm value: evidence from the Korean financial crisis”. *Journal of Financial Economics*: 265–313.

Bargeron, L., K. Lehn, and C. Zutter. 2010. “Sarbanes-Oxley and Corporate Risk-Taking”. *Journal of Accounting and Economics*. 49: 34–52.

Battalio, R., B. Hatch, and T. Loughran. 2011. “Who benefited from the disclosure mandates of the 1964 Securities Act Amendments”. *Journal of Corporate Finance*. 17: 1047–1063.

Bennedsen, M., K. Nielssen, F. Perez-Gonzalez, and D. Wolfenzon. 2007. “Inside the Family Firm: The Role of Families in Succession Decisions and Performance”. *Quarterly Journal of Economics*. 122: 647–691.

Berger, E., A. Butler, E. Hu, and M. Zekhnini. 2015. “Credit Be Dammed: The Impact of Banking Deregulation on Economic Growth”. Working paper. URL: http://ssrn.com/abstract=2139679.

Berkman, H., R. Cole, and L. Fu. 2010. “Political Connections and Minority-Shareholder, Protection: Evidence from Securities-Market Regulation in China.” *Journal of Financial and Quantitative Analysis*: 1391–1417.

Bertrand, M., E. Duflo, and S. Mullainathan. 2004. “How Much Should We Trust Differences-in-Differences Estimates?” *Quarterly Journal of Economics*. 119: 249–75.

Bertrand, M. and S. Mullainathan. 2003. “Enjoying the Quiet Life? Corporate Governance and Managerial Preferences”. *Journal of Political Economy*. 111: 1043–1075.

Besley, T. and A. Case. 2000. “Unnatural Experiments? Estimating the Incidence of Endogenous Policies”. *Economic Journal*. 110: F672–F694.

Bhagat, S. and R. Romano. 2002a. "Event Studies and the Law: Part I – Technique and Corporate Litigation". *American Law and Economics Review*. 4: 141–167.

Bhagat, S. and R. Romano. 2002b. "Event Studies and the Law: Part II – Empirical Studies of Corporate Law". *American Law and Economics Review*. 4: 380–423.

Black, B., A. de Carvalho, V. Khanna, W. Kim, and B. Yurtoglu. 2014. "Methods for Multicountry Studies of Corporate Governance and Evidence from the BRIKT Countries". *Journal of Econometrics*. 183: 230–240.

Black, B., H. Jang, and W. Kim. 2006. "Does Corporate Governance Affect Firms' Market Values? Evidence from Korea". *Journal of Law, Economics and Organization*. 22: 366–413.

Black, B. and V. Khanna. 2007. "Can Corporate Governance Reforms Increase Firm Market Values? Event Study Evidence from India". *Journal of Empirical Legal Studies*. 4: 749–796.

Black, B. and W. Kim. 2012. "The Effect of Board Structure on Firm Value: A Multiple Identification Strategy Approach Using Korean Data". *Journal of Financial Economics*. 103: 203–226.

Bowen, D., L. Fresard, and J. Taillard. 2016. "What's Your Identification Strategy? Innovation in Corporate Finance Research". *Management Science*. Forthcoming.

Brochet, F. 2010. "Information Content of Insider Trades before and after the Sarbanes-Oxley Act". *The Accounting Review*. 85: 419–446.

Broughman, B. and J. Fried. 2010. "Renegotiation of Cash Flow Rights in the Sale of VC-backed Firms". *Journal of Financial Economics*. 95: 384–399.

Brown, J., N. Liang, and S. Weisbenner. 2007. "Executive Financial Incentives and Payout Policy: Firm Responses to the 2003 Dividend Tax Cut." *Journal of Finance*. 62: 1935–1965.

Brown, P., W. Beekes, and P. Verhoeven. 2011. "Corporate Governance, Accounting and Finance: A Review". *Accounting and Finance*. 51: 96–172.

Busso, M., J. DiNardo, and J. McCrary. 2014. "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators". *Review of Economics and Statistics*. 96: 885–897.

Cameron, C., J. Gelbach, and D. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors". *Review of Economics and Statistics*. 90: 414–427.

Cameron, C., J. Gelbach, and D. Miller. 2011. "Robust Inference with Multiway Clustering". *Journal of Business and Economic Statistics*. 29: 238–249.

Carvalhal da Silva, A. and A. Subrahmanyam. 2007. "Dual-class Premium, Corporate Governance, and the Mandatory Bid Rule: Evidence from the Brazilian Stock Market". *Journal of Corporate Finance*. 13: 1–24.

Catan, E. and M. Kahan. 2016. "The Law and Finance of Anti-Takeover Statutes". *Stanford Law Review*. Forthcoming.

Chang, H., J. Chen, and W. Liao. 2006. "Birendra K. Mishra, CEOs'/CFOs' Swearing by the Numbers: Does It Impact Share Price of the Firm?" *The Accounting Review*. 81: 1–27.

Chava, S. and A. Purnanandam. 2010. “CEOs versus CFOs: Incentives and corporate policies”. *Journal of Financial Economics*. 97: 263–278.

Chen, S., R. Chou, and S. Chou. 2009. “The Impact of Investment Opportunities and Free Cash Flow on Financial Liberalization: A Cross-Firm Analysis of Emerging Economies”. *Financial Management*: 543–566.

Cheng, Q. and D. Farber. 2008. “Earnings Restatements, Changes in CEO Compensation, and Firm Performance”. *The Accounting Review*. 83: 1217–1250.

Chetty, R. and E. Saez. 2005. “Dividend Taxes and Corporate Behavior: Evidence from the 2003 Dividend Tax Cut”. *Quarterly Journal of Economics*. 120: 791–833.

Chhaochharia, V. and Y. Grinstein. 2007. “Corporate Governance and Firm Value: The Impact of the 2002 Governance Rules”. *Journal of Finance*. 62: 1789–1825.

Chhaochharia, V. and Y. Grinstein. 2009. “CEO Compensation and Board Structure”. *Journal of Finance*. 64: 231–261.

Choi, J., S. Park, and S. Yoo. 2007. “The Value of Outside Directors: Evidence from Corporate Governance Reform in Korea”. *Journal of Financial and Quantitative Analysis*. 42: 941–962.

Cicero, D. 2009. “The Manipulation of Executive Stock Option Exercise Strategies: Information Timing and Backdating”. *Journal of Finance*. 64: 2627–2663.

Claessens, S. and B. Yurtoglu. 2013. “Corporate Governance in Emerging Markets: A Survey”. *Emerging Markets Review*. 15: 1–33.

Coates, J. 2000. “Takeover Defenses in the Shadow of the Pill: A Critique of the Scientific Evidence”. *Texas Law Review*. 79: 271–382.

Coles, J. L., M. L. Lemmon, and J. Felix Meschke. 2012. “Structural models and endogeneity in corporate finance: The link between managerial ownership and corporate performance”. *Journal of Financial Economics, Elsevier*. 103(1): 149–168.

Conley, T., C. Hansen, and P. Rossi. 2012. “Plausibly Exogenous”. *Review of Economics and Statistics*. 94: 260–272.

Conley, T. and C. Taber. 2011. “Inference with “Difference in Differences” with a Small Number of Policy Changes”. *Review of Economics and Statistics*. 93: 113–125.

Cremers, M. and A. Ferrell. 2014. “Thirty Years of Shareholder Rights and Firm Valuation”. *Journal of Finance*. 69: 1167–1196.

Crump, R., V. Hotz, G. Imbens, and O. Mitnik. 2009. “Dealing with Limited Overlap in Estimation of Average Treatment Effects”. *Biometrika*. 96: 187–199.

Dahya, J. and J. McConnell. 2007. “Board Composition, Corporate Performance, and the Cadbury Committee Recommendation”. *Journal of Financial and Quantitative Analysis*. 42: 535–564.

Dahya, J., J. McConnell, and N. Travlos. 2002. “The Cadbury Committee, Corporate Performance, and Top Management Turnover”. *Journal of Finance*. 57: 461–483.

Daske, H., L. Hail, C. Leuz, and R. Verdi. 2008. “Mandatory IFRS Reporting around the World: Early Evidence on the Economic Consequences”. *Journal of Accounting Research*. 46: 1085–1141.

DeFond, M., X. Hu, M. Hung, and S. Li. 2011. “The Impact of Mandatory IFRS Adoption on Foreign Mutual Fund Ownership: The Role of Comparability”. *Journal of Accounting and Economics*. 51: 240–258.

Demirguc-Kunt, A., I. Love, and V. Maksimovic. 2006. “Business Environment and the Incorporation Decision”. *Journal of Banking and Finance*. 30: 2967–2993.

Desai, M. and D. Dharmapala. 2009. “Corporate Tax Avoidance and Firm Value”. *Review of Economics and Statistics*. 91: 537–546.

Dharmapala, D., F. Foley, and K. Forbes. 2011. “Watch What I Do, Not What I Say. The Unintended Consequences of the Homeland Investment Act”. *Journal of Finance*. 66: 753–787.

Dharmapala, D. and V. Khanna. 2013. “Corporate Governance, Enforcement, and Firm Value: Evidence from India”. *Journal of Law, Economics and Organization*. 29: 1056–1082.

Dinc, S. 2005. “Politicians and Banks: Political Influences on Government-Owned Banks in Emerging Markets”. *Journal of Financial Economics*. 77: 453–479.

Doidge, C., G. Karolyi, and R. Stulz. 2009. “Has New York Become Less Competitive than London in Global Markets? Evaluating Foreign Listing Choices Over Time”. *Journal of Financial Economics*. 91: 253–277.

Doidge, C., G. Karolyi, and R. Stulz. 2010. “Why Do Foreign Firms Leave U.S. Equity Markets?” *Journal of Finance*. 65: 1507–1553.

Donelson, D. C., J. M. McInnis, and R. D. Mergenthaler. 2012. “Rules-Based Accounting Standards and Litigation”. *The Accounting Review*. 87(4): 1247–1279.

Duchin, R., J. Matsusaka, and O. Ozbas. 2010. “When Are Outside Directors Effective?” *Journal of Financial Economics*. 95: 195–214.

Duflo, E., R. Hanna, and S. Ryan. 2012. “Incentives Work: Getting Teachers to Come to School”. *American Economic Review*. 102: 1241–1278.

Dunning, T. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press.

Engel, E., R. Hayes, and X. Wang. 2007. “The Sarbanes–Oxley Act and Firms’ Going-Private Decisions”. *Journal of Accounting and Economics*. 44: 116–145.

Engel, E., R. Hayes, and X. Wang. 2010. “Audit Committee Compensation and the Demand for Monitoring of the Financial Reporting Process”. *Journal of Accounting and Economics*. 49: 136–154.

Faccio, M. and D. Parsley. 2009. “Sudden Deaths: Taking Stock of Geographic Ties”. *Journal of Financial and Quantitative Analysis*. 44: 683–718.

Fahlenbrach, R., A. Low, and R. Stulz. 2010. “Why Do Firms Appoint CEOs as Outside Directors”. *Journal of Financial Economics*. 97: 12–32.

Fama, E., L. Fisher, M. Jensen, and R. Roll. 1969. “The Adjustment of Stock Prices to New Information”. *International Economic Review*. 10: 1–21.

Fernandes, N., U. Lel, and D. Miller. 2010. "Escape from New York: The market impact of loosening disclosure requirements." *Journal of Financial Economics*. 95: 129–147.

Fich, E. and A. Shivdasani. 2007. "Financial fraud, director reputation, and shareholder wealth". *Journal of Financial Economics*. 86: 306–336.

Francis, B., I. Hasan, K. John, and L. Song. 2011. "Corporate Governance and Dividend Payout Policy: A Test Using Antitakeover Legislation". *Financial Management*. 37: 83–112.

Francis, B., I. Hasan, K. John, and M. Waisman. 2010. "The Effect of State Antitakeover Laws on the Firm's Bond Holders". *Journal of Financial Economics*. 96: 127–154.

Frangakis, C. and D. Rubin. 2002. "Principal Stratification in Causal Inference". *Biometrics*. 58: 21–29.

Frumento, P., F. Mealli, B. Pacini, and D. Rubin. 2012. "Evaluating the Effect of Training on Wages in the Presence of Noncompliance, Nonemployment, and Missing Outcome Data". *Journal of the American Statistical Association*. 107: 450–466.

Gagnon, L. and G. Karolyi. 2012. "The Economic Consequences of the U.S. Supreme Court's *Morrison v. National Australia Bank* Decision for Foreign Stocks Cross-Listed in U.S. Markets". Working paper. URL: http://ssrn.com/abstract=1961178.

Galiania, S., A. Murphy, and J. Pantano. 2015. "Estimating Neighborhood Choice Models: Lessons from the Moving to Opportunity Experiment". *American Economic Review*. 105: 3385–3415.

Gao, F., J. Wu, and J. Zimmerman. 2009. "Unintended Consequences of Granting Small Firms Exemptions from Securities Regulation: Evidence from the Sarbanes-Oxley Act". *Journal of Accounting Research*. 47: 459–506.

Gassen, J. 2014. "Causal Inference in Empirical Archival Financial Accounting Research". *Accounting, Organizations and Society*. 39: 535–544.

Gelbach, J., E. Helland, and J. Klick. 2013. "Valid Inference in Single-Firm, Single-Event Studies". *American Law and Economics Review*. 15: 495–541.

Gerber, A. and D. Green. 2012. *Field Experiments: Design, Analysis and Interpretation*. W. W. Norton & Company, Inc.

Ghosh, C., J. Harding, and B. Phani. 2008. "Does Liberalization Reduce Agency Costs? Evidence from the Indian Banking Sector". *Journal of Banking and Finance*. 32: 405–419.

Giannetti, M. and L. Laeven. 2009. "Pension Reform, Ownership Structure and Corporate Governance: Evidence from a Natural Experiment". *Review of Financial Studies*. 22: 4092–4127.

Gilson, R. and B. Black. 1995. *The Law and Finance of Corporate Acquisitions*. 2nd edition. Foundation Press.

Gippel, J., T. Smith, and Y. Zhu. 2015. "Endogeneity in Accounting and Finance Research: Natural Experiments as a State-of-the-Art Solution". *Abacus*. 51: 143–168.

Giroud, X. and H. Mueller. 2010. "Does Corporate Governance Matter in Competitive Industries?" *Journal of Financial Economics*. (2010) 95: 312–331.

Glaeser, E., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer. 2010. "Do Institutions Cause Growth?" *Journal of Economic Growth*. 9: 271–303.

Glaeser, E. L. 2006. "Researcher Incentives and Empirical Methods (2006). Harvard Institute of Economic Research Discussion Paper No. 2122". URL: https://ssrn.com/abstract=934557.

Goldman, E., J. Rocholl, and J. So. 2009. "Do Politically Connected Boards Affect Firm Value?" *Review of Financial Studies*. 22: 2331–2360.

Gomes, A., G. Gorton, and L. Madureira. 2007. "SEC Regulation Fair Disclosure, Information, and the Cost of Capital". *Journal of Corporate Finance*. 113: 300–334.

Gompers, P., J. Ishii, and A. Metrick. 2003. "Corporate Governance and Equity Prices". *Quarterly Journal of Economics*. 118: 107–155.

Greenstone, M., P. Oyer, and A. Vissing-Jorgensen. 2006. "Mandated Disclosure, Stock Returns, and the 1964 Securities Acts Amendments". *Quarterly Journal of Economics*. 121: 399–460.

Grullon, G., S. Michenaud, and J. Weston. 2015. "The Real Effects of Short-Selling Constraints". *Review of Financial Studies*. 28: 1737–1767.

Guner, B., U. Malmendier, and J. Tate. 2008. "Financial Expertise of Directors". *Journal of Financial Economics*. 88: 323–354.

Hail, L. and C. Leuz. 2009. "Cost of Capital Effects and Changes in Growth Expectations around U.S. Cross-Listings". *Journal of Financial Economics*. 93: 428–454.

Harvey, C., Y. Liu, and H. Zhu. 2016. ". . . and the Cross-Section of Expected Returns". *Review of Financial Studies*. 29: 5–28.

Heckman, J., H. Ichimura, and P. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme". *Review of Economic Studies*. 64: 605–654.

Hennessy, C. A. and I. A. Strebulaev. 2015. "Beyond Random Assignment: Credible Inference of Causal Effects in Dynamic Economies". Working paper. URL: http://ssrn.com/abstract=2564828.

Hirano, K. and G. Imbens. 2004. "The Propensity Score with Continuous Treatments, in Andrew Gelman and X.-L. Meng eds". *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*: 73–84.

Ho, D. and K. Imai. 2005. "Randomization Inference with Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election". *Journal of the American Statistical Association*. 101: 888–900.

Hochberg, Y. and L. Lindsey. 2010. “Incentives, Targeting and Firm Performance: An Analysis of Non-executive Stock Options”. *Review of Financial Studies*. 23: 4148–4186.

Hochberg, Y., P. Sapienza, and A. Vissing-Jorgensen. 2009. “A Lobbying Approach to Evaluating the Sarbanes-Oxley Act of 2002”. *Journal of Accounting Research*. 47: 519–583.

Holland, P. 1986. “Statistics and Causal Inference”. *Journal of the American Statistical Association*. 81: 945–960.

Hope, O. and W. Thomas. 2008. “Managerial Empire Building and Firm Disclosure”. *Journal of Accounting Research*. 46: 591–626.

Hosman, C., B. Hansen, and P. Holland. 2010. “The Sensitivity of Linear Regression Coefficients’ Confidence Limits to the Omission of a Confounder”. *Annals of Applied Statistics*. 4: 849–870.

Huang, J., Y. Shen, and Q. Sun. 2011. “Nonnegotiable Shares, Controlling Shareholders, and Dividend Payments in China”. *Journal of Corporate Finance*. 17: 122–133.

Iliev, P. 2010. “The Effect of SOX Section 404: Costs, Earnings Quality, and Stock Prices”. *Journal of Finance*. 65: 1163–1196.

Imai, K. and D. Van Dyk. 2004. “Causal Inference with General Treatment Regimes: Generalizing the Propensity Score”. *Journal of the American Statistical Association*. 99: 854–866.

Imbens, G. W. and J. M. Wooldridge. 2009. “Recent Developments in the Econometrics of Program Evaluation”. *Journal of Economic Literature*. 47: 5–86.

Imbens, G. and T. Lemieux. 2008. “Regression Discontinuity Designs: A Guide to Practice”. *Journal of Econometrics*. 142: 615–635.

Imbens, G. and P. Rosenbaum. 2005. “Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education”. *Journal of the Royal Statistical Society A*. 168: 109–126.

Imbens, G. and D. Rubin. 2015. “Causal Inference in Statistics, Social, and Biomedical Sciences”.

John, K. and L. Litov. 2010. “Managerial Entrenchment and Capital Structure: New Evidence”. *Journal of Empirical Legal Studies*. 7: 693–742.

Kamar, E., P. Karaca-Mandic, and E. Talley. 2008. “Going-Private Decisions and the Sarbanes-Oxley Act of 2002: A Cross-Country Analysis.” *Journal of Law, Economics, and Organization*. 25: 107–133.

Karpoff, J. M. and M. D. Wittry. 2015. “Test identification with legal changes: The case of state antitakeover laws”. Working paper. URL: http://ssrn.com/%20abstract=2493913.

Kezdi, G. 2004. “Robust Standard Error Estimation in Fixed-Effects Panel Models”. *Hungarian Statistical Review*. 9: 95–116.

Kinney, W. and M. Shepardson. 2011. “Do Control Effectiveness Disclosures Require SOX 404(b) Internal Control Audits? A Natural Experiment with Small U.S. Public Companies”. *Journal of Accounting Research*. 49: 413–448.

La Porta, R., F. Lopez-de-Silanes, and A. Shleifer. 2008. "The Economic Consequences of Legal Origins". *Journal of Economic Literature*. 46: 285–332.

Larcker, D., G. Ormazabal, and D. Taylor. 2011. "The Market Reaction to Corporate Governance Regulation". *Journal of Financial Economics*. 101: 431–448.

Larcker, D. and T. Rusticus. 2010. "On the Use of Instrumental Variables in Accounting Research". *Journal of Accounting and Economics*. 49: 186–205.

Leamer, E. 1978. "Specification Searches: Ad Hoc Inference with Nonexperimental Data".

Lee, D. and T. Lemieux. 2010. "Regression Discontinuity Designs in Economics". *Journal of Economic Literature*. 48: 281–355.

Lee, D. and K. Suh Park. 2009. "Does Institutional Activism Increase Shareholder Wealth? Evidence from Spillovers on Non-Target Companies". *Journal of Corporate Finance*. 15: 488–504.

Lemieux, T. and K. Milligan. 2008. "Incentive Effects of Social Assistance: A Regression Discontinuity Approach". *Journal of Econometrics*. 142: 807–828.

Lemmon, M. and K. Lins. 2003. "Ownership Structure, Corporate Governance, and Firm Value: Evidence from the East Asian Financial Crisis". *Journal of Finance*. 58: 1445–1468.

Lennox, C., J. Francis, and Z. Wang. 2012. "Selection Models in Accounting Research". *The Accounting Review*. 87: 589–616.

Li, H., M. Pincus, and S. Olhoft Rego. 2008. "Market Reaction to Events Surrounding the Sarbanes-Oxley Act of 2002 and Earnings Management". *Journal of Law and Economics*. 51: 111–134.

Lin, C. and D. Su. 2008. "Industrial Diversification, Partial Privatization, and Firm Valuation: Evidence from Publicly Listed Firms in China". *Journal of Corporate Finance*. 14: 405–417.

Litvak, K. 2007. "The Effect of the Sarbanes-Oxley Act on Non-US Companies Cross-Listed in the US". *Journal of Corporate Finance*. 13: 195–228.

Litvak, K. 2008. "Long-Term Effect of Sarbanes-Oxley on Cross-Listing Premia". *European Financial Management*. 14: 875–920.

Litvak, K. 2014. "The Effect of Jurisdictional Competition for Corporate Charters on Financial and Accounting Measures of Firm Performance". Working paper. URL: http://ssrn.com/abstract=1990805.

Lo, K. 2003. "Economic Consequences of Regulated Changes in Disclosure: the Case of Executive Compensation". *Journal of Accounting and Economics*. 35: 285–314.

Low, A. 2009. "Managerial Risk-Taking Behavior and Equity-Based Compensation". *Journal of Financial Economics*. 92: 470–490.

MacKinlay, A. 1997. "Event Studies in Economics and Finance". *Journal of Economic Literature*. 35: 13–39.

MacKinnon, J. G. and M. D. Webb. 2016. "Wild Bootstrap Inferences for Wildly Different Cluster Sizes". *Journal of Applied Econometrics*. Forthcoming.

Malani, A. and J. Reif. 2015. "Interpreting Pre-trends as Anticipation: Impact on Estimated Treatment Effects from Tort Reform". *Journal of Public Economics*. Forthcoming.

Malmendier, U. and G. Tate. 2009. "Superstar CEOs". *Quarterly Journal of Economics*. 124: 1593–1638.

McCrary, J. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test". *Journal of Econometrics*. 142: 698–714.

Mitchell, M. and J. Netter. 1989. "Triggering the 1987 Stock Market Crash: Antitakeover Provisions in the Proposed House Ways and Means Tax Bill?" *Journal of Financial Economics*. 24: 37–68.

Morck, R., D. Yavuz, and B. Yeung. 2011. "Banking System Control, Capital Allocation, and Economic Performance". *Journal of Financial Economics*. 100: 264–283.

Morgan, S. and C. Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd edition. Cambridge University Press.

Mulherin, H. 2007. "Measuring the costs and benefits of regulation: Conceptual issues in securities markets". *Journal of Corporate Finance*. 13: 421–437.

Muravyev, A. 2013. "Investor Protection and the Value of Shares: Evidence from Statutory Rules Governing Variations of Shareholders' Class Rights in an Emerging Market". *Journal of Law, Economics, and Organization*. 29: 1344–1383.

Murphy, K. and T. Sandino. 2010. "Executive Pay and "Independent" Compensation Consultants". *Journal of Accounting and Economics*. 49: 247–262.

Nagar, V., D. Nanda, and P. Wysocki. 2003. "Discretionary Disclosure and Stock-Based Incentives". *Journal of Accounting and Economics*. 34: 283–309.

Nasev, J., B. Black, and W. Kim. 2016. "How Does Corporate Governance Affect Firm Behavior? Shock-Based versus Panel Data Models". Working paper. URL: http://ssrn.com/abstract=2133283.

Nenova, T. 2005. "Control Values and Changes in Corporate Law in Brazil". *Latin American Business Review*. 6: 1–37.

Nguyen, B. D. and K. M. Nielsen. 2010. "The Value of Independent Directors: Evidence from Sudden Deaths". *Journal of Financial Economics*. 98: 550–567.

Oster, E. 2013. "Unobservable Selection and Coefficient Stability: Theory and Validation in Public Health". Working paper. URL: http://ssrn.com/%20abstract=2266720.

Paik, M., B. Black, and D. Hyman. 2013. "The Receding Tide of Medical Malpractice Litigation Part 2: Effect of Damage Caps". *Journal of Empirical Legal Studies*. 10: 639–670.

Paik, M., B. Black, and D. A. Hyman. 2016. "Damage Caps and the Labor Supply of Physicians: Evidence from the Third Reform Wave". *American Law and Economic Review*. Forthcoming.

Paik, M., B. Black, and D. A. Hyman. 2017. "Damage Caps and Defensive Medicine, Revisited". *Journal of Health Economics*. Forthcoming.

Perez-Gonzalez, F. 2006. "Inherited Control and Firm Performance". *American Economic Review*. 96: 1559–1588.

Petersen, M. 2009. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches". *Review of Financial Studies*. 22: 435–480.

Qiu, J. and F. Yu. 2009. "The Market for Corporate Control and the Cost of Debt". *Journal of Financial Economics*. 93: 505–524.

Rauh, J. 2006. "Own Company Stock in Defined Contribution Pension Plans: A Takeover Defense?" *Journal of Financial Economics*. 81: 379–410.

Roberts, M. and T. Whited. 2013. "Endogeneity in Empirical Corporate Finance, in George M. Constantinides, Milton Harris, and Rene M. Stulz eds." *Handbook of the Economics of Finance,* vol. 2A, 493–572.

Rosenbaum, P. R. 2010. *Design of Observational Studies*. Springer-Verlag New York.

Rosenzweig, M. and K. Wolpin. 2000. "Natural 'Natural Experiments' in Economics". *Journal of Economic Literature*. 38: 827–874.

Rubin, D. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". *Journal of Educational Psychology*. 66: 688–701.

Ryan Jr., H. and R. Wiggins III. 2004. "Who is in whose pocket? Director compensation, board independence, and barriers to effective monitoring". *Journal of Financial Economics*. 73: 497–524.

Salas, J. 2010. "Entrenchment, Governance, and the Stock Price Reaction to Sudden Executive Deaths". *Journal Banking and Finance*. 34: 656–666.

Shadish, W., T. Cook, and D. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Thomas Cook Publishing.

Shanken, J. and A. Tamayo. 2012. "Payout Yield, Risk, and Mispricing: A Bayesian Analysis". *Journal of Financial Economics*. 105: 131–152.

Small, D. and P. Rosenbaum. 2008. "War and Wages: The Strength of Instrumental Variables and their Sensitivity to Unobserved Biases". *Journal of the American Statistical Association*. 103: 924–933.

Smith, J. and P. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators". *Journal of Econometrics*. 125: 305–353.

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

Strebulaev, I. and T. Whited. 2013. "Dynamic Models and Structural Estimation in Corporate Finance". *Foundations and Trends in Finance*. 6: 1–163.

Stuart, T. and S. Yim. 2010. "Board Interlocks and the Propensity to Be Targeted in Private Equity Transactions". *Journal of Financial Economics*. 97: 174–189.

Subramanian, G. 2004. "The Disappearing Delaware Effect". *Journal of Law, Economics, and Organization*. 20: 32–59.

Thompson, S. 2010. "Simple Formulas for Standard Errors that Cluster by Both Firm and Time". *Journal of Financial Economics*. 99: 1–10.

Wang, X. 2010. "Increased Disclosure Requirements and Corporate Governance Decisions: Evidence from Chief Financial Officers in the Pre- and Post–Sarbanes-Oxley Periods". *Journal of Accounting Research*. 48: 885–920.

Welch, I. 2012. "A Critique of Recent Quantitative and Deep-Structure Modeling in Capital Structure Research and Beyond". *Critical Finance Review*. 2: 131–172.

Wintoki, M. B. 2007. "Corporate boards and regulation: The Effect of the Sarbanes–Oxley Act and the Exchange Listing Requirements on Firm Value". *Journal of Corporate Finance*. 13: 229–250.

Wooldridge, J. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd edition. MIT Press.

Wooldridge, J. 2012. *Introductory Econometrics: A Modern Approach*. 5th edition. Cencage Learning.

Zhang, I. X. 2007. "Economic Consequences of the Sarbanes–Oxley Act of 2002". *Journal of Accounting and Economics*. 44: 74–115.