Recognizing that cross-sectional data are often insufficient to address the identification problems associated with estimating the effect of government taxation or spending, economists engaged in public finance research often utilize longitudinal data that span the period over which a policy change occurred. As economic data have proliferated over the last decade, uses of the difference-in-differences design and its variations have become more numerous. Nevertheless, published research that invokes difference-in-differences commonly fails to present evidence and reasoning that enable the reader to properly evaluate the causal claims under investigation. In this paper, we examine the threats to internal validity that exist when using difference-in-differences for causal inference and review variations of the design that can be used to address these threats. Next, we survey the public finance literature in order to examine the ways that these threats are addressed in practice. We conclude by proposing a number of recommendations for researchers to consider as they implement difference-in-differences as an empirical strategy.

Keywords: difference-in-differences, interrupted time series, research design, causal inference

JEL Codes: C22, C23

I. INTRODUCTION

Recognizing that cross-sectional data are often insufficient to address the identification problems associated with estimating the effect of government taxation or spending, economists engaged in public finance research frequently utilize longitudinal data that span the period over which a policy change occurred. To cite one common example, marginal tax rates are a function of taxable income, and estimates of the effect of taxation on labor supply that fail to correctly account for income effects will be biased (Triest, 1998). As a result, researchers often look for sources of variation, in the tax code and...
elsewhere, to help resolve the identification problem. If there is a change in marginal tax rates (as a result of tax reform, for example), as well as a suitable comparison group, the policy change can be analyzed by examining data from before and after the intervention: the difference-in-differences design.

As financial and administrative data have proliferated over the last decade, uses of the difference-in-differences (D-in-D) design have become more numerous. Moreover, researchers have increasingly used extensions of the traditional D-in-D design, such as the triple-difference estimator, to add robustness to the basic D-in-D design. Nevertheless, published research that invokes the D-in-D design commonly fails to present evidence and reasoning that enable the reader to properly evaluate the causal claims under investigation. The purpose of this paper is to call attention to challenges in the use of the D-in-D design as it is applied to work in public finance and to provide a set of best practices for researchers to consider as they make use of difference-in-differences as an identification strategy.

Although a body of research exists concerning the econometric issues involved with difference-in-differences estimation, many of these papers deal with technical issues that are narrow in scope. Bertrand, Duflo, and Mullianathan (2004), Donald and Lang (2007), and Conley and Taber (2011) focus on the correct estimation of standard errors rather than the bias of the estimates. Abadie (2005) demonstrates a strategy for estimating the average effect of the treatment on the treated when non-parallel outcome dynamics exist for the treated and untreated groups. Athey and Imbens (2006) propose a D-in-D estimator that allows for treatment effect heterogeneity and provide conditions under which the model is identified nonparametrically. Those writing more generally about inference with finance panel data sets have concerned themselves with approaches to standard error estimation (Petersen, 2009) or the choice of estimator (Skoulakis, 2008).

Although we touch on some of these econometric concerns in this paper, we focus instead on broader questions of research design and attempt to synthesize the recent literature in order to provide a roadmap to best practices. Specifically, we examine the threats to internal validity that exist when using the difference-in-differences design for the purpose of causal inference and variations of the design that can be used to address these threats, with particular emphasis on questions relevant to public finance. In doing so, we draw on the Cook/Campbell research tradition, in which valid causal estimates depend on researchers being able to rule out alternative explanations through the use of design elements.

The rest of this paper proceeds as follows. First, we review various methods related to difference-in-differences and discuss the threats to validity that exist under each and how they differ. We call attention to the strengths and weaknesses of each approach by discussing a noteworthy application. Next, we survey the public finance literature in order to examine the ways in which these threats are addressed in practice. The concluding section proposes a number of recommendations for researchers to consider as they implement difference-in-differences as an empirical strategy.
II. OVERVIEW OF THE DIFFERENCE-IN-DIFFERENCES DESIGN: DESCRIPTION AND THREATS TO INTERNAL VALIDITY

The primary focus of the tradition begun by Campbell (1957) and extended by Campbell and Stanley (1963) and Cook and Campbell (1979) is how to use experimental design to improve causal inference. Under this tradition, researchers making causal claims must be able to rule out a specific set of plausible alternative explanations, or “threats to internal validity.” The difference-in-differences design with many time points is itself a more sophisticated version of an interrupted time series, in which a treatment group (without a comparison time series) is analyzed before and after the treatment occurs. Accordingly, for the sake of exposition we first review the threats to internal validity that exist in the case of a simple interrupted time series design: history, selection, instrumentation, regression, and maturation (Shadish, Cook, and Campbell, 2002).

A. Simple Interrupted Time Series

*History* refers to the possibility that confounding events occurred during the same time interval as the policy change under study. This is a frequent danger in public finance research as broader trends in the economy are likely to impact income levels, and causal estimates that fail to take account of these trends will be biased. *Selection* bias can occur if the composition of the treatment group shifts in anticipation of the change in policy. For instance, high-wage earners may shift the nature of their compensation in anticipation of a tax change. This may often occur in the case of aggregate data where it is not possible to identify the exact composition of the treatment group. *Regression* to the mean occurs when the treatment group is composed primarily of subjects with extreme values of the outcome measure, such as earners in the top 1 percent of the income distribution. As these values will tend to become less extreme in subsequent measurements, regressing on time series data may lead to spurious estimates of a treatment effect. *Instrumentation* refers to changes in the measurement of the outcome, a threat that is typically of less concern to public finance researchers as income and revenue measures are reasonably well-defined and consistent. *Maturation* refers to changes that occur naturally over time even in the absence of treatment.

Because of the difficulty of ruling out all of these possible threats, in particular that historical developments other than the measure of interest will influence the time series, the simple interrupted time series (ITS) design is generally considered to be unreliable for the purpose of causal estimation (Shadish, Cook, and Campbell, 2002). The one exception is when treatment effects are very large relative to prior intertemporal variation in the dependent variable and are exhibited instantaneously rather than with a delay. This is typically the only instance in which a researcher can make the strong case using only a one-group design that the change in the outcome measure is related to the intervention under study.
Chetty and Saez (2005) provide an example of a one-group design in a study on the effect of the 2003 tax cut on dividend payments. Using quarterly data spanning the years 1980–2004, they document a surge in dividend initiations immediately following the reform as well as an increase in regular dividend payments. Figure 1, taken from their paper, shows the effect of the reform, separately, on total regular dividends, regular dividend payments by the largest 20 regular dividend payers, and special dividends. Regular dividends are periodic and recurring payments, whereas special dividends are one-time and non-recurring. Both would be expected to be affected by the reform, and hence all three lines would be expected to increase following the tax cut. By decomposing the treatment group into separate categories, the authors give a sense for the relative size of regular vs. special dividends and the magnitude of the reform’s effect on each.

However, the authors do not rely on the one-group design alone; even with a long time series and a relatively clear break (especially in the number of regular dividend initiations, which they show separately), their results rest on a strong assumption — that no other event apart from the tax cut caused the surge in dividend payments. In order to address this threat, they supplement their simple ITS design with a comparison-group design, using firms owned primarily by nontaxable institutions as a comparison group. The addition of the comparison-group design enables them to more credibly address the threat of history than the use of the simple one-group design alone.

![Figure 1](image-url)

*Figure 1*

A Simple Interrupted Time Series Design

Source: Reproduced from Chetty and Saez (2005), Figure I, p. 801
B. Difference-in-Differences

Because of the strong identifying assumptions required by the simple ITS design, researchers typically supplement the treatment group with a comparison time series: the difference-in-differences design. The presence of a comparison group changes the threats to validity that a researcher must rule out. Whereas in the case of the simple ITS design, the set of alternative explanations includes any event that acted on the treatment group during the same time period as the intervention under study, now the threats include history, selection, regression, instrumentation, and maturation as they act differentially on the treatment and comparison groups over time. This set of threats is presumably smaller and thus easier to rule out than the unconditional threats that operate under the simple ITS design. However, the choice of comparison group is non-trivial and should be subject to as many of the same economic conditions as the treatment group as possible. Lemieux and Milligan (2008) show that D-in-D estimators perform well when the comparison group is in the same labor market as the treatment group, but that estimates can diverge greatly otherwise. The most basic version of the difference-in-differences design is

\[
Y = \alpha + \beta_1 \text{Treatment} + \beta_2 \text{Post} + \beta_3 \text{Treatment} \times \text{Post} + \xi X + \varepsilon,
\]

where Treatment represents individuals or groups (states, for example) that received an intervention, Post is a dummy variable equal to 1 in the time period following an intervention, and \( \beta_3 \) indicates the estimated average treatment effect. Because it compares only a difference in means, this basic difference-in-differences design requires a minimum of only two time points: one prior to the intervention under study, and one after. Additional time points prior to the intervention may be useful in obtaining a more precise measure of the pre-intervention mean of the outcome variable. Additional post-test time points may similarly assist in assessing the post-intervention mean or estimating the treatment effects at different points in time, obtained through the use of multiple dummy variables for different years of the intervention. As noted above, the threats to validity remain the same as those in the simple ITS design, but only insofar as they operate differentially on the treatment and comparison groups over the time period spanning the intervention, commonly referred to as the “parallel-paths” or “parallel-trends” assumption (Meyer, 1995). The parallel trends assumption is more likely to be violated when the dependent variable is one that exhibits a high degree of volatility, and researchers should draw on past research to call attention to the stability of the variable.

If researchers do use multiple time points to estimate the pre- and post-intervention means, then they must take care to account for auto-correlation in calculating the standard errors; this has been one of the main critiques leveled at the difference-in-differences design (Bertrand, Duflo, and Mullianathan, 2004), and the issue of serial correlation in data with a group structure remains an area of study. However, as Angrist and Pischke (2009) point out, this problem is now typically addressed through higher-level clustering.
Finkelstein (2002) provides an illustrative example of a difference-in-differences design that uses only two time points. Examining the effect of a 1993 tax change that reduced the subsidy to employer-provided supplementary health insurance in Quebec, Finkelstein uses a D-in-D methodology to compare the change in employer-provided supplementary health coverage in Quebec to changes in health coverage in other Canadian provinces not affected by the law. The study uses two years of repeated cross-sections formed from the 1991 and 1994 Canadian General Social Surveys (GSS) and a binary dependent variable indicating whether an individual received supplementary health insurance from his or her employer. To account for the possibility that compositional changes in the treatment group relative to the comparison group could be driving the results, Finkelstein includes a set of covariates in some of her specifications and also presents descriptive data to illuminate any differences between Quebec and the set of comparison provinces. Finkelstein writes that, “With the covariates included in the regression, the identifying assumption is that, conditional on X, there is no Quebec-specific time trend in coverage by employer-provided supplementary health insurance” (Finkelstein, 2002, p. 315).

Finkelstein uses two falsification exercises to examine the validity of this assumption: (1) a pretreatment specification test, and (2) dependent variables not subject to the treatment, sometimes referred to as non-equivalent dependent variables. With few time points, placebo/falsification tests represent a way for researchers to demonstrate the validity of the parallel trends assumption and credibly address threats to internal validity. Using data from 1989, Finkelstein re-estimates her regression using 1989 as the “before” period and 1991 as the “after” period. The presumption is that the regression will not yield significant results, and thus confirm that the two groups — Quebec and all other provinces — followed parallel trends prior to the change in the law. As expected, she reports coefficients that are close to zero.

As a second falsification exercise, Finkelstein re-estimates her main equation on two variables that were presumably not affected by the change in the law: coverage by employer-provided retirement pensions and coverage by employer-provided paid parental leave. These two variables are conceptually similar to the outcome of interest — coverage by employer-provided supplementary health insurance — and thus in theory provide evidence of how health insurance coverage would have trended in the absence of the law. Hence, this test also addresses the concern that a shock specific to the treatment group — in this case, workers in Quebec — was responsible for the drop in coverage rather than the law. Indeed, Finkelstein again reports coefficients on the variable of interest that are close to zero and not statistically significant. She further demonstrates this result by graphing the trend lines of one of these variables, illustrating the clear discrepancy between the outcome that was affected by the law and the other that was not.

This second falsification exercise illustrates the use of non-equivalent dependent variables. A non-equivalent dependent variable time series is one that is theoretically subject to all known threats to validity except for the treatment and can thus serve as a placebo test. By using the same empirical model to test whether the treatment has
an effect on a distinct outcome measure, researchers can examine pertinent threats to validity, in particular history. Unlike falsification exercises that involve manipulation of the time period under analysis or the sample under study, a non-equivalent dependent variable uses the sample treatment and comparison groups as well as the same time period as the main D-in-D specification and so presumably represents a strong counterfactual (Shadish, Cook, and Campbell, 2002). Nevertheless, strong assumptions are also required; in order for a non-equivalent dependent variable to be valid for the purpose of ruling out explanations, it must be conceptually related to the primary dependent variable. If the two variables are not closely aligned, then they will not necessarily be subject to the same threats, and the non-equivalent time series becomes meaningless.

C. Difference-in-Difference-in-Differences

A further variation on the difference-in-differences design is the difference-in-difference-in-differences (D-in-D-in-D) design, or the triple difference estimator. This approach rests on the premise that an additional source of variation can help to better isolate and identify the effects of a program. This is typically accomplished by identifying a second comparison group that differs from the first comparison group along a particular dimension, thereby helping to rule out threats to validity that the initial comparison alone cannot. We have confined our discussion so far to longitudinal D-in-D designs in which time is one of the dimensions of comparison. However, this need not be the case, and D-in-D-in-D designs illustrate that time may be one of several useful sources of variation. The triple difference design can be estimated as

\[
Y = \alpha + \beta_1 \text{Treatment} + \beta_2 \text{Post} + \beta_3 \text{Treatstate} + \beta_4 \text{Treatment} \times \text{Post} + \beta_5 \text{Treatment} \times \text{Treatstate} + \beta_6 \text{Treatment} \times \text{Treatstate} \times \text{Post} + \beta_7 \text{Treatment} \times \text{Treatstate} \times \text{Post} + \xi X + \varepsilon,
\]

where \text{Treatstate} in this case indicates, for instance, the state in which a policy change occurred, and \text{Treatment} indicates the treatment group within the state where the change occurred. In this case, \(\beta_7\) is equal to the difference in estimated impact on treatment versus control individuals in the state where the policy change occurred relative to the difference observed in individuals that are comparable to the treatment and control units in a nearby comparison state. The triple difference design is particularly useful in addressing selection bias; it accounts for trends that affect all individuals that are similar to the treatment group in the same way. If the out-of-state “treatment” individuals are also affected by these trends, then the spurious effect will be differenced out.

Cancian and Levinson (2006) provide an illustrative example. They examine the effect of the earned income tax credit (EITC) on labor supply by measuring the employment outcomes of single mothers in Wisconsin, which provides a generous supplement to the federal EITC for families with three children. Because the supplemental benefit was much larger for families with three children than families with two children, they
are able to exploit three sources of variation: (1) single mothers with three children in Wisconsin relative to single mothers in Wisconsin with two children, (2) single mothers with three children in Wisconsin relative to single mothers with three children in states that do not supplement the federal EITC, and (3) single mothers with three children in Wisconsin before and after the change to the supplement rate. Thus, the authors are able to isolate an additional source of variation because of the specificity of the treatment assignment; single mothers with three children received a notably larger benefit than single mothers with two children, a group that otherwise would most likely be fairly comparable.

However, the additional comparison group does not remove the possibility of shocks that affect only the treatment group in the treatment states in the post-intervention time period. That is, with triple difference designs researchers must still address the possibility of trends that differentially affect the treatment group relative to both comparison groups and thereby bias measurements of the treatment effect.

D. Comparative Interrupted Time Series

One way to address this possibility is to model the unobserved bias using trend variables. In fact, although some economists refer to all two-group designs as D-in-D designs, others have distinguished between two types of D-in-D designs: (1) one that represents a comparison of pre-and post-intervention means, i.e., the difference between the pre- and post- comparison in the treatment group and the pre- and post- comparison in the comparison group, and (2) one that examines not just the pre- and post-interventions means in the outcome measure, but also introduces time trends in order to account for the differences in the slopes of the two time series. Somers et al. (2012) and St. Clair, Cook, and Hallberg (2014) distinguish this type of design from the difference-in-differences design by referring to it as the comparative interrupted time series (CITS) design. This design has several advantages over the more simple two-group difference-in-differences design. First, it alters the threats to internal validity such that they are presumably easier to rule out. No longer are the threats history, selection, instrumentation, regression, and maturation as they operate differentially across the treatment and comparison groups. Now the concern is that these threats operate differentially such that they are not adequately captured by time trends, linear or otherwise. For example, researchers might add two terms to the more simple difference-in-differences model: one term for the pre-treatment trend and another for the interaction between treatment group and the pre-treatment trend, or

\[ Y = \alpha + \beta_1 \text{Treatment} + \beta_2 \text{Post} + \beta_3 \text{Time} + \beta_4 \text{Treatment} \times \text{Time} + \beta_5 \text{Treatment} \times \text{Post} + \xi X + \varepsilon, \]

where $\text{Time}$ represents the pre-treatment trend, and $\text{Treatment} \times \text{Time}$ accounts for the differences between treatment and comparison groups that change linearly over time prior to the intervention. Equation (3) assumes that there is only one post-intervention time point, but additional trend variables can be introduced to differentiate between
the long-term and short-term effects of a policy or to account for trends that affected only specific groups.

However, this design comes with disadvantages as well. The use of trend variables imposes more stringent data requirements — data must be available for at least three time points prior to the intervention to estimate the baseline mean and trend — and requires assumptions about the parametric form of the time-series process. With more time points, it is of course possible for researchers to more directly examine the functional form of the pre-test trend and test a wider range of assumptions, either by modeling the time series non-parametrically or using non-linear functions.

Acemoglu and Angrist (2001) illustrate the examination of pre-treatment trends in a study measuring the effect of the Americans with Disabilities Act (ADA) on the employment and earnings of the disabled. The authors had 10 years of data from the Current Population Survey (CPS), five years before the law came into effect (1987–1991), one year they regarded as a “transition” year (1992), and four post-treatment years (1993–1996). They used the non-disabled as a comparison group for the disabled.

Figure 2 reproduces one diagram from their paper. The 10 years of data enable Acemoglu and Angrist to clearly demonstrate the pre- and post-treatment employment trends of their treatment and comparison groups and give a sense of the magnitude of

---

**Figure 2**

Visual Inspection of the Parallel Trends Assumption

![Graph showing weeks worked vs. year for non-disabled and disabled men and women](source: Reproduced from Acemoglu and Angrist (2001), Figure 2, p. 930)
the effect. The graph also enables them to visualize the functional form of the group trends and inspect whether or not the trends move in parallel. The graph (illustrating workers aged 40–58, with younger workers illustrated separately) shows a slight decline in the hours worked of disabled men relative to nondisabled men in 1993. Prior to the enactment of the law in 1992, the two trends moved roughly in parallel. (There is no visible effect for women in the age range, with the lines representing disabled and nondisabled both rising slightly in 1993 and 1994.)

Because they have five years of pretreatment data, the authors are able to test this assumption by interacting treatment status (in this case, disability) with a linear trend. None of the coefficients on this term are statistically significant, a result that confirms what is evident from the figure: the difference between the lines is reasonably stable. Having 10 years of data also enables the authors to interact year dummies with treatment status and thus estimate the treatment effect separately by year. The pre-treatment years serve as pre-treatment specification tests; as in Finkelstein’s study, they serve as placebo tests to show that no “treatment” effect was evident in the years prior to the enactment of the law. The post-treatment interactions demonstrate whether the effect varied over time.

Although Acemoglu and Angrist (2001) examine the possibility that changes in their dependent variable can be explained by the extrapolation of differential trends in their treatment and comparison time series, it is also possible to test for the existence of a trend in the treatment effect, something that Gorodnichenko, Martinez-Vasquez, and Peter (2009) do in their study of the Russian flat tax. Figure 3 reproduces a figure from their study, illustrating the difference in linear trends between the treatment and comparison groups in the post-reform period. Although the fit of the lines appears suspect, there is some evidence of the treatment group decreasing relative to the comparison group post-treatment.

E. Matching

Another approach to reducing bias in the D-in-D design is through the use of matching. Matching procedures pair treatment and comparison units on baseline characteristics to create comparison groups that are as similar to the treatment group in observables as they would be under randomization. The goal is to reduce selection bias relative to the design that simply uses a comparison group of convenience. Although only a few studies specifically examine matching with longitudinal data (St. Clair, Cook, and Hallberg, 2014; Cook, Shadish, and Wong, 2008), these studies as well as prior work on the topic suggest that matching on pretest measures of the outcome and on geographic proximity can yield reliable effect estimates, though it is unclear whether matching alone offers significant advantages relative to standard regression methods (i.e., with covariates but without matching adjustments) (Glazerman, Levy, and Myers 2003; Smith and Todd, 2005; Michalopolus, Bloom, and Hill, 2004). And though various statistical procedures for matching exist, the selection of covariates and their reliable measurement is thought to be more important than the choice of matching strategy (Steiner and Cook, 2013).
Where matching does offer advantages is as a means of pre-processing data, in which case much of the adjustment for control variables is non-parametric, producing causal estimates that are less model-dependent than the standard parametric statistical procedures (Ho et al., 2007). This is particularly relevant in situations when the distributions of pre-treatment covariates differ across the treatment and comparison groups or fail to overlap. On the other hand, matching may be inefficient if it discards observations that are within the range of overlap (Gelman and Hill, 2007).

Insofar as the choice of comparison groups is often done informally and without sufficient consideration of what the treatment group would have looked like in the absence of a policy intervention, matching represents a more coherent strategy for selecting a comparison group. Matching may be particularly useful with time-series data when the functional form of the time-series is difficult to model. A related approach is to use synthetic case control, a method of estimating treatment effects when the units are large aggregates like states. By using weighted averages of potential comparison units to create synthetic comparison units, researchers can ensure that pre-treatment trends for the treatment state and its synthetic control are similar (Abadie, Diamond, and Hainmueller, 2010).
In his examination of a Danish credit market reform, Leth-Peterson (2010) compares two groups of households, one of which was likely to be affected by the reform (low liquid asset households) and another that was not (high liquid asset households). As with all studies that compare two groups from along a distribution, it is not clear that low liquid asset households represent a valid counterfactual for high liquid asset households. Recognizing that “regression estimators rely on extrapolation in areas with no overlap in the covariate distribution between low liquid asset households and high liquid asset households,” he uses propensity score matching in order to balance the covariates between the two groups (Leth-Peterson, 2010, p. 1088).

In order to demonstrate that matching is successful in making the groups more comparable, he presents t-tests for all the explanatory variables in the propensity score model, showing that there is no evidence of differences between the two groups in observables. Although matching does not reduce the need to address other threats to internal validity, such as events other than the treatment that may have influenced borrowing by the two groups, it presumably reduces the sensitivity of the regression model to misspecification. Furthermore, Leth-Peterson takes advantage of the individual estimates that matching produces (from each matched pairing) in order to explore the heterogeneity of treatment responses; by regressing nonparametrically these individual estimates of the reform against the ages of the oldest members of the household, he is able to show that the effect of the reform was strongest among younger households.

Other variations on the difference-in-differences design exist in addition to those discussed above. For reasons of exposition, we have focused on cases where the treatment: (1) affects all units in the treatment group at a single moment in time; (2) is represented by a continuous variable; and (3) is not removed or re-introduced. However, the logic of the design can easily be extended along any one of these three dimensions.

By pooling a number of different treatment-comparison groups across different time periods, researchers can expand the number of cases in their analysis. In doing so, however, the threats to internal validity become more numerous and difficult to rule out; the assumption that all untreated units represent a valid counterfactual for all treated units is particularly difficult to examine. Moreover, researchers must differentiate between the immediate impact of a treatment and the potentially cumulative effect of treatment histories (Blackwell and Glynn, 2014). The use of time-series cross-sectional designs for causal inference remains an ongoing area of study.

The treatment variable can also be continuous; there is no reason why Leth-Peterson had to divide his sample into only two groups. It would have been just as possible to use a continuous variable representing the level of liquid assets per household prior to the policy change. The assumption then is that the magnitude of treatment in fact varies linearly with the level of liquid assets. This represents a stricter assumption than assuming that one group was constrained prior to the policy change while another group was not, and explains why it can sometimes be useful to form groupings according to the relative intensity of treatment as a complement to using a continuous measure. Card (1992) illustrates such an approach. In his study on the effect of the federal minimum wage, Card exploits the fact that the effect of the policy change varied across states based
on the fraction of workers who previously earned less than the new minimum. Hence, in one set of regressions, he treats this fraction as a continuous treatment variable, and examines employment and wages in all states before and after the increase in the federal minimum wage. However, in a separate preliminary analysis, he groups states into “low-wage,” “medium-wage,” and “high-wage” states based on their fraction of low-wage workers, demonstrating clearly that the wage increase affected “low-wage” states more than “medium-wage” states, and “medium-wage” states more than “high-wage states”.

Finally, in situations where the researcher can more actively manipulate the independent variable (i.e., the treatment), as often occurs in medical or education research, it is possible to remove treatment or stagger the introduction of the treatment across multiple points of time (Kratochwill et al., 2010). However, as these so-called “single-case designs” require researcher-control of the treatment, they are more rare in the public finance literature.

III. A SURVEY OF THE LITERATURE

In this section, we describe the use of the D-in-D design in public finance research and discuss the various forms in which the design appears. We analyze articles from five leading journals — The American Economic Review, The Quarterly Journal of Economics, The Journal of Political Economy, the Journal of Public Economics, and the National Tax Journal — over the period 2000–2012. We limit our search to articles that utilize a D-in-D design where the independent variable of interest is a binary variable representing a policy change and the dependent variable is an economic outcome. As noted in the previous section, the term “difference-in-differences” is sometimes used to refer to a wider array of designs than we have discussed here, including instances where the treatment variable is continuous, or to more general longitudinal designs where policy interventions occur in different regions at different points in time. For the sake of clarity, and in order to ensure consistency across our sample, we confine our search to the basic D-in-D design. Our search terms included “difference in difference,” “differences in differences,” and “interrupted time series.” This search yielded 60 articles across the five journals.¹

We classified the articles along seven different categories: (1) “Simple ITS,” (2) “Random Assignment,” (3) “Triple Difference,” (4) “Differential Trends,” (5) “Non-Equivalent DV,” (6) “Placebo/Falsification Test,” and (7) “Matching.”² The “Simple ITS” category indicates that the article performed at least one test without a comparison group — the simple interrupted time series design — that was separate from the main D-in-D specification. “Random Assignment” indicates that treatment status was

¹ An online appendix providing a full list of the articles is available at http://faculty.publicpolicy.umd.edu/tstclair/publications.
² In order to bolster the reliability of our findings, a research assistant performed a similar content analysis of a random subset of the articles on our list (20 out of 60 articles, or a third of the sample). His findings were entirely consistent with those of the first author.
randomly assigned. This is obviously a rare occurrence with policy interventions and only occurred on five occasions, most of which involved large-scale field experiments that were conducted with the blessing of state or federal tax authorities. The category “Triple Difference” indicates that a triple difference design was used, as in (2). We found only eight instances of the triple difference estimator. The rarity of this design is perhaps not surprising as it requires two unique sources of variation in treatment status.

“Differential Trends” indicates that the authors included in their analysis separate time trends for the treatment and comparison groups (typically accomplished through an interaction term between treatment status and a time trend, as in (3) or attempted to model in some way the pre-treatment differences between the two groups. Because this requires at least three pre-intervention time points, not all of the articles surveyed could have implemented such an approach, even if they had wanted to. This may explain in part why so few (3 out of 60) introduced differential trends into their regression analysis.

The “Non-Equivalent DV” category includes articles that used a non-equivalent dependent variable as a robustness check. “Placebo/Falsification Test” includes articles that used some other form of placebo test or falsification test. The two categories are similar in concept; both enable researchers to test whether the observed outcome was due to the treatment under study and not some alternative explanation. However, whereas a non-equivalent dependent variable uses the same treatment and comparison groups as the main analysis, a placebo/falsification test involves manipulating either the treatment sample or the time period of the analysis. While there may not always exist plausible non-equivalent dependent variables, it should be possible for most authors to conduct some form of falsification exercise.

The “Matching” category indicates that researchers limited the comparison group by matching treatment and comparison cases on observables. We found only six articles that utilized matching in the D-in-D context.

Table 1 demonstrates several noteworthy patterns. Only three articles attempted to model the pre-treatment differences between the treatment and comparison group. As noted above, this finding may be explained in part by data limitations. Nonetheless, where data are available, it should be fairly simple to add additional terms to a pre-existing empirical specification, if only for the purpose of a robustness check, and the low number suggests that this practice is infrequent. While a non-equivalent dependent variable may also not always be available, and its justification is not necessarily intuitive, the infrequent use (three out of 60 articles) is similarly grounds for concern. On the other hand, it appears that other types of placebo tests are more common. Overall, however, very few articles used supplemental design features, and as a result we conclude that there is much room for improvement in the use of strategies to strengthen causal inference with the D-in-D design.

IV. A SET OF RECOMMENDATIONS

With this concern in mind as well as the threats to validity discussed in Section II, we present a set of recommendations for researchers to consider as they employ difference-
in-differences as an identification strategy. As noted above, these recommendations apply to the proper choice of research design as a means of obtaining unbiased estimates. These recommendations are meant as a complement to, and not a substitute for, prior work on the precision of difference-in-differences estimates, such as that contained in Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2007), and Conley and Taber (2011).

1. **Examine graphically the parallel trends assumption and, if possible, test for a statistical interaction.** This is of course easier when more time points are available. But regardless of the number of pre-treatment time points, researchers should provide a plot of the temporal variation in the data so as to give the reader a sense of the size of the treatment effect relative to the prior intertemporal variance. With three or more pretreatment time points, researchers should interact a time trend — linear or otherwise — with the treatment dummy as a specification test.

2. **Use falsification tests to rule out threats to validity.** These tests should include a placebo test as well as a non-equivalent dependent variable if possible. Placebo tests can be conducted using either a treatment sample that was not in fact subject to the treatment or by conducting the D-in-D on the pretreatment time period. A non-equivalent dependent variable is subject to all of the same conditions as the treatment and therefore represents a strong check of a model’s assumptions, including the possibility of shocks that differentially affect the treatment and comparison groups prior to treatment. Nevertheless,

<table>
<thead>
<tr>
<th></th>
<th>AER</th>
<th>QJE</th>
<th>JPE</th>
<th>JPubEcon</th>
<th>NTJ</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple ITS</td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Random assignment</td>
<td>2</td>
<td></td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Triple difference</td>
<td>3</td>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Differential trends</td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Non-equivalent DV</td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Placebo/falsification test</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Matching</td>
<td>1</td>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Number of articles examined</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>31</td>
<td>10</td>
<td>60</td>
</tr>
</tbody>
</table>

Notes: Table 1 summarizes articles from *The American Economic Review, The Quarterly Journal of Economics, The Journal of Political Economy, the Journal of Public Economics,* and the *National Tax Journal.* The articles all include a difference-in-differences design in which the independent variable of interest is a policy change represented by a binary variable and the dependent variable is an economic outcome.
the assumptions that undergird its validity may be more difficult to meet; the variable must be conceptually similar to the true dependent variable for the test to hold.

3. **With data in the form of repeated cross-sections, probe for compositional change.** Otherwise, researchers leave themselves open to the criticism that some individuals self-selected into or out of treatment. This criticism has been levied frequently at studies that group observations according to wage income for the purpose of estimating the effect of tax reform (Triest, 1998). For instance, Heckman (1996) criticizes Eissa (1996) for grouping observations according to an income variable in her study of the Economic Recovery Tax Act of 1981, arguing that some members will switch groups as a result of the tax reform, leading to biased estimates of its impact. In the case that data limitations make it difficult to probe the composition of the groups, researchers should run specification checks using alternative definitions of treatment status, as in LaLumia (2009).

4. **Justify the choice of comparison group.** Even apart from the compositional shifts that may occur as a result of the policy change, there may be reason to believe that the comparison group in a study is not subject to the same economic conditions as the treatment group, complicating identification even when panel data are available. From a research design perspective, many of the challenges involved in estimating the effect of “endogenous policies” come down to the appropriate selection of a comparison group. Policies may be brought about by a specific set of economic conditions, which may have effects on the outcome of interest that are independent of the policy change (Besley and Case, 2000). Moreover, there may be reason to believe that the comparison group in a study will respond differentially to changes in economic conditions than the treatment group.

As Besley and Case point out, workers in one part of the earnings distribution are often used as controls for workers in another part of the distribution that are directly affected by a policy change, such as when a change to a benefit cap affects only higher wage workers. However, low skilled workers respond differently to changes in economic conditions than more highly skilled workers do. If an economic shock was the precipitating event for a change in the law, then there may be real reason to believe that workers at one end of the income distribution are not in fact a valid comparison group for workers at the other end.

When grouping treatment and comparison units according to where they fall along a distribution, researchers must also be concerned about regression to the mean. If one focuses on a panel of high-income individuals, on average their income will decrease the following year, while the income in the low-income group will increase. This concern particularly bedevils estimates of the elasticity of taxable income (ETI) where high-income taxpayers are often
implicitly compared to lower-income taxpayers using continuous treatment D-in-D designs. Weber (2014) derives the conditions under which a consistent estimate of the ETI is identified, and proposes an instrument that provides consistent estimates under certain reasonable assumptions.

5. **Where the choice of comparison group is unclear, matching can present a useful alternative.** The choice of a comparison group may be unclear either due to an ambiguous functional form in the pretreatment trend, lack of balance in the distribution of covariates between the treatment and comparison groups, or lack of overlap. Matching may be particularly useful in cases where researchers are reluctant to impose a functional form on time-series data. As noted above, the statistical process of matching is less important than the covariates that are used; ideally the covariates used for matching should capture the selection differences between the treatment and comparison groups.

6. **Where an additional source of variation can be found, a triple-difference-estimator can provide a useful identification strategy.** However, recommendation (4) above still holds; researchers must justify the choice of comparison groups and explain how the additional source of variation helps to rule out alternative explanations of the treatment effect. In most cases, triple-difference estimates should be presented alongside regular D-in-D estimates as a robustness check rather than as a main result given their strong assumptions, namely that two sets of comparison groups are both appropriate as counterfactuals.

7. **Be careful not to condition on post-treatment outcomes.** Although we have focused primarily on design strategies rather than modeling choices, covariates are frequently employed in D-in-D estimation to account for selection or to increase precision. While providing useful information in some cases (Glynn and Kashin, 2014), post-treatment covariates, i.e., variables that are also subject to change as a result of the treatment, should ordinarily not be included in an analysis lest they block the effect of interest (Gelman and Hill, 2007).

Although there are undoubtedly other considerations to bear in mind as well, we present these seven recommendations as the most crucial for causal inference with difference-in-differences. Although not every analysis will be able to follow all of these recommendations, we believe that this list offers a roadmap toward improvements in the quality of research that is produced. Difference-in-differences has been and will remain a useful estimation strategy, particularly as more and more economic data become widely available. When executed properly, this common-sense design represents an extremely useful tool in the arsenal of applied researchers attempting to make causal claims from observational data. Nevertheless, as with any quasi-experimental design, its efficacy depends on the clarity and transparency with which the results are presented. The most convincing uses of the design are those that demonstrate, through judicious use of design elements, that the treatment is the only valid explanation for the obtained results.
DISCLOSURES

The authors have no financial arrangements that might give rise to conflicts of interest with respect to the research reported in this paper.

REFERENCES


