INTERNAL AND EXTERNAL VALIDITY OF THE COMPARATIVE INTERRUPTED TIME-SERIES DESIGN: A META-ANALYSIS

Jared Coopersmith, Thomas D. Cook, Jelena Zurovac, Duncan Chaplin, and Lauren V. Forrow

Abstract

This paper meta-analyzes 12 heterogeneous studies that examine bias in the comparative interrupted time-series design (CITS) that is often used to evaluate the effects of social policy interventions. To measure bias, each CITS impact estimate was differenced from the estimate derived from a theoretically unbiased causal benchmark study that tested the same hypothesis with the same treatment group, outcome data, and estimand. In 10 studies, the benchmark was a randomized experiment and in the other two it was a regression-discontinuity study. Analyses revealed the average standardized CITS bias to be between −0.01 and 0.042 standard deviations; and all but one bias estimate from individual studies fell within 0.10 standard deviations of its benchmark, indicating that the near zero mean bias did not result from averaging many large single study differences. The low mean and generally tight distribution of individual bias estimates suggest that CITS studies are worth recommending for future causal hypothesis tests because: (1) over the studies examined, they generally resulted in high internal validity; and (2) they also promise high external validity because the empirical tests we synthesized occurred across a wide variety of settings, times, interventions, and outcomes. © 2021 by the Association for Public Policy Analysis and Management

INTRODUCTION

Study Purpose

The basic interrupted time-series design (ITS) evaluates the effect of an intervention by observing the same outcome over time in a single population prior to estimating differences in mean or slope from before to after the intervention. The comparative interrupted time-series (CITS) improves on ITS by adding an *untreated comparison group*. This allows trend differences observed during the baseline period to be extrapolated into the post-intervention period where they form the null hypothesis against

Journal of Policy Analysis and Management, Vol. 41, No. 1, 252–277 (2022)

Published by Wiley Periodicals, Inc.View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.22361

which observed data in the treated and comparison groups are then evaluated. Comparing change in two non-equivalent groups from before to after an intervention is a framework CITS shares with difference-in-difference designs in general, event studies, synthetic control designs, and two-way fixed effect models. What is special about CITS is its laser-sharp focus on controlling for differences in time trends among any other differences there might be between the treatment and comparison group populations.

In most CITS studies, the comparison group is composed of persons from a different population. But other alternatives also exist. For instance, in Cook, Tang, and Seidman Diamond (2014), the treatment series charted all the crime types that a bail-bond reduction intervention targeted, while the comparison time series indexed the more serious crime categories to which the intervention did not apply. CITS data can be analyzed in many different ways. Some preserve the initially observed pretest differences and count on baseline trend differences being so clear that they can be modeled well, as with two-way fixed effects models, generalized difference-in-difference models, and event-study regressions (Wing, Simon, & Bello-Gomez, 2018). Other approaches seek to match the baseline treatment and comparison groups either directly or synthetically (Abadie, Diamond, & Hainmueller, 2010). When they are matched, the matching can be on just the immediate pre-treatment mean, or on the grand mean across all pretest time points, or on both the baseline mean and slope of the populations being contrasted. Whatever the analysis chosen, the added comparison data create a counterfactual that is predicated on a difference-in-differences (DiD) identification strategy that contrasts pre- and post-intervention data in two or more non-equivalent groups, though CITS necessarily includes pre-intervention time trend differences in its estimation.

The advantages of explicitly controlling for group baseline time trends can be illustrated by comparing the number and plausibility of the internal validity threats that operate with CITS versus standard ITS. With CITS, no treatment-correlated historical event can be a confound if it affects the treatment and non-equivalent comparison populations equally; nor can a change in instrumentation be a confound if it affects both groups equally; nor can selection bias be a problem if the treatment and comparison groups change similarly over time; nor can statistical regression be a threat if the same pre-intervention dip (Ashenfelter & Card, 1985) is observed in the treatment and comparison group; nor is a hard-to-model pretest functional form a problem if the trends are similar in the groups being compared. However, *differential* internal validity threats from these same sources are still possible. Local history describes a sudden historical shift in outcome-correlated events that occurs in one time series more than another after treatment onset. With measurement changes, the differential threat is that the outcome measurement abruptly changes more in one group than another and exactly at treatment onset. With selection bias, the differential threat is that the rate of change in group composition over the baseline period abruptly changes at treatment onset. With statistical regression, the differential threat is of a sudden pre-intervention “dip” that is larger in one group than another, a possibility that can be readily observed in the data. With hard-to-model baseline trend forms, the differential form of the threat only arises if the complexities in one group trend differ from those in the other, since it is easy to difference out functional form complexities when they are similar. Thus, the inferential strength of CITS follows from internal validity threats being less common in their differential than their general form. An added advantage is that some differential threats can be directly observed in the data, while others can be assessed by adding specific measures to a study—say, of historical events that co-occur with the intervention. Still, there can be no guarantee that every relevant differential threat will be ruled out in any CITS study.

Nonetheless, CITS bias might be rare in practice or minimal in its consequences when it does occur. We do not know. So, this paper estimates both the likelihood and size of CITS bias. It does so across 12 quite heterogeneous studies that directly difference mean effects from a CITS study and a presumably unbiased benchmark study. In 10 cases, this benchmark is a randomized controlled trial (RCT), and in the other two it is a regression-discontinuity study (RD). Each of these 12 studies contrasts CITS and benchmarks estimates when the intervention, outcome, and estimand are held constant between the designs. Thus, if the differential threats we have outlined are infrequent or cause minimal bias, then the difference in CITS and benchmark causal estimates should be close to zero in any one test. And with multiple tests, the average bias should be close to zero and the dispersion of study-specific design effect differences should be modest. If such a pattern of results were achieved in a synthesis of 12 individual studies, our hope is that it might induce social scientists and their sponsors to advocate for CITS methods with more conviction than today, perhaps even raising its profile among all the other causal methods social scientists currently use.

Single studies comparing causal estimates from a benchmark and observational study are called design experiments (LaLonde, 1986) or within-study comparisons (Cook, Shadish, & Wong, 2008); we will use the former. To estimate average CITS bias and its distribution, we conduct a meta-analysis of 12 design experiments. Especially important is to describe whether the mean bias is close to zero, and whether the distribution of single study estimates around this mean is leptokurtic, for the latter indicates that the low mean bias does not result from CITS studies that are individually quite biased in different directions. On the other hand, mean bias far from zero indicates that CITS is not a useful tool for the policy sciences because it cannot dependably produce internally valid causal results. And if the average bias is close to zero, but the distribution of study-specific effect sizes is dispersed, this too undermines the utility of CITS. While it might be unbiased over many trials, the results from any one trial cannot be trusted.

In what follows, we first discuss the CITS design and the standing we think it currently enjoys. Then we describe the strengths and limitations of design experiments and explain how meta-analyzing many of them helps reduce most of the limitations we identify. Next, we describe the data sets and synthesis methods we use before presenting the bias results. We find that average bias is about 0.03 standard deviation units, and that the distribution of study-specific bias estimates is basically leptokurtic, though one study does produce a major outlier among the three bias estimates it provides. Finally, we discuss the implications of the findings for using CITS more often and more confidently when an important causal hypothesis is to be tested.

The Comparative Interrupted Time-Series Design

CITS has probably received less discussion and research use than designs such as RCT, RD, single-group ITS, instrumental variables (IV), and DiD with matching. Nonetheless, some social science books deal exclusively with interrupted time-series and include considerable detail on the design, implementation, analysis, and interpretation of CITS, in particular (McDowall, McCleary, & Bartos, 2019). Some books that deal with nonexperimental causal methods in general also include discussion of CITS (e.g., Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). Time-series research is found in various academic disciplines. It is common in Statistics, though it receives more attention there for forecasting and standard error estimation (due to serial autocorrelation) than for causal identification (Box et al., 2015); and when some statisticians do discuss impact estimation within a time-series framework each pre-intervention time point is treated as an additional covariate within a propensity score or multivariate framework rather than as contributing to a between-group difference in intercept or slope (e.g., Rubin, 2006, 2007). Applied econometricians also recognize CITS, but they tend to consider it as another incidence of causal identification via DiD. CITS is also used in Psychology, but often with researchers collecting their own data in controlled laboratory settings that enable assessing, say, brain functioning or eyelid blinks within very short measurement intervals. CITS is much rarer in Psychology outside of laboratory settings, though some examples can be found in Educational and Community Psychology (e.g., Cook, Tang, & Seidman Diamond, 2014; Wong, Cook, & Steiner, 2015).

Some evidence about CITS' modest profile and reputation can be adduced from how it is treated in clearinghouses of effective social policies, programs, and practices. Wadhwa, Zheng, and Cook (forthcoming) considered 22 clearinghouses that detail which quasi-experimental methods they deem acceptable for certifying interventions as effective or not (Table A1). All 22 treat RCT as an acceptable design and discuss it in detail, and 20 of these are willing to accept some nonexperimental results as evidence-based. But only five explicitly deal with CITS, either by name or as a non-equivalent comparison group design with multiple baseline time points. The *What Works Clearinghouse* (WWC) provides the most detail about CITS, but its manual of 116 pages devotes only one and one-half pages to it. CITS is treated as one of three repeated measures designs, the others being growth curve and DiD analysis. Generally speaking, WWC applies the same standards to CITS as other quasi-experimental designs and so excludes them from garnering the clearinghouse's highest quality rating. This suggests a weak endorsement of CITS' ability to assess educational interventions relative to other design options.

This weak endorsement is shared by other clearinghouses too. Of the 20 clearinghouses that accept some nonexperimental designs, only five explicitly deal with CITS. The others treat it implicitly under some very general heading like DiD rather than as a design in its own right or even as a subcase of generic ITS. And of the clearinghouses that deal with CITS, none ascribes its highest study quality rating to the design or devotes as much space to describing it as compared to RCT, RD, or observational designs with some form of individual or statistical case matching. The present study probes whether CITS deserves a more unique and higher profile than this. It does so by synthesizing the results of 12 design experiments, each of which computes CITS and benchmark mean estimates that are then differenced to achieve a study-specific CITS bias estimate. The mean and distribution of these bias estimates are then computed over the 12 studies. If mean bias is low, and the distribution of estimates is tight, then researchers and their sponsors might want to consider raising their level of confidence about using CITS studies.

Design Experiments and the Special Advantages of Meta-Analyzing Their Results

A single design experiment computes bias as the difference in causal estimates between an observational study and an unbiased benchmark when each design tests exactly the same causal hypothesis in essentially the same way. The only difference should be that the treatment assignment process is fully known in the benchmark group, but not in the observational study design under test—here CITS. To meet these conditions, only RCT and RD studies can currently serve as benchmarks, and the treatment group, outcome, and estimand must be identical in the CITS and the RCT or RD study. In this set-up, similar causal estimates will only be achieved if the presumptively biased CITS comparison group is made identical to the randomly formed RCT control group by whatever sampling, design, or analysis strategies are used to adjust for all sources of observed *and unobserved* bias in the CITS study. Notice that this framing omits the treatment group altogether since it is a constant; only data from the control and comparison groups are needed—for an example of this, see Michalopoulos, Bloom, and Hill (2004). However, most design experiments do include the treatment group, and that is true of 11 of the 12 studies we meta-analyze here.

The technical quality of a design experiment depends on many features. One is that the benchmark fulfills its role by providing an unbiased and precise estimate of the intervention effect. In theory, we would like to have the exact causal parameter value. But, in practice, the RCT estimate is all we have, and it can be imprecise due to small sample sizes or can be biased if an incorrect random assignment procedure is selected or if there is treatment-related attrition or contamination (Rubin, 2008). We consider an RCT to be biased if there are differences of 0.10 standard deviations or more on key baseline characteristics related to the main study outcomes (Cook, Shadish, & Wong, 2008). Obviously, a bias estimate that depends on differencing two impact estimates will be all the more uncertain because of sampling and measurement error in both the benchmark and observational study. The single number representing CITS bias will be the product of compounded errors as well as true bias. Fortunately, such concerns are diminished when multiple design experiments are meta-analyzed. For one, any modest treatment/comparison group pretest differences should tend to cancel out across studies. For another, the overall bias due to any one RCT application should increasingly diminish as more design experiments, and thus RCTs, are added. Third, while systematic biases due to episodic improper randomization or treatment-correlated attrition or contamination are still possible in a research synthesis, they must cumulatively operate more in one causal direction than another if they are to constitute an internal validity threat. This is less plausible in the synthesis context than in a single study where the total bias can operate in only one direction. Thus, the evidence that RCT benchmarks provide is likely to be even less biased in meta-analyses than single studies.

However, two of the studies we examine use an RD rather than an RCT benchmark. Although RD is unbiased in theoretical expectation (Imbens & Lemieux, 2008), it has two major practical limitations relative to the RCT that detract from its use as a benchmark: (1) its impacts are only unbiased at the very local treatment cutoff point, and (2) its implementation requires meeting more assumptions, especially as regards functional form assumptions in parametric data analyses and bandwidth assumptions in non-parametric ones. The first limitation means that analysts should not confound a CITS estimate for all those treated with an RD estimate of all those treated who score at the treatment cutoff—except in the rare case when the RD regression lines have zero slope, and the RCT and RD models are therefore identical (Tang et al., 2017). One design experiment with an RD benchmark actually tested and met this condition (Jacob et al., 2016), while the other is less clear (Patnaik, 2019). As for analytic complications, Patnaik (2019) used non-parametric analyses whose assumptions about bandwidth are weaker than the functional form assumptions parametric analyses require. Jacob et al. (2016) used parametric methods, but also included multiple specification and functional form tests to evaluate the validity of their parametric model. We are confident, therefore, about the technical integrity of one RD benchmark but less sure about another. So, we include it in some but not all analyses and test what difference its inclusion makes. (Its inclusion makes no substantive difference to the bias findings.)

A second problem is that bias estimates from design experiments reflect the role, not just of true bias, but also of sampling and measurement error in both the observational and benchmark studies. Such error leads to the need to specify a region of practical equivalence (ROPE) within which the CITS and benchmark estimates are considered not to differ substantively even if they do differ numerically (Wilde & Hollister, 2007). Specifying such bounds is difficult with single design experiments. To conduct statistical tests of the difference between estimates entails obvious power issues when sample sizes are not large, predisposing the results towards a conclusion of no bias. To rely on causal sign equivalence has problems if the same-signed estimates are nonetheless far apart and have different practical implications. It would be laudable to estimate the costs and benefits of each impact in order to describe the substantive differences between them. But this is usually practically dubious because of the many assumptions benefit-cost analyses require (Wilde & Hollister, 2007). To conduct formal statistical tests of equivalence or non-equivalence is also laudatory, but this strategy requires very large sample sizes in order to test whether both confidence intervals of the bias estimate fall completely within those of the equivalence standard. None of the design experiments we have located to date has explicitly conducted such equivalence or non-equivalence tests. Instead, they specify a bound for acceptable bias, usually 0.10 standard deviations, and then they describe whether the obtained bias falls above or below this without reference to confidence intervals.

Testing for design equivalence is easier in meta-analysis because average bias is usually more stably estimated than study-specific bias, and also because we can conduct a formal ROPE analysis to describe the probability of obtaining bias of whatever equivalence standard is preferred. For instance, a recent meta-analysis of 15 study-specific estimates of RD bias at the treatment cutoff concluded that the average bias was of 0.01 standard deviation units and that the probability of bias falling within 0.10 standard deviations of no bias was 99 percent and of falling within 0.05 standard deviations was 97 percent (Chaplin et al., 2018). The results we present later will show the probability of bias of different levels and, in the absence of a professional consensus about which levels are worth tolerating, readers can substitute their own preferred level for those we report.

Third, design experiments seek to test the internal validity of specific quasi-experimental design and analysis options, making it imperative that these options are not confounded with other study irrelevancies that might affect the study outcome. That is one reason why the designs being contrasted should have the same treatment groups, outcome measures, and causal estimands. It is also why analysts should be blind to whether a causal estimate is from the CITS or benchmark study, thereby ruling out confirmation biases linked to researcher expectations about the relative merits of observational studies versus theoretically unbiased designs. Unfortunately, fully blind design experiments are rare in single design experiments (see Shadish et al., 2011, for an exception). But they are more plausible in meta-analyses of single design experiments, for then the shared bias of a single research team is not in play. Instead, it is any predominance in the collective bias of all those researchers who have conducted design experiments on whatever quasi-experimental practice that is under test. Some past researchers may have preferred observational studies to fare better than others did.[1] Nonetheless, a predominant direction of confirmation bias is still possible, even in the meta-analysis context; the argument is merely that its effects will generally be smaller than in a single design experiment where confirmation bias, if it exists, will tend to operate in a single direction. Since we estimate bias over multiple studies with different outcomes, we examine it in standard score form, scaling bias so that higher scores reflect more CITS bias in the direction of corroborating the hypothesis the RCT or RD study set out to test. This direction was always easy to assess, since all 12 studies set out to test potential solutions to a recognized social problem. Thus, we test whether CITS bias tends to overestimate

[1] Publication bias is also a concern. To the extent possible, we accounted for it by obtaining non-published studies and by asking researchers who had done design experiments to point us to any non-published studies they knew about.

the positive impact of social programs relative to an RCT or RD study (more on this below).

Fourth, if the benchmark has been carefully scrutinized to ascertain that it can serve as a valid causal estimate, there is little point to contrasting such a high quality RCT with a poorly designed or executed observational study; this merely confounds the design difference of interest with design differences in quality or execution. Instead, design experiments should be limited to testing only the highest quality quasi-experimental design or analysis options that are executed well, albeit within whatever limits the nature of the nonrandomized option imposes. Past design experiments indicate many design features that are ineffective in eliminating bias by themselves, even if they often reduce it—e.g., use of comparison groups that are local to the treated one, use of a pretest measure of the study outcome, or use of a "rich" set of other covariates—for a summary of the evidence, see Cook et al. (2020). Thus, there is now little point to design experiments that examine these design features alone, unless it is to probe hypotheses about the conditions under which each is more robustly effective. CITS is a stronger quasi-experimental design than any of the three features above when they are examined alone. It is worthy of design experimental tests because the pre-intervention time series has the potential to rule out the effects of time-varying selection and maturational processes in a way that is not possible with the single wave of pre-intervention data that characterize many DiD studies or with a single group longitudinal design such as ITS. Contingent alternative interpretations still remain to be ruled out, of course, including novel historical events that occurred more in one of the studied populations than the other. The rationale for claiming that CITS is worth study is not that it is "bias-free in asymptotic theory" or will be "invariably bias-free" in research practice. Rather, the rationale is that its careful design and analysis might make it "dependably free of intolerable levels of bias."

Fifth, single design experiments are inevitably weak in external validity, as are most single studies on any social or behavioral topic. With a single design experiment, it is not clear to what extent the results apply to other treatments, outcomes, populations of treatment providers and recipients, settings, times, or even different variants of the same causal design, its implementation, and how its analyses are conducted. Meta-analysis promises to be superior since average bias is estimated *across* whatever range of factors the studies collectively contain—here across 12 design experiments and the 454 treatment/comparison contrasts that arise because some studies have more than one dependent variable or examine multiple ways of analyzing the data. The studies and contrasts available touch on many sources of background heterogeneity. One is the field of application—education (Jacob et al., 2016; St. Clair, Cook, & Hallberg, 2014), health policy (Anglin et al., forthcoming; Schneeweiss et al., 2004), job training (Michalopoulos, Bloom, & Hill, 2004), family leave policy (Patnaik, 2015) and water usage (Ferraro & Miranda, 2014). Another source of heterogeneity is the sample size of respondents—from 79 to 222,261. Yet other sources are the number of baseline observation points—from four to 13; the year the design experiment was conducted—2004 to 2019; and whether the pretest observation points were matched or not—in 65 percent of the contrasts they were and in 35 percent they were not.

Six studies were associated with the same Northwestern University research team—three are in St. Clair et al. (2014) and St. Clair, Hallberg, and Cook (2016), and three others are in Hallberg et al. (2018). Given the history of quasi-experimentation at Northwestern, it seems warranted to test whether the bias estimates this team produces are systematically different from those that other researchers have produced. However, the fact that there are only 12 studies limits the number of other moderator variables we can explore. We decided on two others—the number of pretest time points, since this speaks to the stability of pre-intervention time trend estimates; and also whether the analysis matches pretest time points or not. No single study matched on just the immediate pre-treatment measurement wave; all sought to match on pretest means and slopes. So, we compare this form of matching to model-based analytic approaches that depend on validly identifying baseline trend differences.

Although this study is the first to meta-analyze design experiments that estimate bias in research using the CITS design, we are not the first to meta-analyze design experiments. Glazerman, Levy and Myers (2003) compared 12 RCT and non-time-series DiD designs in the context of job training and welfare evaluations, concluding that the bias estimates they obtained were too different from zero to recommend such simple DiD designs. Chaplin et al. (2018) compared 15 design experiments contrasting RCT and RD results at the cutoff determining treatment in RD. Their average bias estimates differed by 0.01 standard deviation units, and no study-specific shrunken bias estimate exceeded 0.10. Of course, statistical theory predicts zero difference when RCT and RD estimates are compared at the cutoff. However, RD's implementation and analysis are far from unproblematic so that the extremely modest average bias suggests that RD effect estimates are trustworthy in practice as well as in theory. Like Chaplin et al. (2018), we will describe: (1) the size of the average bias and (2) how tightly the study-specific design experiment effects are distributed around the average bias. But unlike with RD, CITS precludes a theoretical expectation of zero bias, so that the test we propose here is brutally empirical. Unlike in RCTs and RDs where the assignment mechanism in fully known, in CITS valid estimation relies on correctly modeling the pre-intervention group trends and on the assumption that the observed difference would have continued the same way into the post-intervention period absent the intervention. As a result, it is only if the study-specific estimates cluster tightly around a mean bias of zero that we can countenance recommending CITS as a practical causal method. The rationale is not theoretical in the usual statistical sense; rather, it is empirical and predicated on dependably producing minimally biased causal results over a series of studies that are heterogeneous in their interventions, outcomes, human populations, social settings, and times of intervention. Obviously, large mean bias, or widely dispersed study-specific bias effects, would make it impossible to recommend CITS.

METHODS

Identifying Relevant Studies

We sought to discover social science design experiments that evaluated the internal validity of CITS, specifying four as the minimum number of required pre-intervention time points and one as the minimum number of posttest time points. While the latter allows us to estimate changes in intercept, it precludes estimating higher-order effects such as changes in slope or variation patterns. The search for studies was conducted via journal databases and direct communication with researchers who had published any kind of design experiment prior to March 2015, with new studies identified and added as late as November 2019.

We initially identified 15 relevant manuscripts. Five were eventually excluded. Bell et al. (2016) examined the external validity of CITS, and to this end used CITS designs for both the benchmark and observational study. Fretheim et al. (2013) and Fretheim et al. (2015) used a cluster-based RCT to compare the results of an information campaign targeting primary care physicians. Most of their analyses were of single-group ITS. When they turned to CITS, their analysis used the same randomized control group as the original RCT, merely adding a time series to the analysis and so creating a study of efficiency rather than bias. We also excluded Rindskopf, Shadish, and Clark (2018) and Shadish, Rindskopf, and Boyajian (2016) since these clinical studies tested different doses of immunoglobulin for patients with immune deficiencies, and so did not meet our criterion as being of interest to social scientists.

Of the 10 remaining manuscripts on CITS bias, nine are published in peer-reviewed journals, and the other is a working paper (Anglin et al., forthcoming). One paper (Hallberg et al., 2018) used three independent data sets that we treat as independent studies. Two other papers included two data sets each (St. Clair et al., 2014; St. Clair et al., 2016), but one data set was the same in both papers, so it is included only once in this meta-analysis. Ferraro and Miranda (2017) and Wichman and Ferraro (2017) are unique published papers, but one has a single comparison group and the other has two, one of which is also used in the other publication. We treat the two publications as a single study for analysis purposes. Twelve design experiment data sets result and are described in Table 1. We refer to them henceforth as the 12 Main Sample studies, even though analyses of different data sets within the same publication are not completely independent. Most of the CITS studies estimated effects using models accounting for the baseline trends in the study outcome. Two studies used fixed effects panel estimators (Ferraro & Miranda, 2014; and Michalopoulos, Bloom, & Hill, 2004), and one study used difference-in-difference with year fixed effects (Patnaik, 2019).

Two studies have features that we belatedly discovered cast doubt on their suitability for meta-analysis. In Schneeweiss et al. (2004), the treatment group in the RCT is not the same as in the CITS. Instead, a much larger group of treated but non-randomized doctors was added to the CITS treatment population. As a result, the RCT and CITS treatment groups differ on some pre-intervention characteristics—in the case of one baseline variable by as much as 0.65 standard deviation units. Having different treatment populations in the RCT and CITS designs violates an important assumption of the theory of design experiments. The other case for exclusion involves Patnaik (2019). As its forcing mechanism, the RD benchmark uses the time lapse between date of birth and the date of implementing a new family leave policy. The CITS design also uses the policy implementation date as the intervention point, but it adds pre-intervention time points and comparison units from a neighboring province. Therefore, each design has the same treatment assignment mechanism, a point in time, even though design experiments are supposed to test nonexperimental designs whose assignment mechanism is different from the benchmark. We discovered these limitations after beginning the analysis. To avoid the possibility of cherry-picking design experiments for inclusion in the meta-analysis, we included these two studies in the Main Sample but excluded them from the Sensitivity Sample of 10 studies. Excluding them made no practical difference to the conclusions we draw.

Contrasts Per Study

The design experiments vary from four to 114 in the number of contrasts estimating CITS bias. In total there are 454 contrasts in the Main Sample and 442 in the Sensitivity Sample. As Table 1 shows, the excluded studies are characterized by fewer internal contrasts than in the other studies. Most of the contrasts result from examining more than one outcome construct per study—e.g., reading, math, and social study performance in Education—and from varying more than one analytic feature—e.g., the number of pre- and post-intervention time points, how pre-intervention matching is conducted, or how the authors adjust for serial correlation of the errors. Excluded are all contrasts based on population subgroup estimates or sensitivity tests whose results are not presented in the main tables or figures reporting estimates of bias.

Table 1. Design experiments in this meta-analysis.

Study	Publication date	Topic domain	Benchmark type	Sample sizes Benchmark/ CITS	Number of contrasts	Pretest time points	Matching used	Northwestern University influence	Parallel trends
Anglin et al.	forthcoming	Health care utilization	RCT	1,183/1,282	24	12	Both	Yes	Mixed
Ferraro and Miranda; Wichman and Ferraro	2014; 2017	Water usage	RCT	83,319/138,942	20	13	Both		Yes
Hallberg et al.	2018								
data set 1: Algebra		Education	RCT	45/5,112	96	4	Both	Yes	Mixed
data set 2: College		Education	RCT	47/736	30	10	Both	Yes	Yes
data set 3: Literacy		Education	RCT	45/1,554	114	4 or 6	Both	Yes	Mixed
Jacob et al.	2016	Education	RD	168/680	16	6	Both		Yes
Michalopoulos et al.	2004	Job training	RCT	4,015/12,382	88	8	Both		Mixed
St. Clair et al.	2014								
data set 1: Indiana 1		Education	RCT	63/1,118	14	4-6	Both	Yes	Mixed
St. Clair et al.	2016								
		Education	RCT	63/1,118	12	4-6	Both	Yes	Mixed
data set 2: Indiana 2		Education	RCT	63/1,040	10	6	Both	Yes	Mixed
data set 3: Florida		Education	RCT	64/1,825	18	7	Both	Yes	Mixed

Table 1. (Continued).

Study	Publication date	Topic domain	Benchmark type	Sample sizes Benchmark/ CITS	Number of contrasts	Pretest time points	Matching used	Northwestern University influence	Parallel trends
Studies with weak design experiment qualities									
Patnaik	2019	Family leave	RD	1,781/8,907	8	4	No		Yes
Schneeweiss et al.	2004	Healthcare utilization	RCT	835/13,354	4	6	No		Mixed

Notes: Each dataset is treated as a study. Thus, the use of "data set 1: Indiana 1" by St. Clair et al. (2014) and St. Clair et al. (2016) is treated as a single study even though the contrasts were published in two separate papers.

Coding

Coding involved three stages: initial coding, quality control (QC) of the coding, and quality assurance (QA) of the QC process. In the first stage, a reviewer was assigned a single manuscript in order to judge whether it included one or more design experiments and then to complete all study- and contrast-level codes before sending their completed coding sheet to the research team member serving as their QC for this manuscript. That person then read the study in question and closely reviewed the coding choices of the primary reviewer. Once this was completed, a reconciliation meeting was held where coding discrepancies were discussed and a reconciliation was achieved, usually after detailed examination of the published specifics. The resulting codes were then sent to a different research team member tasked with the QA role. This person read the study and examined the most important codes—items that had needed reconciliation, whether all eligible contrasts were included, whether causal estimates had been correctly calculated and recorded, whether the direction of CITS bias was correctly recorded, and details of the analytic approach in the CITS and in the study as a whole. Except for the CITS-specific codes, the categories and processes used were like those in Chaplin et al. (2018).

The RCT-relevant codes register whether random assignment was at the individual or cluster level, whether there was any known threat to the random assignment mechanism, whether the authors assessed baseline balance, what the results of this assessment were, and which kinds of control variables were used in the RCT impact analysis. The RD-relevant codes verified that the analysis controlled for the running variable, described whether the analysis used a parametric or non-parametric model, if an optimal model specification algorithm was used, and if the authors tested for changes in density at the cutoff score. Other benchmark codes were common to the RCT and RD benchmarks and recorded the number of units observed, the impact estimate and standard error, and the estimand type. We coded the pooled standard deviation of the RCT when it was provided. If it was not, we estimated it from the standard deviations for the separate treated and untreated units.

The CITS design codes included the number of observed time points in the pre- and post-intervention periods, what type of analytic model was used to model the error structure, whether the control group was geographically local, if case-matching was used to create or refine the control group, what types of covariates were used in the analysis, and how seasonality was modeled where relevant. We also recorded the stability of the pre-intervention outcome trends, using the authors' stated judgment and our own examination of the available data, preferring the latter if there was any conflict. The types of internal validity threats we recorded included evidence for differential statistical regression at or around the time of treatment, a local history difference at that time, a change in instrumentation, and a change in the sample compositions.

The final codes assess how the correspondence between the benchmark and CITS study was assessed—whether the same covariates were used in each analysis, whether the estimands were the same, whether impacts were estimated at the same time points, and whether the benchmark and CITS used the same cross-sectional or longitudinal method for collecting their time-varying data.

To describe the effect size for each contrast, we recorded the benchmark and CITS causal estimates in standard score form. Most outcomes were continuous, and their standard score transformation used standard meta-analytic methods (Cooper, Hedges, & Valentine, 2019). The CITS and benchmark estimate differences were scaled so that the CITS option always scored higher in the direction of desired study outcomes—e.g., higher income or educational achievement—thereby allowing us to examine the claim that observational study bias generally produces more positive study outcomes than theoretically better warranted methods might achieve

(Iaonnides, 2005). The chance to test this would be obscured if we were to describe mean absolute bias by summing over all bias estimates irrespective of their causal sign. So, in calculating mean bias, we respect the causal signs attached to the various contrast estimates. However, we also report study-specific estimates of mean bias, each with its own causal sign and variation, thus allowing readers to judge the direction of bias in each study.

Given only 12 design experiment studies, few study-level covariates can be used in the analysis. We settled on three: (1) The number of pretest time points, because trend differences are better estimated, and matching is more complete, the longer the pre-intervention time series is; (2) Whether the CITS data analysis did or did not match pretest means and trends, because the What Works Clearinghouse suggests one simple form of matching to be necessary for acceptable CITS studies; and (3) Whether Northwestern personnel conducted the design experiment or not, in case the studies in St. Clair et al. (2014, 2016) and Hallberg (2018) might be subject to a confirmation bias that some commentators might want to link to that university's history of research on the theory of quasi-experimentation (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979; and Shadish, Cook, & Campbell, 2002). The Northwestern covariate varies only between studies; the number of pretest time points varies within only two of the studies—the Hallberg et al. (2018) Literacy study and the St. Clair et al. (2014) Indiana study—while pretest matching varies within most studies and accounts for 65 percent of all bias estimates at the contrast level.

Modeling Approach

In overviewing our modeling approach, note that we seek to estimate three properties: the bias and variation in bias in individual studies, the mean and variation in bias across studies, and the probability that the next study-level CITS bias estimate will fall within the region of practical equivalence (ROPE) that we set to be 0.10 standard deviations on either side of the RCT estimate. To achieve these ends, we estimate essentially the same model, but do so using three different analytic techniques—a fully Bayesian, an empirical Bayesian, and a frequentist analysis. The Bayesian analyses have the advantage of "borrowing" strength across studies and thus of reducing the imprecision of estimates based on smaller, or otherwise less precise, individual studies. But such pooling requires additional assumptions, not all of them transparent to many readers or acceptable to those with a preference for methods that make fewer and more transparent assumptions. That is why we include a frequentist meta-analysis, even while preferring the fully Bayesian one.

The two Bayesian models differ in how they account for uncertainty and prior distributions. The fully Bayesian model accounts for uncertainty in the estimate of the variance across study-specific bias estimates that (below) we call σ_a^2. In contrast, the empirical Bayesian model treats this component as fixed and does not propagate its estimation uncertainty throughout the analysis as the fully Bayesian model does. Consequently, the fully Bayesian model produces study-specific bias estimates with greater uncertainty than those from the empirical Bayesian model. (Being based on more data, mean bias estimates are less affected by this uncertainty than are study-specific ones.) The fully Bayesian and empirical Bayesian results both provide estimates on the extent of bias that can be expected in future CITS studies, based on the estimated variation in bias, and they both improve the prediction of bias in the worst powered studies using the more precisely measured studies. The fully Bayesian model also accounts for uncertainty in the estimation of the variance in bias across studies. As a result, the fully Bayesian estimates more fully reflect uncertainty than the empirical Bayesian. We provide both sets of results as readers may differ on their relative comfort with the two methods.

Both Bayesian analyses have the same regression equation:

$$y_{ij} = \alpha + \beta X_i + a_j + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N\left(0,\ \sigma^2/w_{ij}\right).$$

In this model, y_{ij} is the standardized bias of contrast i from study j. Average standardized bias is reported as the sum of an overall intercept term α, which represents the average standardized bias across contrasts and studies at the mean of the covariates. The study-specific effect a_j represents the difference between this overall standardized bias, α, and the average standardized bias across contrasts within study j. β is a vector of coefficients for the three background covariates (X); and the error is ε_{ij}. The second formula above shows that the variance of ε_{ij} is a function of the variation in bias, including measurement error, and study weights w_{ij}. These weights are proportional to the sample size of the comparison group in the RCT and CITS, a proxy for the precision of the standardized bias estimate.[2] The formula for the weights is:

$$w_{ij} = 1/\left(\frac{1}{N^C_{RCT_{ij}}} + \frac{1}{N^C_{CITS_{ij}}}\right).$$

The first term in the denominator approximates the precision of the RCT arm through the sample size in the control group, while the second term approximates the precision of the CITS arm through the sample size of its comparison group.

In both Bayesian versions, we borrow strength across studies according to the formula:

$$a_j \sim N\left(0,\ \sigma_a^2\right).$$

Thus, we conceive of study-specific bias estimates as coming from a common distribution with variance σ_a^2, which is estimated from the data. When a particular study estimates bias imprecisely, its a_j will be adjusted towards the across-studies estimate of average standardized bias. In contrast, the more precise estimates of a_j will hardly differ from what their unadjusted estimate would have been in a non-Bayesian model. The fully Bayesian model, unlike its empirical Bayesian counterpart, also requires specifying prior distributions for all the terms in the model. To this end, we used the following default prior distributions, as recommended in Gelman et al. (2008).

$$\alpha \sim t_7\,(0,\ 2.5)$$

$$\beta \sim t_7\,(0,\ 2.5)$$

$$\sigma,\ \sigma_a \sim N^+\,(0,\ 1)$$

Fully Bayesian analyses were fit via Markov Chain Monte Carlo using the Stan programming language's R interface, *rstan* (Carpenter et al., 2017). Empirical Bayesian analyses were fit via restricted maximum likelihood using the R package lme4 (Bates et al., 2015). A third, frequentist analysis was also conducted using the same model but following the procedures outlined in Tipton and Pustejovsky (2015) that were

[2] Results reported below do not differ substantively based on the use of weights.

Table 2. Estimated average bias across all studies as a function of three data-analytic strategies and two samples.

Fully Bayesian		Empirical Bayes		Frequentist	
Main	Sensitivity	Main	Sensitivity	Main	Sensitivity
−0.003 (0.068)	0.013(0.075)	−0.041 (0.068)	−0.002 (0.096)	−0.03(0.033)	0.01(0.010)

Note: Results are based on models controlling for three covariates as described in text.

specifically designed for correlated contrasts when the sample of studies is small—factors that characterize this set of 12 studies. We use this frequentist approach to again estimate average CITS bias and its confidence intervals. The analysis was conducted using the software package *robumeta* in R (Fisher, Tipton, & Zhipeng, 2019; R Core Team, 2018).

To summarize, the main difference between these three analytic methods is the assumptions they make about the relationships among studies and the implications that follow from these different assumptions. The frequentist approach assumes that each study is largely independent of the others. The empirical Bayesian approach allows for some pooling of information across studies, but it does not account for uncertainty in the model's assessment of how much pooling to induce. The fully Bayesian approach pools information across studies, accounts for uncertainty in how much pooling the model deems appropriate, and propagates this uncertainty through to the final estimates of study-specific bias that will be less precise than with the two other methods. Some researchers will like maximizing uncertainty because the bias estimates are used to compute the probability of bias in future CITS scenarios that are like the old ones but cannot be identical to them. However, other researchers might prefer to describe past findings without aspiring to extrapolate, albeit with some pooling as in the empirical Bayesian approach or with even less pooling as in the frequentist approach. These different assumptions and their consequences incline us to assess the robustness of bias results across all three analyses, each of which uses the same set of contrasts that come either from 12 studies in the Main Sample or from 10 in the Sensitivity Sample.

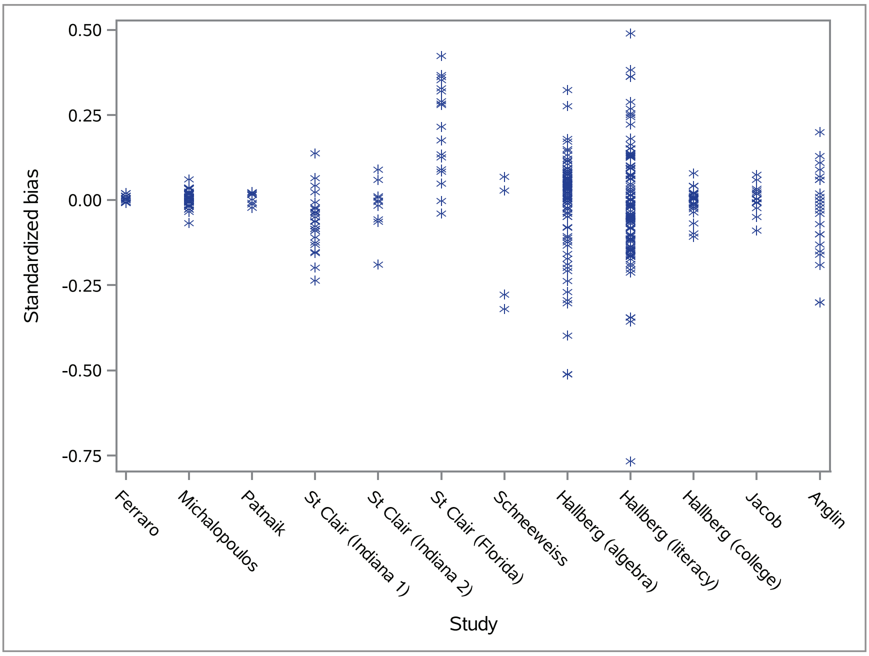
RESULTS

Mean Bias Estimates

Table 2 presents the mean bias results from the three types of analysis and two non-independent samples. All Bayesian estimates fall between 0.041 and 0.002 standard deviations, and none differs from zero at the 5 percent level. The frequentist results are similar—mean standardized CITS bias of −0.03 (se = 0.05) with the Main Sample, and of 0.01 (se = 0.01) with the Sensitivity Sample.

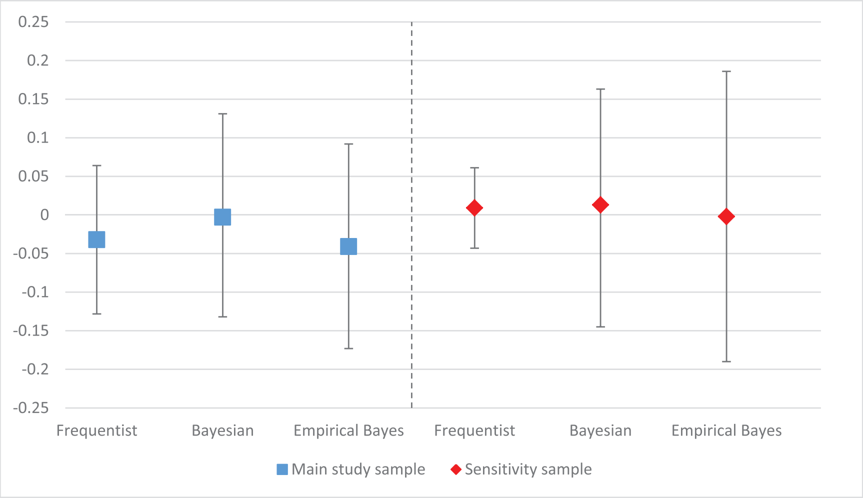
Distribution of Study-Specific Effects

Figure 1 presents the raw data by contrast within each study. It shows that the contrast-specific estimates are very variable by study but are generally normally distributed around a mean close to zero except in the one St. Clair et al. (2016) example in Florida. Figure 2 presents 95 percent credible bounds for the Bayesian analyses and 95 percent confidence intervals for the frequentist analyses. These estimates of variation tend to be higher in the fully Bayesian work and lowest in the frequentist, and all but one exceed the ROPE standard of 0.10 standard deviations.



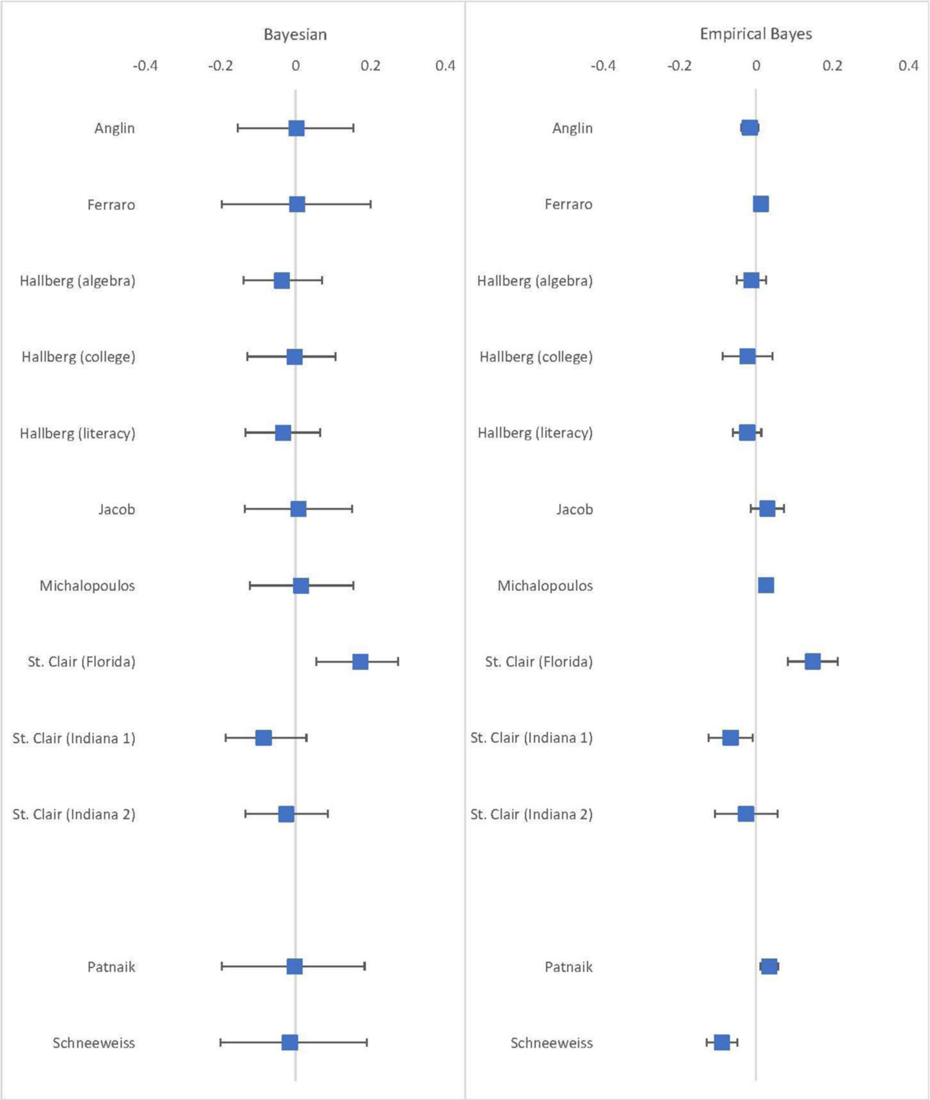
Notes: Based on results presented in Table 2.

Figure 1. Distribution of Standardized Bias by Contrast Within Studies.



Note: Bars represent the 95 percent confidence intervals and the 95 percent credible intervals for the frequentists and Bayesian analyses, respectively.

Figure 2. Average Standardized Bias as a Function of Three Data-Analytic Strategies and Type of Sample.
[Color figure can be viewed at wileyonlinelibrary.com]



Note: The bars represent the 95 percent credible intervals.

Figure 3. Study-Specific Effects for the Main Study Sample.

Figures 3 and 4 are more relevant and present the study-specific shrunken bias estimates and their credible bounds. In the fully Bayesian model, these estimates were calculated from the posterior distribution for each study-specific term in the regression model (a_j); and the values presented are the mean and 95 percent credible interval bounds of those distributions. For the empirical Bayesian study-specific effects, the figure presents the mean and 95 percent credible interval based on the random effects component of the model. The results show a generally tight distribution of study-specific effects around zero, and only one value exceeds 0.10 SDs—the Florida study of St. Clair et al. (2016) where the bias estimate is 0.174 (se = 0.055) in the fully Bayesian analysis and is 0.149 (se = 0.033) in the empirical Bayesian analysis. In contrast, the next highest single estimate is for the St. Clair et al. (2014)

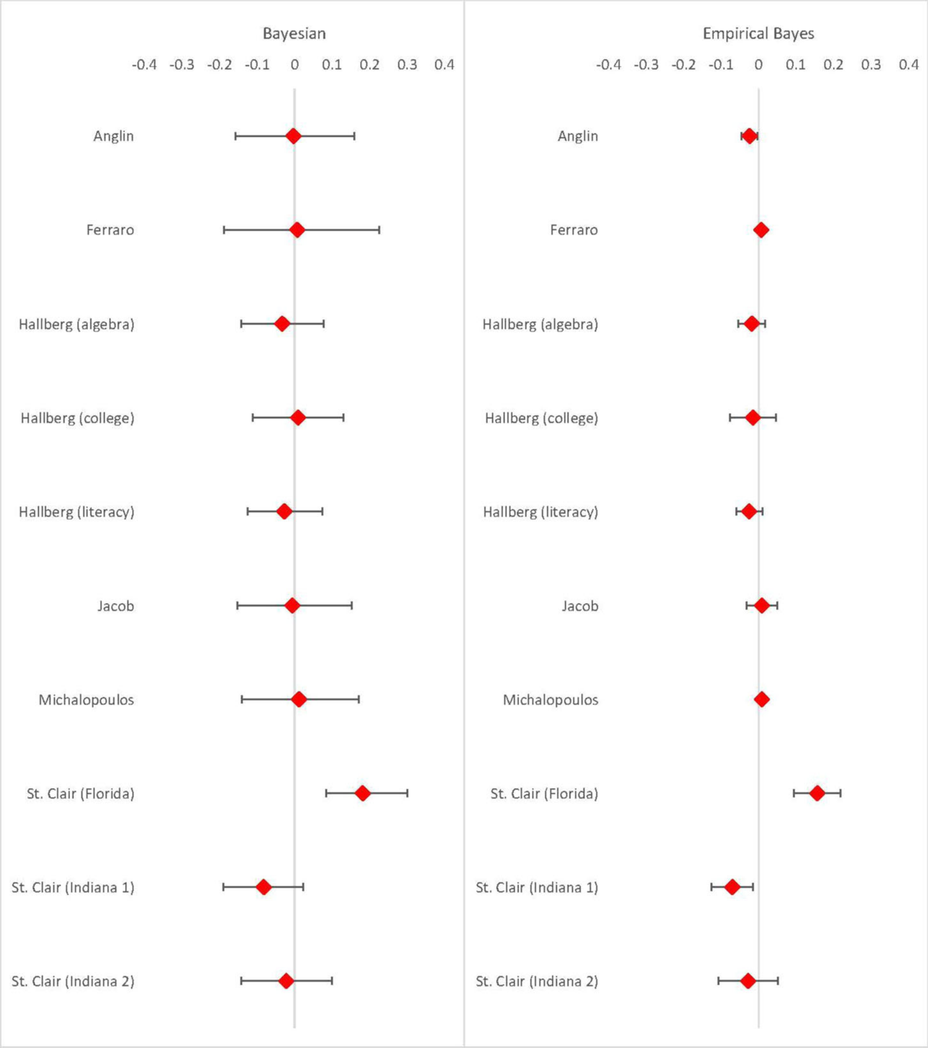


Figure 4. Study-Specific Effects for the Sensitivity Study Sample.

Indiana I study where the value is −0.085 (se = 0.051) in one analysis and −0.066 (se = 0.03) in the other. Thus, all but one of the 12 study-specific estimates are under 0.10 SDs.

Figure 5 reports the ROPE results based on the Bayesian posterior probability distribution. It depicts the estimated probability that the CITS and benchmark results will differ in the next study by a specific amount in standard deviation units. In the Main Sample, the probability of bias under 0.10 is 86.5 percent, indicating that in almost 11 out of 12 future CITS studies the estimate will differ from the RCT's by 0.10 SD or less. The corresponding value with the Sensitivity Sample is 81.4 percent, indicating that just over eight of 10 CITS estimates will differ from the RCT by this much. If we lower the ROPE to 0.05 standard deviations, then the corresponding probabilities are just over 50 percent with each sample.

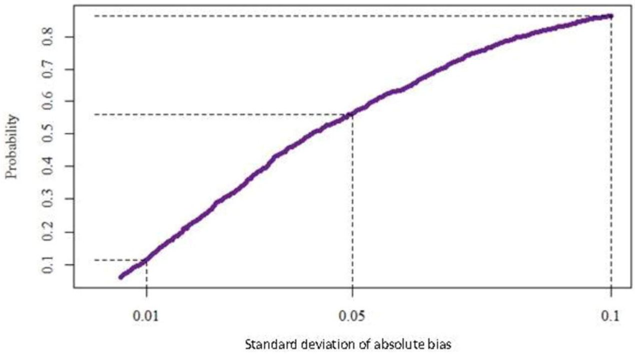


Figure 5. The ROPE Function for the Main Sample.

Table 3. Associations between bias and pretest time points, matching, and Northwestern influence.

	Bayesian		Empirical Bayes	
Parameter	Coefficient	SE	Coefficient	SE
Number of pretest time points	−0.007	0.010	0.002	0.007
Used matching	−0.018	0.017	0.001	0.003
Northwestern influence	0.033	0.082	0.046	0.046

Determinants of Variation in Bias Levels

Table 3 presents the bias estimates associated with the three covariates, though the analyses of only 12 studies have modest statistical power. Since none of the three covariates has either a large or statistically significant relationship to bias, bias is not likely to have been moderated by the number of pretest observation points, by matching versus modeling baseline trend differences, or whether the design experiment is conducted by researchers associated with Northwestern University.

DISCUSSION AND CONCLUSIONS

Results for both the Main and Sensitivity Samples show that bias was not related to whether the meta-analysis included 10 or 12 design experiments. Notwithstanding, the results we discuss below are limited to the 10 studies since each of them meets the theoretical requirements for a high-quality design experiment. Since bias was also not related to whether the three covariates were in the impact model, the results we discuss below are limited to those from models with covariates.

One main substantive finding is that average CITS bias is low, being under 0.05 standard deviations in every analysis with this particular sample of design experiments, and model-based extrapolations from the current studies to similar future CITS applications resulted in a probability of 81.4 percent that the CITS impact would be biased by as much as 0.10 SDs. By the more stringent 0.05 SD criterion, the probability was 51.7 percent. Since the distribution of study-specific estimates determines the practical relevance of low average bias, our second main finding was that these estimates were generally tightly distributed around the mean. Of the 10 study-specific estimates in the Sensitivity Sample, nine were below 0.10 SDs. The exception was the Florida study of St. Clair et al. (2016) and, when it was omitted for the record, mean bias fell to under 0.016 SDs in both Bayesian analyses while the probability of obtaining bias below 0.10 SDs rose from 81.4 percent to 96.8 percent for the nine remaining studies. Thus, the distribution of study-specific effects can be characterized as generally (but not universally) tight around a low estimate of average bias.

Why was the St. Clair et al. (2016) CITS estimate in Florida such an outlier? We cannot be certain, but the published paper noted after-the-fact that the baseline slope difference between the treatment and comparison series was not clear and that the matching of both groups that was then necessitated was itself only marginally successful. The implication is that a CITS that was technically weak by current CITS design standards might have been tested against an RCT that was tested and shown to be strong by its own design standards. Even so, this one published study could not have been omitted from the present meta-analysis because such after-the-fact cherry-picking undermines the extrapolations to future scenarios that the fully Bayesian analysis sought to achieve.

The Florida outlier may make some commentators doubt whether the present findings are dependable enough to support a strong conclusion about the priority CITS deserves when decisions are made about causal method choice. However, accepting the bias result from the Florida study leads us to frame the ROPE analyses as follows. How willing are you to endorse CITS if there is an 81 percent probability of achieving a study average causal estimate within 1/10th of a standard deviation of the estimate from an RCT that tests the same causal hypothesis within the same external validity context represented by this particular sample of design experiments? In educational research with academic achievement outcomes, 1/10th of a standard deviation equates to raising test scores from the median to the 58th percentile. If this is held to be too wide and a ROPE of 1/20th of a standard deviation is accepted instead, then the question becomes: How willing you would be to endorse CITS if it has a 56 percent chance of agreement within four percentile points in fully Bayesian models that seek to maximize sources of uncertainty in the estimate? No magic answer to these questions is possible, and stakeholders in the social and behavioral sciences will probably differ in their answers. But we contend that the present results are *at least* promising for making the case that CITS' profile and reputation are worth raising and that it should be recommended for wider and more confident use in causal research.

Of course, some well-conducted CITS studies will inevitably result in intolerable levels of bias, just as some single RCT studies do because of sampling and measurement error or because of treatment contamination and attrition issues. And the dependability of CITS applications with minimal bias can be increased even further, including by future design experiments that seek to identify even more of the design, sampling, and analysis specifics that reduce bias in CITS research, thus permitting even better specification of what constitutes high quality CITS research.

The design experiments we examined are heterogeneous on many attributes that might affect bias. Some are methodological—e.g., the number of pretest observation points, whether the analysis involved matching or not, and whether the analysts might be more likely to favor observational designs. Though the small sample size makes statistical significance results of these potential moderator variables suspect, the estimates are all small and suggest that bias is not strongly related to such methodological features. Other features that might affect bias are more substantive. They involve sampling attributes like the types of intervention and outcomes studied (see Table 1), the primary authors' academic discipline—mostly Education, Economics, and Political Science, the study populations selected, the physical settings studied, and when studies were published as they span more than a decade. Such heterogeneity boosts confidence in the external validity of the internal validity of CITS. Obviously, if an even more heterogeneous set of design experiments replicated the high internal validity of CITS, this would further extend the design's external validity.

Inevitably, the present study has limitations. One relates to the population of CITS studies. We examined "all CITS studies in the social sciences that have been evaluated against an RCT or RD." But this has an obscure relationship to our implicit population of interest—"all possible CITS studies in the social sciences." The possibility of mismatch between obtained and targeted populations can be levelled against almost all causal studies evaluating manipulated interventions (Cook, Shadish, & Wong, 2008). Few of them select their samples of persons, settings, and times at random from clearly designated universes, and few replicate their results across the multiple versions of a theoretical cause that are possible. Instead, samples and operational definitions have an opportunistic, purposive flavor that makes their referent populations and constructs hard to pin down. And even when many heterogeneous tests of the same hypothesis are available, the operative population is still of all studies *that could be found* rather than all studies that *could be done*. Causal generalization always requires an extrapolative leap from the samples at hand to the populations of interest, and the claim we make here is modest. It is that the population of 10 opportunistic design experiments we analyzed is more diverse than any single study can achieve, suggesting that the very small bias we found is dependably replicable across diverse research contexts, though it is not inevitably to be found—see the one exception obtained here.

A second limitation to the generalization of the present findings is more focused. By definition, design experiments cannot deal with CITS designs that are undertaken because an RCT on the same topic is not possible. For instance, many CITS studies are undertaken to examine universal policy changes that are made at a specific time for a given population in a context where political realities preclude randomly assigning individuals, households, communities, or states from within the population. While the studies we examined are heterogeneous on many attributes, they cannot vary whether an RCT or RD is feasible. One of them has to be feasible; otherwise, a design experiment cannot be conducted. No definitive empirical or theoretical argument can fully counter the claim that the present findings will not hold with CITS studies that are undertaken in settings where an RCT or RD is truly impossible. At best, it is possible to argue inductively that the present findings serve to increase the shared subjective odds that little CITS bias will occur in contexts where RCT and RD are truly impossible.

A third limitation to the external validity of the present findings would arise if the CITS studies in this meta-analysis are better designed, implemented, and analyzed than the CITS studies that other researchers carry out. This could happen if the researchers conducting design experiments are more interested in causal methodology and know more about it than other research practitioners, including about CITS. The CITS analyses examined here were quite standard and very heterogeneous. Three always used time fixed effects, another four used time fixed effects in some analyses but not in others, and four studies modeled the outcome using a baseline mean or the baseline mean and trend. The remaining studies used other analytic approaches. As for design features, only two studies always used a fully local comparison group, another four used either a local comparison group or a mix of local and non-local comparisons, while the remainder used non-local comparison units. One study exhibited evidence of statistical regression, and another two studies had mixed evidence about a pre-intervention "dip." No study exhibited evidence of historical confounds, of instrumentation changes, or of selection bias that might affect the outcome trends. Finally, while one study used only the pretest time trend of the outcome to estimate treatment effects, four others added covariates to estimate the treatment effect, and seven varied how the covariate data were used. The present studies did not reflect a single favored theory of how to design and analyze CITS studies and, while all analyses seemed competent to us, none seemed to be better than what researchers commonly do to conduct CITS studies in contexts that are irrelevant to design experiments. Nonetheless, some commentators may still fear that the present findings will encourage some researchers to conduct CITS studies whose methodological quality is lower than the worst in this synthesis. But this is an argument for better education and professional development rather than a source of doubt about what this meta-analysis teaches us about the internal and external validity of CITS.

A fourth limitation follows from the uncontested truth that the five differential internal validity threats we described earlier *can* operate with CITS. The present study was concerned with how often they operate and with the total bias they might cause when they do. In the purposive sample of design experiments examined here, we found that the threats in question had little effect. From an armchair, they could have caused more bias; but in reality, they did not. From the perspective of sampling theory, RCT is clearly stronger than CITS because of the unbiased causal expectations its random assignment feature justifies. RCT is always to be preferred. However, the results presented here show that, to date, there is little to choose between the two designs when they are evaluated against each other empirically. Of course, in the real world of social and behavioral research, some researchers will always prefer theoretically justified methods over empirically warranted ones. And, more importantly perhaps, RCT and CITS designs rarely compete in the way that design experiments require. CITS is more often used when policy changes are made at a known date and apply to all members of the target population, making it difficult to mount an RCT. Conversely, RCTs command a control over treatment allocation processes that is often not feasible when CITS studies are undertaken today. Notwithstanding, the present findings indicate that researchers need have little apprehension about the internal validity of the findings from well-designed CITS studies, and we suggest such studies can now be done with less concern so long as the quality of the CITS design, implementation, and analysis details does not fall below those of the studies synthesized here.

The final limitation of the current study is that the analyses are based on either 12 or 10 studies. This is obviously preferable to analyzing a single study, but it is still a smaller sample than in many other meta-analyses and it reduces the statistical precision of estimates. However, it was not a study priority to compute null hypothesis tests of whether the mean bias estimates differed from zero. Rather, the study focused more on estimating the expected mean bias and on describing the potential for variation around this mean. Precision is relevant to the confidence intervals and credible bound estimates in Figure 2, but is most crucial for study purposes in the estimates of study-specific bias in Figures 3 and 4 and in the predicted probability of study-specific bias of a given magnitude in Figure 5. These findings are stable enough to tell a consistent story of low bias in all but one of the individual design experiments. It is worth noting here, though, that the present findings concern study-level bias estimates and not necessarily contrast-level ones. The study-level of analysis asks how much bias there is averaged across all of the contrasts in a study irrespective of whether there is a positive or negative causal sign attached to them—signs that could indicate sampling error or true heterogeneity in bias. The former is more plausible in contexts such as here where the mean bias was close to zero and the contrasts were symmetrically distributed around this mean, as shown in Figure 1. Moreover, most of the contrasts are formed by ex-post-facto internal analyses of different groups of respondents, settings, and times, by different variations of the same intervention, and by different ways of analyzing time-series data with non-equivalent groups. Very few emanate from different outcomes in the same study, and none from different outcomes deliberately chosen to test for both positive and negative true effects. Instead, contrasts were generally chosen to probe marginal differences in the same positive sign. Even so, this paper does not directly speak to the likelihood of obtaining unbiased effects for all the contrasts that a CITS might examine, of which there were about 44 per study here. It only speaks to the average bias over all the contrasts in a study.

We paid relatively little attention to the data analysis choices that CITS requires and that are important for both technical and policy reasons. One technical issue concerns the bias in standard errors due to the correlation between adjacent and nearby errors in time-series work. Our meta-analysis accepted whatever methods the original analysts used to adjust for serial correlation, and it is important to note that every analysis used one method or another for this. The implication here is that current time-series practice acknowledges and deals with the possibility of correlated errors, even if the strategies for accomplishing this differ, but these differences do not seem to be consequential in the studies examined. The other main analytic issue concerns whether achieving well-balanced baseline means is necessary for responsible causal inference, as WWC assumes, or whether approaches are as good that model baseline trend differences without losing cases when no matches can be found. Of the 454 bias contrasts analyzed, 65 percent involved matching and 35 percent did not and there were no meaningful differences in bias estimates between them. In addition, eight studies reported internal analyses with and without matching, and the unweighted average of these within-study differences was −0.01 standard deviations. So, matching was not associated with greater bias reduction, though some form of it is required when baseline trends are so variable that no reasonable estimate of trend difference is possible. But it was not difficult to predict trend differences in all but one of the cases tested here, suggesting that we should question whether model-based approaches to CITS analysis should be banned. Indeed, we think that all clearinghouses of effective social programs, policies, and practices should critically examine (1) whether to allow in CITS results at all—we think they should; (2) whether to allow in only CITS results from analyses with matching—we think not, although checks of the stability of baseline population trend differences might be worth insisting on; (3) whether to allow in only studies that match on the pre-intervention time point closest to the intervention point—we think not, since the rest of the baseline time series is useful too; and (4) whether all CITS studies automatically deserve quality ratings lower than those from studies with random assignment or regression discontinuity—we think not, again, given the results reported here.

It is time to reassess the standing of CITS. Its *theoretical warrant* is obviously more problematic than when the treatment assignment process is fully known. But *its current empirical warrant* is much stronger, since the present results indicate that the main internal validity threats that *can* bias CITS estimates may operate only rarely in meaningful ways. Here, all but one of 10 heterogeneous studies produced acceptably low bias as judged against an RCT counterpart that was unbiased in theory and likely also as implemented in practice.

JARED COOPERSMITH is a Senior Statistician at Mathematica, 1100 First Street, NE, 12th Floor, Washington, DC 20002 (e-mail: JCoopersmith@mathematica-mpr.com).

THOMAS D. COOK is Professor Emeritus of Sociology at Northwestern University in Evanston, IL, and Research Professor at the Trachtenberg School of Public Policy and Public Administration at The George Washington University in Washington, DC (e-mail: tomcook6@gwu.edu).

JELENA ZUROVAC is a Senior Researcher at Mathematica, 1100 First Street, NE, 12th Floor, Washington, DC 20002 (e-mail: jzurovac @mathematica-mpr.com).

DUNCAN CHAPLIN is a Principal Researcher at Mathematica, 1100 First Street, NE, 12th Floor, Washington, DC 20002 (e-mail: DChaplin@mathematica-mpr.com).

LAUREN V. FORROW is an Independent Consultant for Mathematica, 955 Massachusetts Avenue, Suite 800, Cambridge, MA 02139 (e-mail: lvollmer@mathematica-mpr.com).

ACKNOWLEDGMENTS

The research reported here was supported by NSF grant 1760458 to Thomas D. Cook.

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (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.

Ashenfelter, O., & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. The Review of Economics and Statistics, 67, 648–660.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67, 1–48.

Bell, S., Olsen, R., Orr, L., & Stuart, E. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. Educational Evaluation and Policy Analysis, 38, 1–18.

Box, G., Jenkins, G., Reinsel, G., & Ljung, G. (2015). Time series analysis: Forecasting and control. Hoboken, NJ: Wiley.

Campbell, D., & Stanley, J. (1963). Experimental and quasi-experimental designs for research. Boston, MA: Houghton Mifflin Company.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Brubaker Riddell, A. (2017). Stan: A probabilistic programming language. Journal of Statistical Software, 76, 1–32.

Chaplin, D., Cook, T., Zurovac, J., Coopersmith, J., Finucane, M., Vollmer, L., & Morris, R. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. Journal of Policy Analysis and Management, 37, 403–429.

Cook, T., & Campbell, D. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton Mifflin Company.

Cook, T., Shadish, W., & Wong, V. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management, 27, 724–750.

Cook, T., Tang, Y., & Seidman Diamond, S. (2014). Causally valid relationships that invoke the wrong causal agent: Construct validity of the cause in policy research. Journal of the Society for Social Work & Research, 5, 379–414.

Cook, T., Zhu, N., Klein, A., Starkey, P., & Thomas, J. (2020). How much bias results if a quasi-experimental design combines local comparison groups, a pretest outcome measure and other covariates? A within-study comparison of pre-school effects. Psychological Methods, 25, 726–746.

Cooper, H., Hedges, L., & Valentine, J. (Eds.). (2019). The handbook of research synthesis and meta-analysis, 3rd Edition. New York, NY: Russell Sage Foundation.

Fisher, Z., Tipton, E., & Zhipeng, H. (2019). robumeta: Robust variance meta-regression. R package version 2.0.

Fretheim, A., Soumerai, S., Zhang, F., Oxman, A., & Ross-Degnan, D. (2013). Interrupted time-series analysis yielded an effect estimate concordant with the cluster-randomized controlled trial result. Journal of Clinical Epidemiology, 66, 883–887.

Fretheim, A., Zhang, F., Ross-Degnan, D., Oxman, A., Cheyne, H., Foy, R., … Soumerai, S. B. (2015). A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation. Journal of Clinical Epidemiology, 68, 324–333.

Gelman, A., Jakulin, A., Pittau, M., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics, 2, 1360–1383.

Glazerman, S., Levy, D., & Myers, D. (2003) Nonexperimental versus experimental estimates of earnings impact. The Annals of the American Academy of Political and Social Science, 589, 63–93.

Iaonnides, J. (2005). Why most published research findings are false. Chance, 18, 40–47.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142, 615–635.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, 76, 604–620.

McDowall, D., McCleary, R., & Bartos, B. (2019). Interrupted time series analysis. New York, NY: Oxford University Press.

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available at. https://www.R-project.org/.

Rindskopf, D., Shadish, W., & Clark, M. (2018). Using Bayesian correspondence criteria to compare results from a randomized experiment and a quasi-experiment allowing self-selection. Evaluation Review, 42, 248–280.

Rubin, D. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. Statistical Science, 21, 299–309.

Rubin, D. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. Statistics in Medicine, 26, 20–36.

Rubin, D. (2008). The design and analysis of gold standard randomized experiments. Comment on "Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment" by Shadish, Clark, & Steiner. The Journal of the American Statistical Association, 103, 350–1353.

Shadish, W., Cook, T., & Campbell, D. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin Company.

Shadish, W., Galindo, R., Wong, V., Steiner, P., & Cook, T. (2011). A randomized experiment comparing random to cutoff-based assignment. Psychological Methods, 16, 179–191.

Shadish, W., Rindskopf, D., & Boyajian, J. (2016). Single-case experimental design yielded an effect estimate corresponding to a randomized controlled trial. Journal of Clinical Epidemiology, 76, 82–88.

Tang, Y., Cook, T., Kisbu-Sakarya, Y., Hock, H., & Chiang, H. (2017). The comparative regression discontinuity (CRD) design: An overview and demonstration of its performance relative to basic RD and the randomized experiment. Advances in Econometrics, 38, 237–279.

Tipton, E., & Pustejovsky, J. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. Journal of Educational and Behavioral Statistics, 40, 604–634.

Wadhwa, M., Zheng, J., & Cook, T. (2021). What does "evidence-based" mean in actual research practice? An analysis of 22 clearinghouses disseminating information about evidence-based social programs. (Forthcoming).

Wilde, E., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. Journal of Policy Analysis and Management, 26, 455–477.

Wing, C., Simon, K., & Bello-Gomez, R. (2018). Designing difference in difference studies: Best practices for public health policy research. Annual Review of Public Health, 39, 453–469.

Wong, M., Cook, T., & Steiner, P. (2015). Adding design elements to short interrupted time series when evaluating national programs: No Child Left Behind as an example of pattern-matching. Journal of Research on Educational Effectiveness, 8, 245–279.

LIST OF META-ANALYZED PUBLICATIONS

Anglin, K., Wong, V., Wing, C., Miller-Bains, K., & McConeghy, K. (Forthcoming). Only time will tell? The validity of causal claims with repeated measures designs.

Ferraro, P., & Miranda, J. (2014). The performance on non-experimental designs in the evaluation of environmental programs: A design-replication study using a large-scale randomized experiment as a benchmark. Journal of Economic Behavior & Organization, 107, 344–365.

Hallberg, K., Williams, R., Swanlund, A., & Eno, J. (2018). Short comparative interrupted time series using aggregate school-level data in education research. Educational Researcher, 47, 295–306.

Jacob, R., Somers, M.-A., Zhu, P., & Bloom, H. (2016). The validity of the comparative interrupted time series design for evaluating the effect of school-level interventions. Evaluation Review, 40, 167–198.

Michalopoulos, C., Bloom, H., & Hill, C. (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory Welfare-to-work programs? The Review of Economics and Statistics, 86, 156–179.

Patnaik, A. (2019). Reserving time for daddy: The consequences of fathers' quotas. Journal of Labor Economics, 37, 1009–1059.

Schneeweiss, S., Maclure, M., Carleton, B., Glynn, R., & Avorn, J. (2004). Clinical and economic consequences of a reimbursement restriction of nebulized respiratory therapy in adults: Direct comparison of randomised and observational evaluations. BMJ, 328, 560.

St. Clair, T., Cook, T., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. American Journal of Evaluation, 35, 311–327.

St. Clair, T., Hallberg, K., & Cook, T. (2016). The validity and precision of the comparative interrupted time-series design: Three within-study comparisons. Journal of Educational and Behavioral Statistics, 41, 269–299.

Wichman, C., & Ferraro, P. (2017). A cautionary tale on using panel data estimators to measure program impacts. Economics Letters, 151, 82–90.

APPENDIX

Table A1. List of clearinghouses and dimensions of heterogeneity.

Name (URL)	Primary Field	Target Population	Funding Organization	Object of quality assessment	Whether they use same standards for study, programs, policies, practices?
Blueprints (https://www.blueprintsprograms.org/)	Multi-focus: social and behavioral development, education, health	Youth	Nonprofit	Programs	N/A
California Evidence Based Clearinghouse for Child Economic Welfare (https://www.cebc4cw.org/)	Multi-focus: social and behavioral development, health, economic welfare, education	Children and family	Public	Programs	N/A
Promising Practices Network (https://www.rand.org/well-being/social-and-behavioral-policy/projects/promising-practices.html)	Multi-focus: social and behavioral development, health, education, economic welfare	Children and family	Nonprofit	Programs	N/A
Best Evidence Encyclopedia (http://www.bestevidence.org/)	Education	Students	Public	Programs	N/A
National Dropout Prevention Center (http://dropoutprevention.org/)	Education	Students	Nonprofit	Programs	N/A

Table A1. Continued.

Name (URL)	Primary Field	Target Population	Funding Organization	Object of quality assessment	Whether they use same standards for study, programs, policies, practices?
Social Programs that Work (https://evidencebasedprograms.org/)	Multi-focus: education, health, economic welfare, labor, social and behavioral development	All	Nonprofit	Programs	N/A
Clearinghouse for Military Family Readiness - Continuum of Evidence (https://militaryfamilies.psu.edu/)	Multi-focus: health, education, social and behavioral development, labor, education	Military Family	Public	Programs	N/A
Collaborative for Academic, Social, and Emotional Learning Guide (https://casel.org/)	Education	Students	Nonprofit	Programs	N/A
Research-Tested Intervention Programs (https://rtips.cancer.gov/rtips/index.do)	Health	All	Public	Programs	N/A
Home Visiting Evidence of Effectiveness (https://homvee.acf.hhs.gov/)	Multi-focus: health, education, economic welfare, social and behavioral development	Families with pregnant women and children from birth to kindergarten entry (that is, up through age 5)	Public	Programs; study	Separate standards for study and programs

Table A1. Continued.

Name (URL)	Primary Field	Target Population	Funding Organization	Object of quality assessment	Whether they use same standards for study, programs, policies, practices?
What Works for Health Wisconsin (http: //whatworksforhealth. wisc.edu/)	Health	All	Nonprofit	Programs and policies	Yes
Conduent Promising Practices Database (https://cdc.thehcn.net/ index.php?module = promisepractice&controller = index&action = index)	Health	All	Private	Programs, policies and practices	Yes
Promise Neighborhoods Research Consortium (http:// promiseneighborhoods. org/index.html)	Multi-focus: education, social and behavioral development, health, economic welfare, education	Youth	Public	Programs, policies and practices	Only standards for policies
CrimeSolutions.gov (https: //www.crimesolutions. gov/default.aspx)	Social and behavioral development	All	Public	Programs and practices; study	Separate standards for study, programs and practices

Table A1. Continued.

Name (URL)	Primary Field	Target Population	Funding Organization	Object of quality assessment	Whether they use same standards for study, programs, policies, practices?
What Works Clearinghouse (https://ies.ed.gov/ncee/wwc/)	Education	Students	Public	Programs + Policies + Practices; study	Separate standards for study and (programs + policies + practices)
Clearinghouse for Labor Evaluation and Research (https://clear.dol.gov/)	Labor	All	Public	Study	N/A
Teen Pregnancy Prevention Evidence Review (https://tppevidencereview.youth.gov/)	Health	Youth	Public	Study	N/A
Strengthening Families Evidence Review (https://www.mathematica.org/our-publications-and-findings/projects/strengthening-families-evidence-review-sfer)	Multi-focus: labor, economic welfare, social and behavioral development	Children and family	Public	Study	N/A

Table A1. Continued.

Name (URL)	Primary Field	Target Population	Funding Organization	Object of quality assessment	Whether they use same standards for study, programs, policies, practices?
Employment Strategies for Low Income Adults Evidence Review (https://www.acf.hhs.gov/opre/resource/employment-strategies-for-low-income-adults-evidence-review-standards-and-methods)	Labor	Low-income adults	Public	Study	N/A
The Community Guide (https://www.thecommunityguide.org/)	Health	All	Public	Interventions	N/A
Cochrane Collaboration (https: //www.cochrane.org/)	Health	All	Nonprofit	Study; Interventions	Separate standards for study and interventions
Campbell Collaboration (https://campbellcollaboration.org/)	Multi-focus: education, social and behavioral development, economic welfare, health, labor	All	Nonprofit	Study; Interventions	Separate standards for study and interventions

Notes: Some clearinghouses have a major focus field while some clearinghouses have a focus on multiple fields but solely deal with a certain group of the population. We include a target population column to capture this information.