

NATIONAL
ACADEMY
of
EDUCATION

Workshop to Examine Current and Potential Uses of
NCES Longitudinal Surveys
by the Education Research Community

Testing Causal Hypotheses Using Longitudinal Survey Data: A Modest Proposal for Modest Improvement

Thomas D. Cook

Northwestern University & Mathematica Policy Research, Inc.



National Academy of Education

***Workshop to Examine Current and Potential Uses of NCES Longitudinal Surveys
by the Education Research Community***

**Testing Causal Hypotheses Using Longitudinal Survey Data:
A Modest Proposal for Modest Improvement**

Thomas D. Cook
*Northwestern University &
Mathematica Policy Research, Inc.*

December 2014

This paper was prepared for the National Academy of Education's *Workshop to Examine Current and Potential Uses of NCES Longitudinal Surveys by the Education Research Community*, held on November 5-6, 2013, in Washington, DC. The workshop was organized to provide input to the National Center for Education Statistics (NCES) on how its longitudinal survey program could be enhanced to serve the changing needs of the education research community in light of (1) changing technological opportunities, (2) changing data availability (with an increasing focus on the classroom and learning outcomes), and (3) a changing U.S. population. The collection of commissioned papers reflects the views of the authors and not necessarily those of the National Academy of Education or the U.S. Department of Education. This paper was supported with funding from the Institute of Education Sciences.

INTRODUCTION

One-shot surveys were developed to describe populations and longitudinal surveys were developed to describe changes in populations. Relative to the range of all questions that the community of educational researchers asks, the scientific function of surveys is limited, just as the scientific function of other methods is also limited. Surveys were not developed to describe or explain causal connections between interventions and outcomes. There are better methods for this, such as experiments, but they are generally weaker than the survey when it comes to population description. It is the research purpose that should determine method choice. But since science values many different kinds of purpose, no method can be superior to all the others for all the relevant purposes to which research is put.

This truism can get lost from view, buried by pragmatism, opportunism, and ungrounded hope. Surveys tend to be expensive and their results time dependent. The temptation is common, therefore, to seek to justify the use of surveys in terms of their ability to warrant causal knowledge as well as to describe populations and changes in them. The present paper is one of a long series dating back to the end of World War II in which attempts are made to probe how the use of surveys might be extended to include the testing of hypotheses about bivariate causal relationships.

The hurdle that surveys have to jump to do this becomes clear by comparison with the main methods that have been explicitly developed to test causal hypotheses. These causally privileged methods include (a) random assignment, as in the randomized controlled trial (RCT); (b) other studies with completely known processes of selection into treatment, such as the regression-discontinuity design (RDD); or (c) case-study designs with repeated treatment applications, removals, and reapplications at known times and in controlled settings under researcher control. Relative to these methods, one-shot surveys are bound to do a worse job of reducing uncertainty about a causal claim. Even longitudinal surveys will be worse since (1) there is rarely certitude that the causal treatment under investigation is exogenous to the process generating the study outcome, and (2) the time-varying causal counterfactual against which change in the treatment group is evaluated is not likely to be unbiased. Survey-based claims about causal relationships will never satisfy advocates of the methods explicitly developed to test causal hypotheses. Yet the need to justify survey budgets by claiming competency for causal purposes keeps reoccurring. To accommodate them is likely to mean lowering the consensual standards of evidence for inferring cause relative to those that currently hold in, for instance, the National Center for Educational Effectiveness or the What Works Clearinghouse.

There are a few exceptions to the assertion that surveys constitute a weaker reed for testing causal hypotheses. One is when a lottery takes place before or during survey waves. The data can then be used to provide measures of the study effects. But this circumstance is so rare as to provide no sound basis for claiming that surveys are useful for testing causal propositions. Another exception is when an RCT or RDD is deliberately introduced into a survey. This is desirable, but to date it has been mostly used to test hypotheses about the consequences of different survey methods or to test the effects of scenarios presented as vignettes. Yet we have little interest here in testing predictions about survey research practice or the hypothetical situations described in vignettes. The final exception is when survey results fortuitously provide outcome data

for an RCT or RDD that was designed independently of the survey. This can provide a rich source of often especially important outcome data. But many experiments are quite local in their reach and, to be useful, survey cases have to be sampled so densely within the RCT's or RDD's catchment area that data can be collected from all or most of those serving in the RCT or RDD. Again, this will rarely be the case.

As valuable as these exceptions are, none gets at what animates this paper—using broad survey data to test a substantive causal hypothesis within the context set by the respondents, variables, and times sampled in the survey. So we will not deal with the use of surveys to examine the consequences of lotteries, of variations in survey research practice and hypothetical vignettes, or with the use of surveys to provide outcome data for RCTs that were originally implemented outside of the survey context. Instead, the paper focuses on longitudinal survey methods for testing educationally substantive causal relationships.

If the paper has any innovation, it is to advocate for the use of a delimited set of non-experimental design (and, to a much lesser extent, analysis) practices identified through a method that is variously called a design experiment or a within-study comparison. Such studies test whether a set of nonexperimental practices results in causal estimates that are acceptably similar to those from an RCT that shares the same treatment group as the nonexperiment. What varies, then, is whether the comparison group was chosen randomly (as in the RCT) or nonrandomly (as in a nonexperiment). Identifying superior nonexperimental design and analytic practices through this method provides one form of external warrant for identifying those longitudinal survey circumstances that test causal hypotheses “adequately” well.

However, the within-study comparison method was designed for testing RCT and quasi-experimental results rather than RCT and survey results. Quasi experiments are like experiments in purpose and in all structural details other than random assignment. So they test hypotheses about exogenous causal agents deliberately introduced into an ongoing system. Surveys, on the other hand, often test hypotheses about less clearly exogenous treatments that might be products of the very causal hypothesis being tested. Moreover, quasi experiments strive for the same control over the research setting, the measurement plan, and the treatment application that characterizes RCTs. Surveys, on the other hand, typically provide less control over the kinds of covariates that can be used to adjust for selection bias; one is often limited by the kinds of variables collected in the survey. Surveys also make it difficult to find very local comparison groups that share many unmeasured attributes that might affect the main study outcome. Few sampling designs allow for dense data collection within local settings. To identify the specific design and analysis strategies that result in RCTs and quasi experiments having comparable causal estimates does not necessarily transfer to the survey context. Again this is because, unlike the quasi-experimental context, the survey context was not specifically designed to test causal hypotheses.

The generic internal validity limitations of the survey have to be weighed against their potential advantages for other kinds of validity. The intellectual taste of the moment is to prioritize on internal validity, buttressed by the assertion that *there is little point to generalizing causal claims if they might be wrong*. This constitutes the warrant for the primacy of internal validity. However, claims about internal validity are never assumption free, and these assumptions may be wrong. RCTs require assump-

tions whose justification is only imperfectly warranted, including assumptions about initial balance, nondifferential attrition, treatment contamination, statistical power, and chance. Moreover, even with RCTs, external validity is still a desideratum. No one wants to take a short-term experiment on self-esteem with college sophomores and generalize it to effects on self-regulation in elementary school students. One rationale for using survey data to test causal relationships is to extend external validity over more representative and more heterogeneous populations. Another is that surveys may often have superior statistical power, given their large sample size and repeat data collection waves. All research involves validity trade-offs. The current vogue is to optimize on internal validity, but if methods can be identified that have often reproduced RCT results and that can also be implemented in surveys, then the presumption is that the internal validity losses that survey methods engender might be tolerable because external validity is enhanced and even smaller effects can be detected. We preserve the current emphasis on internal validity but note that if we were to weight external and statistical conclusion validity more highly relative to internal validity, then the tone and recommendations in this essay would be quite different.

We have not yet described what we mean by causation. It obviously has many meanings (Cook & Campbell, 1979). This essay and the previous discussion use it in the limited sense of identifying the effects of manipulated (or potentially manipulable) causes. This allows us to identify the things we can deliberately vary to see what happens as a result. This sense of cause is quite elementary in evolutionary terms and corresponds with theories that philosophers of science would call “manipulability” or “activity” or “recipe” theories of causation. These are theories of causation in philosophy since they do not necessarily have anything to do with understanding cause in explanatory terms. The latter usually involves identifying either a multivariate system of variables that relate to each other before impacting on the effect of interest or identifying a highly general causal mechanism that is activated by multiple co-causal forces and that will often affect multiple outcomes. This last is the holy grail of science.

Explanatory causation mostly deals with identifying the causes of a given effect or related set of effects. The alternative is to describe the effects of a given cause, as occurs in all experimentation. Explanation commonly lends itself to methods like structural equation modeling rather than RCTs, and it is another area where longitudinal surveys have the edge over RCTs. But the problem is that multivariate explanatory methods are rarely—if ever—definitive, and there is currently a vogue for causal answers with little or no ambiguity, hence the apotheosis of RCTs and even of claims to understand quasi experiments in RCT terms. But few experiments are built to test all the links in an explanatory model, and none have longitudinal data collection calibrated to the known different times when various outcomes are supposed to change before the next outcome in sequence changes. A mismatch exists between the explanatory causation to which science aspires, and the testing of usually bivariate causal hypotheses that experimental structures allow. This paper is limited to testing bivariate causal hypotheses, given how prominent this task is in contemporary educational research.

In the space available, I cannot consider in detail any single example of a causal question, nor any specific National Center for Educational Statistics (NCES) longitudinal survey dataset. Yet each would be necessary for a *grounded* analysis of any real-world cause-probing research possibilities that NCES might undertake. Instead, I

content myself with a more general analysis of the quasi-experimental methods worth pursuing once suitable causal agents have been determined whose effects are to be assessed. The examples I provide are from research with archival data in general rather than with any single NCES dataset.

One other distinction needs elaboration before embarking on the main analysis. Longitudinal surveys have different kinds of sampling design with unique and important implications for causal hypothesis testing. The main survey data structures are (a) true longitudinal datasets—identical units assessed at different times; (b) true cohort datasets—repeat cross sections formed from the same population selected with known probability at each time point; and (c) opportunistic cohorts—groups not identical in composition over time and not randomly formed but nonetheless going through the same organization at different times (e.g., grade 4 samples year by year). Reluctantly, we call all of these “longitudinal surveys.”

To summarize, this essay assumes the following: (1) We are primarily interested in bivariate cause understood as identifying the effects of a given cause. (2) We assume that, if a named cause is sufficient for an effect in a given study, it will be only part of a larger constellation of co-causes that collectively explain why the effect was found. (3) RCTs should be done within surveys whenever this is possible, but that will be quite rare. (4) Many nonexperimental alternatives to RCTs exist, and their causal results are generally poorly identified. (5) Among nonexperimental methods, quasi-experimental alternatives are generally superior because they are predicated on exogenous variation, knowledge of the temporal sequence of intervention and outcome assessment, the availability of comparison groups, and identical measurement on both treatment and comparison units at both posttest and pretest. (6) Empirical knowledge is evolving of quasi-experimental options whose results coincide with RCT results. (7) This correspondence of results creates an external empirical warrant for identifying effective quasi-experimental practices that might also be applicable in some longitudinal survey contexts.

WHICH QUASI-EXPERIMENTAL PRACTICES HAVE CLEARER THEORETICAL OR EMPIRICAL WARRANTS?

Scholars agree that unbiased causal inference results when there is perfect knowledge, measurement, and modeling of that part of the selection process into treatment that is correlated with the study outcome. Such conditions meet the crucial assumption that is variously called the hidden bias, the strong ignorability, or the conditional independence assumption. It is obviously met with properly executed experiments, but also with sharp regression discontinuity (RD) and with other applications so rare that they cannot be relied on for a model of how to do quasi-experimental research (e.g., Diaz & Handa, 2006). As a result, among quasi-experimental methods only RD has a clear warrant in statistical theory. In other quasi-experimental applications, theory is not specific enough to identify when the strong ignorability assumption is met.

Where is an external warrant for quasi-experimental applications other than RD to come from? Traditionally, many social scientists have relied on an alternative warrant that has a background in falsificationism: namely, a causal claim is warranted when the focal research community can come up with no plausible alternative interpretations—a

causal claim then stands unless later proven otherwise. This is not always a satisfactory warrant by itself. For one, what is “plausible”—a rather “squishy” concept? Next, there have been periods in history where the unanimous current wisdom subsequently turned out to be wrong, as with phlogiston. And finally, the disputatious community of scholars often disagrees with new knowledge claims rather than endorses them. A different warrant is needed that, at a minimum, complements the falsificationist one.

Within-study comparisons (WSCs) seek to create an empirical warrant for generally unbiased causal inference. They do so by directly comparing an RCT effect size with that from an adjusted quasi experiment that shares the same treatment group. The mode of comparison group selection is thus the object of study. If the effect sizes from each design are similar, then the conclusion is drawn that the quasi experiment was unbiased in this particular application. Its causal result is, after all, similar to that from an RCT with the same treatment group.

There are currently two main WSC design variants. The three-arm design is in Figure 1. It is composed of the usual two RCT arms plus a third arm from a nonequivalent comparison group selected in some nonrandom way; how it is chosen is very important and is discussed later. The data from this third arm are often adjusted in some way to try to control for selection bias. This adjusted comparison group value is then compared to the value from the same treatment group as in the RCT, producing an adjusted effect size that is then compared to the effect size from the RCT. A judgment is then made about how similar the two effect sizes are. If they are similar enough, the conclusion is drawn that the quasi-experimental result is unbiased, or biased to a tolerably small level.

The four-arm WSC design variant is depicted in Figure 2. Respondents are first randomly assigned to participate in an RCT or a quasi experiment. Within each, there is a treatment group and a comparison group; they are created randomly in the RCT but by some selection process in the quasi experiment (e.g., by self-selection or administrator selection). Treatment cases are handled identically within the RCT and quasi experiment, as are comparison cases; and all treatment and comparison cases experience the same testing regimen. Comparing the RCT and adjusted quasi-experimental effect sizes creates a clean comparison of the bias-related consequences of whether the comparison group was formed randomly or not.

Four-arm designs are much rarer than three-arm designs, and we know of only 3 four-arm studies as opposed to about 30 three-arm studies. To date, the four-arm studies are very experimenter controlled and short lasting and take place in laboratory-like settings. So they are more like analog experiments than real-world experiments. Fortunately, three-arm studies are much more prevalent, and recent attempts have been made to improve their theory and design by laying down standards:

First, WSCs obviously require a well-scrutinized RCT if they are to function as a valid causal benchmark. Thus, a correct random assignment procedure has to have been correctly implemented, and treatment-correlated attrition and treatment crossovers have to have been “adequately” accounted for. Otherwise, the benchmark is bedeviled by both sampling error and small sources of bias whose cumulative impact it is difficult to assess but that manifestly reduce the validity of the indispensable causal benchmark.

Second, also required in a good WSC is the same causal estimand that is involved in the RCT and quasi experiment. There is little point, for example, to comparing the

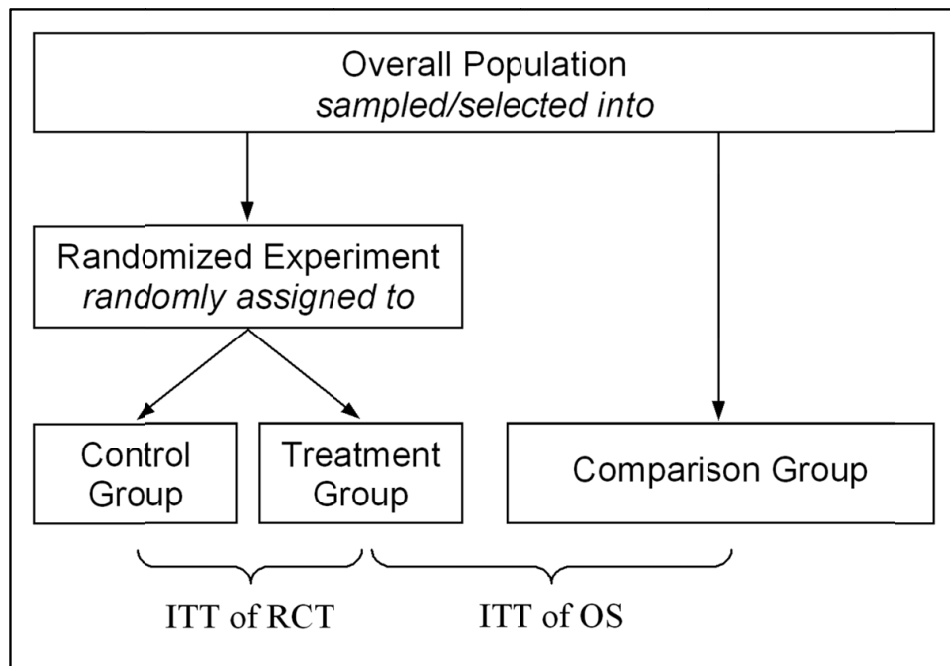


FIGURE 1 WSC three-arm design.

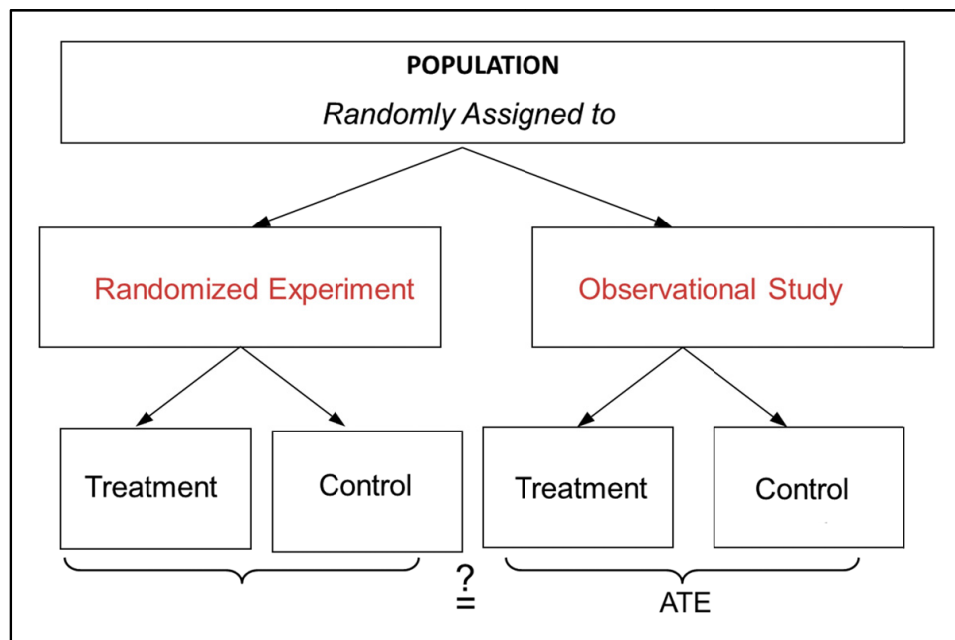


FIGURE 2 WSC four-arm design.

average treatment effects (ATE) from an RCT with the local average treatment effects (LATE) from some simple RD design.

Third, there should not be confounds from the randomized and nonrandomized comparison groups being measured in different ways or otherwise treated in ways that might affect the study outcome.

Fourth, the RCT and quasi-experimental data should be analyzed blind lest knowledge of the results of one analysis lead to additional tests with the other design that capitalize on chance.

And finally, there have to be clearly articulated standards about what level of correspondence is acceptable between the RCT and adjusted quasi-experimental results, given that (a) exact point correspondences are unlikely since the RCT and quasi experiment are each subject to sampling error; (b) identical statistical significance patterns are also unlikely since, with $\beta = .80$, the same pattern of significance would be obtained in only 68 percent of exact replications of the same RCT; (c) interobserver agreement is very unlikely about whether different effect sizes should lead to different policy implications; and (d) equivalence tests as used in medicine and pharmacology require very large sample sizes and assumptions about what (small) difference between effect sizes can be tolerated.

Since no perfect criterion is possible for adjudicating the meaning of RCT and adjusted quasi-experimental effect sizes, we compute and report estimates of the size of these differences in standard deviation (SD) units and then judge how tolerable they are. This last is difficult, even in the preferred context of synthesizing effect size differences across multiple studies rather than analyzing a single one. Nonetheless, we borrow standards from current practice in education where an effect size of .20 SD units is usually considered worth detecting, and so we consider an unadjusted bias of .20 as being unacceptable. This is somewhat arbitrary, of course, and so we remind readers that the absolute size of the differences is usually reported too.

The WSC standards described above have slowly evolved over nearly 30 years and are not yet firmly fixed. Older three-arm WSCs were developed before these standards and now seem embarrassingly quaint. Newer studies have moved closer and closer to meeting the standards. They have also moved away from asking whether quasi experiments can reproduce RCT findings. In theory, this is a trivial question because any quasi experiment that meets the strong ignorability assumption is unbiased and, pragmatically, empirical existence proofs of this are already on hand (e.g., Shadish et al., 2008). So the current orientation is toward examining the conditions under which closer approximations to the RCT results are achieved, entailing focused hypothesis tests rather than explorations of equivalence. What brings about a better approximation is now the key issue, though estimates of how closely RCT and quasi-experimental estimates track each other are inevitably also provided.

WSCs have other limitations than sampling error in both the RCT and the quasi experiment that makes exact point correspondences well-nigh impossible. Each WSC is a single case replete with its specific population, specific versions of the treatment and outcome, and specific setting and time details. Any one of these might affect the level of RCT and quasi-experimental correspondence and so preclude general learning about unbiased quasi-experimental alternatives. This is why heterogeneous replication of WSC results is very important, as is the use of formal hypothesis tests.

Another limitation is that WSCs can only be done on topics where an RCT has already been done, or, in the few cases where the causal standard is an RD estimate, where an RD has been done. But we most want to generalize to settings where an RCT cannot be done; that is when quasi experiments are most needed to test causal hypotheses. If topics on which an RCT can and cannot be done vary in characteristics promoting bias control, then WSC results will be distorting for the context where they are most needed—when only quasi experiments are feasible! No credible analysis of this conundrum currently exists; but if such an analysis strongly suggested that causal heterogeneity is expected along the lines just mentioned, then the utility of WSCs would be powerfully undermined.

Absent this analysis at present, in advising NCES on what can be done to promote stronger causal inference in longitudinal survey work, we rely on the results of WSCs to date, most coming from three- rather than four-arm designs. In particular, we seek to identify those quasi-experimental design practices that often promote causal estimates close to those from an RCT, and we then ask how can they be used with survey data over time. But first we have to examine which kinds of causal agents surveys can most productively examine. This is important because of the strong emphasis statistical theories of cause place on testing treatments as exogenous shocks into ongoing causal processes as opposed to testing treatments that are endogenous components operating within ongoing, time-dependent causal processes in which reciprocal causation might well be implicated.

LIMITING CAUSAL HYPOTHESES TO THE EFFECTS OF EXOGENOUS CAUSAL AGENTS

Within the constraints of the conception of causation we explore here, causal agents should be potentially manipulable and uncorrelated with errors in the outcome. This condition is most clearly met when evaluating the effects of exogenous shocks that are under researcher control and can be made to occur at a time that is known to be prior to assessing the study outcome. Meeting these requirements is easy in experiments and quasi experiments, but some of them cannot be readily met in longitudinal surveys, especially as concerns exogeneity and researcher control. Many variables one might like to consider as causes cannot be, such as fixed individual attributes like race and gender or endogenous variables embedded in complex, ongoing processes where mutual causal influence is likely. In survey work, the search for causal agents is constrained by the need to identify those that are clearly exogenous. Yet surveys were designed to describe populations, while experiments were developed to identify the consequences of deliberately manipulated causal agents. When population description is the priority, as in surveys, causal analysis is secondary; and when causal estimation is dominant, population description takes a back seat, however heroic are the attempts to describe the samples obtained. Expecting any method to perform tasks well for which it was not designed will likely result in disappointment at worst or, at best, in knowledge that is less secure than had the appropriate method been used for the type of question being asked. The key is to “somehow” develop and warrant standards of causal inference and of population description that are “good enough” or “satisficing,” even though they are not “perfect” or “as good as can be achieved with the best method currently available.”

(We take this last to be the RCT for cause, the survey for population description, and the longitudinal survey for descriptions of population change.)

Even so, in education research we can identify some important intervention possibilities that are most likely exogenous and amenable to analysis using archival data. Some examples include studies of the effects of Reading First (Somers et al., 2013), of the No Child Left Behind Act (NCLB) (Dee & Jacob, 2011; Wong et al., in press), and of being retained a grade (Hong & Raudenbush, 2005, 2006). Badly needed are longer lists of completed causal work with archival data where the treatment is clearly exogenous, plus attempts by small groups of researchers to identify other exogenous treatment shocks with important potential consequences that could be explored. The starting point in the Rubin causal model is the causal agent, not the data source. Unless this is recognized and the usual limits of the survey in this regard are acknowledged, seeking to use surveys for even “satisficing” causal work are likely to be limited.

Absent a list of causal agents reliably assessed in surveys, or a list of causal agents with known onset date and conditions of application, it will be very difficult to assess how relevant NCES surveys can be to national knowledge needs and how well “satisficing” answers can be generated. For what it is worth, my own guess is that the current uncertainty would be dramatically reduced and use of NCES datasets would be facilitated by (a) great clarity from NCES about its understanding of and commitment to exogenous cause; (b) NCES developing a provisional list of causal agents of interest to them; (c) providing funds for analysis; (d) streamlining procedures for data linking by individual student or school; and even (e) beginning workshops to teach young faculty and contract researchers about causal methods for use with surveys.

One attribute of causal agents is worth describing. Donald Campbell emphasized the importance of finding, describing, and evaluating what he called “outcroppings” and that applied microeconomists might call the preconditions for “natural experiments.” These are instances where local authorities (or individual school entrepreneurs) have created something that is different from current practice. He contended that these should be the most eager targets for analysis, especially when the onset date of the outcropping is known and the researcher can estimate how the change in units subjected to the outcropping can be judged relative to some estimate of change in comparable units—more about this comparison later. The counterexample is when units are compared without a sudden change at a known date. To evaluate such state differences is much more difficult than evaluating state changes. I know of no current efforts to create a systematic search for such outcroppings so as to guide researchers where and how to look. Without this, I would guess that attempts to use existing NCES datasets for causal purposes will remain sporadic and ad hoc.

However, identifying exogenous causal agents of substantive importance is only the first step. As implied above, also needed is a determination of what is “good enough” causal knowledge. So long as RCT knowledge remains the only acceptable standard—an assumption that is explicitly built into the Rubin causal model—nothing else can be tolerated. Even RD remains a distant second. While its causal inferences at the cutoff are unbiased, they are still less powerful statistically, less general along the assignment variable, and subject to more assumptions—about functional forms in parametric work and bandwidths in nonparametric applications. This paper adopts the criterion that “good enough” causal knowledge results from using quasi-

experimental methods that have often closely replicated experimental results in past high-quality WSCs.

REGRESSION DISCONTINUITY IN LONGITUDINAL SURVEY WORK

There is no doubt theoretically that sharp RD can result in unbiased causal inferences at the cutoff when a small number of testable assumptions are met, only one of which is deeply problematic: when deliberate manipulation of the treatment assignment scores has taken place so that the original assignment score values are unknown. The warrant from statistical theory is buttressed by the results of seven WSCs in different substantive areas within the social sciences. These all show causal claims at the cutoff that hardly vary between the RCT and simple RD design (Aiken et al., 1998; Black et al., 2005; Buddelmeyer & Skoufias, 2004; Gleason et al., 2012; Green et al., 2009; Shadish et al., 2011; Wing & Cook, 2013). More important, but tested only once to date, is the hypothesis that a comparison RD function formed by adding a pretest measure to the usual posttest-only RD (a) increases statistical power; (b) allows functional forms to be compared in the untreated part of the assignment variable; and (c) where the regressions are similar in form, then allows unbiased causal inferences in all the treated area away from the cutoff and not just at it (Wing & Cook, 2013). The implication is that researchers seeking to test causal hypotheses should use comparative RD (CRD) designs rather than simple RD whenever they can and should test whether the similar regression assumption is met in the untreated part of the assignment variable. Just as RCTs do not require pretests but they are nonetheless recommended, so pretests are worth recommending for RD studies even though they are technically not required.

The key question with RD and archived survey data concerns how feasible it is to identify situations where RD is clearly applicable. When can student, classroom, or school treatments be identified whose assignment is uniquely by some quantified indicator of academic merit or need, or by first come/first served, date of birth, social class, or anything else quantified where important outcome data can be collected from units above and below the cutoff? In the real world of allocating scarce educational resources, criteria like these are commonly used, and when one of them is the unique reason for treatment allocation then a sharp RD can and should be used. And it need not be the same cutoff in all sites, nor need there be just one cutoff per site.

However, in many real-world settings treatment allocation is not so sharp, whether by program design because exceptions are built into eligibility requirements or by happenstance—whereby local implementers “adapt” program requirements by supplementing with other assignment priorities. While we should do more in education to encourage sharp treatment assignment in the regular world of school practice, the reality for both of the above reasons is that assignment will often be fuzzy. For instance, students are retained a grade or enter a special education service in part due to prior performance on achievement tests but also in part due to teacher recommendations and even parent wishes. If individual study units can be assigned to each of these allocation processes, then the treatment assignment is still completely deterministic even if dependent on three treatment selection mechanisms rather than one. There is no bias problem then. But if only one of the allocation mechanisms is measured, say prior academic performance in a study of grade retention, then one option is to limit

the study to those sites using just this one allocation principle. Another but more problematic option is a fuzzy RD where the assignment variable is the prior performance measure, there is a cutoff score for treatment assignment, the unobserved other cutoffs are treated as sources of fuzz-inducing misallocation around the cutoff, and the binary intended treatment is then used as an instrumental variable for examining the effects of the treatment actually received. This situation gets much messier as the degree of treatment “misallocation” due to multiple assignment criteria increases and, as in an RCT, it is insurmountable if the treatment misallocation takes the form of deliberate manipulation that obscures the true observed assignment score.

It is easy to be sanguine about such RD procedures. In education, they have been very rarely used with data from a national archive as opposed to a very local one (e.g., Seaver & Quarton, 1976). So it is important to understand the theoretical and implementation experiences of those who have tried to pull off RD studies using data from national datasets. There are such examples in the literature (e.g., Ludwig & Miller, 2007); there are examples of those who have tried and failed to pull off a credible RD (e.g., the American Institutes for Research examined how the students with disabilities (SWD) requirements of NCLB affected academic performance of students with disabilities but the analyses they did were not included in the final report); and there are examples of strong research groups that are trying to get funds for ambitious RD studies with archival data (e.g., Dee & Jacob, 2011). It would be important to get together a work group of those who have tried to pull off an RD using national archives in education so as to better understand the conditions under which RDs are practical with archived survey data. What grounded wisdom do these pioneers have about practice? And how is this incipient wisdom to be shared with the larger body of researchers seeking to do high-quality work with archival rather than prospective data?

INTERRUPTED TIME-SERIES DESIGNS

This is a type of stronger quasi-experimental design that can often be used with longitudinal survey data when the need to evaluate the consequences of an intervention occurs at a known time prior to outcome assessment. This is because (1) it is easy to see if there is statistical regression because the intervention was a response to sudden prior change in performance. (2) Some time intervals are so short that it is not easy to come up with alternative interpretations based on “history”; events that co-occur with treatment and affect the outcome. (3) Even when intervals are longer, it is possible (and we see highly desirable) to supplement the interrupted time-series group with one or more comparison time series, perhaps even matched in terms of pretest means and slopes. (4) Information is routinely available about when outcome measures change in terms of items or the sampling design within which the data are collected, and when comparative interrupted time series (CITSs) are used, this problem is reduced anyway. (5) It is now easy to model the correlated errors that bias standard errors. And (6) there are now more data sets available that permit repeated pretest assessment, probably more so at the school than at the individual-child level since we have annual school data from all states but, in most states, we have student-level data at only certain grades.

The warrant for recommending CITS is that there are now five WSC studies comparing RCT and interrupted time-series (ITS) designs, of which four are CITS designs. All

five (Freitheim et al., 2013; Schneeweiss et al., 2004; Shadish et al., 2013; Somers et al., 2013; St. Clair et al., 2014) claim that the RCT and adjusted ITS results are very similar and, for what it is worth, our reading corroborates this. Although only two of these datasets are in education, and although there is not yet a meta-analysis or an analysis of the file drawer problem, the available WSC results show similar results and suggest a low likelihood of outcome-correlated historical events operating differentially around the intervention in the treatment and control conditions—usually the biggest worry in CITS designs. There is, then, an evolving external warrant for recommending CITS studies for testing causal hypotheses, though there are only two studies to date in education. We suspect that the frequent relevance of CITS has been overlooked in the current design environment in education. For instance, the What Works Clearinghouse does not yet have standards for CITS studies.

The viability of CITS can be quickly illustrated from data of St. Clair et al. (2014). Indiana started a new assessment system at the school level. Teachers were provided with regular feedback about the performance of individual students in anticipation that they would use this feedback to improve their instruction in math or language arts. Using data from the state's achievement monitoring system, schools that received this treatment in the program's second year were compared in two ways to comparison schools. First was a comparison to all the other state schools not involved in the innovation. In math, the observed selection process was such that schools receiving the intervention performed worse at pretest than the comparison schools, but stably so over time. In English Language Arts (ELA), however, the pretest difference between the treatment and comparison schools varied with time, and the schools selected for treatment were getting progressively worse compared to controls. (The two selection patterns are in Figures 3 and 4). In outcome models correcting for these different selection patterns, the resulting estimates were compared to those from an RCT with the same treatment schools. The results relative to the RCT show that one pretest time point suffices to account for the parallel math trend difference, but that the slope is required for language art where the two groups are growing apart prior to treatment. Indeed, using multiple time points without a slope reduces the amount of bias reduction relative to when the slope term is used. So the analysis depends on knowing the temporal selection differences that are fortunately directly observable. It also depends on there being no forces operating differentially by treatment condition at the intervention time point. The WSC comparison with the RCT shows that this was the case in this instance.

Estimating functional forms is not universally appreciated, even when they are observed. So St. Clair et al. also matched treatment schools over six years with a much smaller subset of comparison schools, again recreating estimates like those of the RCT. Somers et al. (2013) also used a CITS school-level matching strategy and again showed similar results for Reading First treatment schools when they were matched over time to those from an archival dataset and the results were then compared to RD (rather than RCT) estimates. All that is needed for CITS analysis, whether with or without matching, is knowledge of when a particular innovation took place in a specific set of schools and archives that permit finding comparison schools; more is given about this comparison group choice later. Such a strategy can be used with any outcome that is repeatedly observed in the data monitoring system, and extensions are viable when

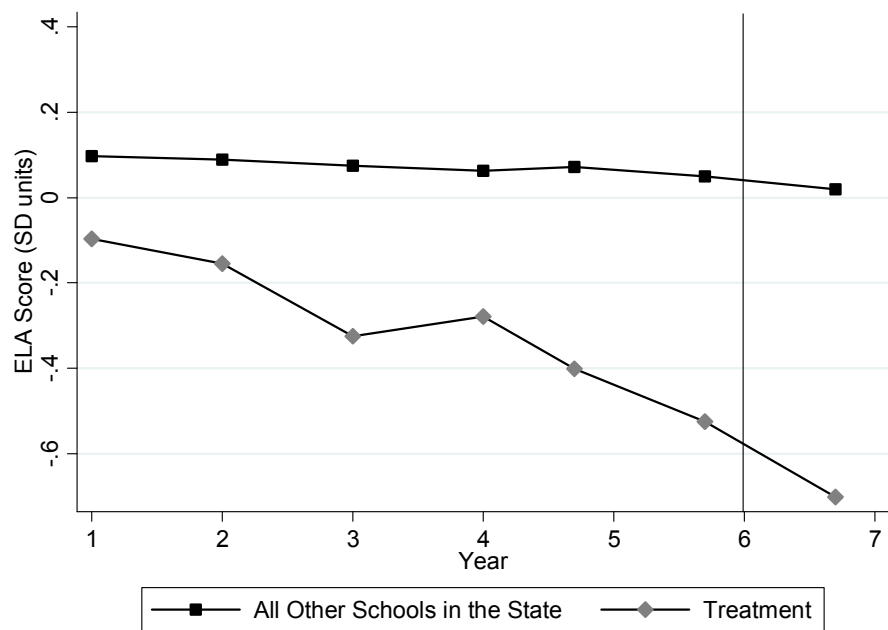


FIGURE 3 Treatment group vs all schools in the state, ELA scores.

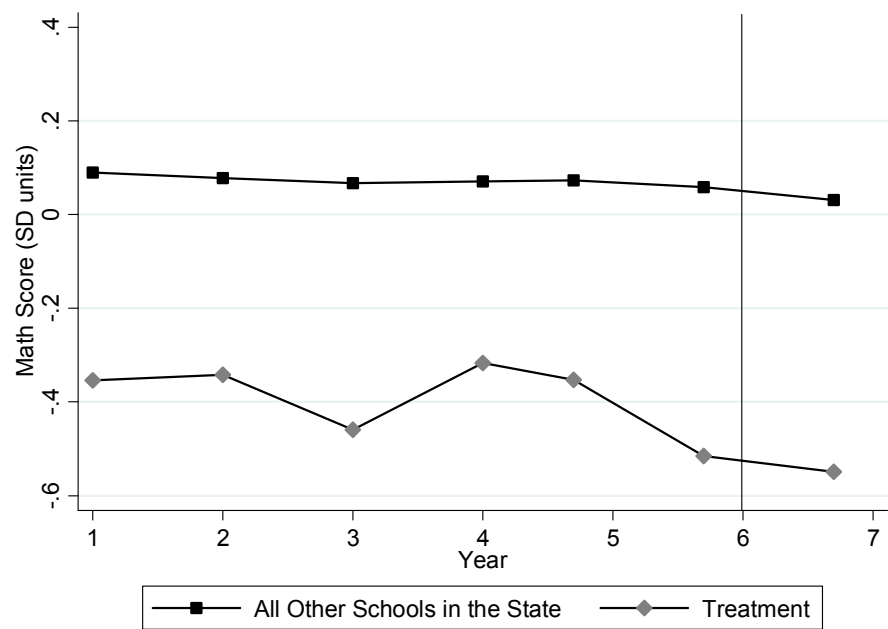


FIGURE 4 Treatment group vs all schools in the state, math scores.

individual student data are also available to add to the school level. Then, multilevel matching becomes a reality (Hallberg & Cook, 2013a).

Another example comes from research on NCLB, introduced in 2002. Dee and Jacob (2011) partitioned states into those with “consequential accountability” systems prior to NCLB and those only receiving such a system through NCLB. Using the grade 4 Main National Assessment of Educational Progress (NAEP), they showed that the two groups of states differed in pretest trend over four time points in a way favoring states with pre-NCLB accountability, but this slope difference changed after the intervention to favor the states newly getting an accountability system via NCLB. Only the presence of NAEP data over time made this analysis possible. The problems here are that this is not an evaluation of the NCLB’s national impact; the math effect may be due to chance, given largely undetectable eighth-grade math effects and nondetected reading effects at either grade. Wong, Cook, & Steiner (in press) used archival Main and Trend NAEP data to explore the same general NCLB issue, but now comparing pretest and post-NCLB time-series means and slopes with national public school data, first using Catholic and then non-Catholic private school data as controls. Every one of the tests show evidence of a change in both mean and slope; so do state-level analyses that partition states, not by their pre-NCLB “consequential accountability” status but by how high they set their standards for making annual yearly progress. The lower the new standards set, the less reform will be required, a criterion that is not correlated with state-level “consequential accountability.” Again, all the relevant data show a change in both intercept and slope. Twelve of them are independent, and they all show the same pattern of results, ruling out chance.

In summary, the CITS analyses showed that something happened in 2002 that affected school performance, most clearly in math. What alternative interpretations are there? One possibility is that the results are due to publicity about sexual abuse in Catholic schools that began in 2002. But we judge it to be implausible that such abuse affected the composition and performance of non-Catholic private schools nationally and was also more prevalent in states with less “consequential accountability” and in states “with higher testing standards.” The National Association of Math Teachers changed its standards in 2000. Can this have changed math performance more than reading two years later, and then more so in public than in private schools and in states with late consequential accountability and stricter testing standards after 2002? This too seems implausible. If only a professional association changing standards could have such a national impact! The third possibility involves deliberate manipulation of NAEP scores for which there was greater motivation in public schools nationally and in states forced into accountability provisions and selecting higher performance standards. The corollary argument is that, while the incentive to manipulate is clear for state achievement tests that directly link to adequate yearly progress (AYP) decisions, it is less so with NAEP tests and their much lower stakes. Nonetheless, the reality of state-level tests with high accountability consequences might have led to a change in school-level culture around testing that led to gaming all achievement tests, whatever their stakes. If there is any truth to this speculation, then it offers a rationale by which the pattern of manipulation might mirror the pattern of obtained achievement results in both of the NCLB papers using CITS methodology with archival achievement data.

I have provided this level of detail to make the obvious points that (a) had an RCT been possible to evaluate NCLB, there would be no need for recourse to these plausibility arguments. Quasi experiments always require more assumptions than experiments and hence usually more modeling, even if it is still less than in most explicit modeling exercises. However, (b) it is not easy to imagine a national RCT on the effects of a national program like NCLB. Some experimental scenarios are explored by Wong et al., but none seems politically or logistically feasible. So this brings us round to the thorny issue that dominates this paper. How can we establish standards for good enough causal inferences? Is it enough to say that a quasi-experimental method has been used that has often produced results similar to an RCT (as with CITS) and/or that judgments have been made by multiple independent reviewers that no alternative interpretations seem plausible among those that have been surfaced to date?

A second summary point is worth making. It is that much CITS research that uses archival data can exceed any reasonable RCT in statistical power. In the study by St. Clair et al. (2013), for example, the ability to have all schools in the state as the comparison added considerably to power, while matching each treatment school to a comparison group over time enabled four adequate comparison matches for each treatment school. Moreover, questions about national programs like NCLB do not lend themselves to RCTs, so it might be possible to answer a greater range of distinctly policy-relevant causal questions quasi-experimentally rather than experimentally. The advantage RCTs have with respect to internal validity might not turn into such an obvious advantage if designs were evaluated not by internal validity criteria alone, but also by external validity, construct validity, and statistical significance criteria.

The third conclusion about CITS is perhaps the most important. It is that school-level CITS studies on educational reforms are easy to conduct with archival data like NCES has. So will some kinds of child-level studies across all states and in those states like Florida, Texas, and North Carolina that already collect extensive child-level data on a routine basis. There will still be implementation issues, of course, but these are presently better understood for CITS than for CRD. Quasi-experimental results will never be quite as clear in internal validity terms as when an RCT is done. But since we have no standards yet for “good enough” causal inferences, narrow-minded advocates of RCTs can always denigrate CITS studies as being more assumption dependent. They are correct, but that is not the issue. The issue is what other observational studies are better than the ones we present here, given that no challenge is offered to the superiority of RCTs when only internal validity criteria are considered.

NONEQUIVALENT CONTROL GROUP DESIGNS WITHOUT A PRETEST TIME SERIES

The design we now consider is almost certainly the most common in educational research. Its basic structure entails the comparison of two or more nonequivalent groups on an outcome measure collected at both pretest and posttest. Much of the research using this type of design is very poor, and the task is to find the sweet spots where a close approximation to RCT results is likely because such a close correspondence has been multiply achieved in the past.

Many WSCs exist in this domain. Glazerman et al. (2003) meta-analyzed 12 in job training; but in education from kindergarten to college, we have only 10: Aiken et al. (1998), Agodini and Dynarski (2004), Wilde and Hollister (2007), Shadish et al. (2008), Pohl et al. (2009), Steiner et al. (2010), Bifulco (2012), Hallberg and Cook (2013a,b), Hallberg et al. (2013), and Fortson et al. (2012). (In other domains, we have Diaz & Handa, 2006; and Peikes et al., 2008). There are also some studies by Heckman and his students that are not formal WSCs but have several of their features.

There is some closure on which design features produce better approximations to RCT results. These include (a) having a pretest measure of outcome (Bifulco, 2012), (b) comparison groups that are local rather than distant—say, from within the same school district (Bloom et al., 2005; Heckman et al., 1997), (c) matching or covariance adjustments that are determined after careful analyses and measurement of the selection process into treatment (Diaz & Handa, 2006; Shadish et al., 2008), (d) matching strategies that use multiple covariates from many different domains and at many different levels, for example, school, classroom, and student (Hallberg & Cook, 2013b), and (e) matching or covariance adjustment strategies that use more reliable covariate assessments (Steiner et al., 2011). The problem is that, while each of these has sometimes reduced all of the initial bias, none has always done so. As a result, knowledge of some causes of bias reduction is more secure than knowledge of the causes of total bias reduction—except in the almost unheard-of case where the outcome-related selection model is fully known and perfectly measured. In the current absence of enough WSC studies to do a meta-analysis of the conditions under which close-enough approximations to RCT results are achieved, what practical advice is possible about how to do a nonequivalent control group design (NECGD) with archival data? Of course, no such advice is warranted unless three preconditions have been met. One is having a causal question; the second is identifying a presumably exogenous causal agent with known onset date; and the third is access to data sources about the outcome of interest. The discussion below assumes these three tasks have already been accomplished.

Creating Comparison Groups: The Hybrid Choice Model

Comparison group selection is the first step. The best current version of this that has been influenced by WSC work requires three coordinated steps. The first prioritizes on sampling local comparison groups, since this equates treatment and comparison schools on all those unobservables related to district-wide policy and practices that might affect the study outcome. There will also be related observed variables measured on each treatment and possible comparison case, of which prior pretest means will usually be one, but only one. It is possible to create estimated propensity scores from all these observed data at multiple levels, and multiple matches are potentially feasible for each treatment case. A caliper has to be set to distinguish tolerable from intolerable matches, and the tolerable matches are retained. The second step applies to matches exceeding the caliper limits. Here, the best nonlocal, focal matches have to be made on all the available covariates that correlate with selection, hopefully as a large and heterogeneous set of variables that are the product of explicit analyses of the selection process (note the plural)—of which more is discussed later. The final step

involves creating a dataset with local matches for some treatment cases and nonlocal but focal matches for those treatment cases where a local match is not possible. This hybrid sampling strategy is suggested by Stuart & Rubin (2007) and has been validated in at least one WSC study (Hallberg et al., 2013, whose study results are in Figure 19). Propensity score (PS) analysis is the obvious analytic tool here, the success of which depends on how well the covariates in the analysis capture the true selection process correlated with the study outcome (Cook et al., 2009).

Pretest Measures of the Outcome in Different Bias-Reducing Contexts

Many presentations of NECGDs focus on the importance for reducing selection bias of pretest measures of the outcome. They are indeed important. But four things have to be remembered about them. First, they are sometimes uncorrelated with the selection process, and there is a real case of this even in studies of educational achievement (St. Clair et al., 2014). Second, when they are correlated with selection, pretests may not be very well correlated so that additional covariates are needed too. Third, in some WSCs pretests alone have produced total bias reduction (Bifulco, 2012), although it is obviously dangerous to assume this result in any one application. And finally, some selection processes involving pretests depend on treatment and comparison group differences in slope rather than mean. Two (very reliable) measures of the pretest at different times allow some estimation of linear time-varying processes, but an ITS design is clearly preferable. Nonetheless, the bias reduction from a single pretest measure has often been considerable to date; the benefit of two measures is marginally better.

Strategies Using Multiple Covariates

Whether conducting prospective or retrospective NECGD studies, it is always useful to conceptualize the selection process into treatment. In this regard, it is especially important to be open to different theories of how selection took place, without assuming any one of them to be true. In archival work, this process alone can help ascertain how likely it is that the available covariates are adequate. It would be counterproductive, for instance, to assume that what is available is necessarily adequate just because some particular data-analytic method like propensity score analysis is used. How the data are analyzed seems to play at best a minor role in bias reduction. To illustrate this, consider Hong & Raudenbush (2005, 2006), who examined how grade retention affected subsequent achievement. Most analyses of retention make it primarily a function of prior academic achievement performance and of teacher judgments about such performance, sometimes invoking levels only and sometimes both levels and rates of change. The covariates Hong and Raudenbush used in the individual-level analysis included achievement in math and literacy as well as teacher evaluations of math and literacy measured at two time points prior to retention. Since there can be no a priori guarantee that these variables fully capture the entire selection process, Hong and Raudenbush also included 136 other measures from the Early Childhood Longitudinal Study–Birth Cohort about a wide variety of heterogeneous covariates assessing attributes of the student, of the school, of parents, and of neighborhoods. As it happened, two waves of either achievement data or teacher estimates of achievement

were enough to recreate the results of all 144 variables; the other 136 variables added nothing (Hallberg & Cook, 2013a). But without these 136 variables it would have been impossible to learn this, so the 144 were justified. Undertaking several different theoretical analyses of selection helps researchers to judge how adequate the available archival dataset is and, if the appropriate variables are not on hand, cautions against advancing with the desired nonexperimental study. However, if many covariates from multiple domains are available, this by itself has sometimes led to well-nigh total bias reduction.

Having Covariates at Multiple Levels, Especially the School and Student Levels

With school-level interventions, comparison cases were so often so closely matched in the past that no need existed for subsequent matching at the student level; this could not improve the match (Cook et al., 2008). But adding school-level variables exists as an option when school-level matches are inadequate and even for partially dealing with any unobservables that might still threaten causal inference in order to ensure that the populations of treatment and comparison schools are more comparable. However, this creates some decrement to external validity since generalization is only possible to those students in a school who can be matched, and thus not to the school at large. Figures 21 and 22 show results from WSCs varying the availability of school- and student-level data. In these applications, the school-level covariates do a good job, but the student-level data sometimes add to the bias reduction achieved and bring it close to total bias reduction, thus recreating the experimental estimate. Whether the intervention is at the school, classroom, or student level, there is always a case for multilevel covariate selection.

Mode of Data Analysis

Theoretically, propensity score analysis is to be preferred over ordinary least squares OLS analysis because the analysis is nonparametric, it avoids “the curse of dimensionality,” and it creates complete group overlap and so no extrapolation is required. Nonetheless, analysts of WSC studies in different substantive fields who have explored both PS and OLS methods have come to the unanimous conclusion that it has not mattered to date how the analysis was done. Quasi experimentation is more about (a) the study sampling design with respect to how treatment and comparison samples are chosen for initial comparability, (b) about the study measurement plan with respect to the kinds of covariates, and, to a lesser degree, (c) on the reliability of the most important covariates.

CONCLUSION

I cannot say anything about the adequacy of the NCES datasets for meeting the criteria above for describing better quasi-experimental tests of causal hypotheses. In part, this is because I do not know the datasets well; but it is also in part because the adequacy of selection controls is heavily tied to (1) particulars of the causal hypothesis under test, (2) the quality of the information for constructing comparison schools or comparison students or both, (3) the relevance of the available data for mirroring various conceptions of the treatment selection process, (4) the data quality for select-

ing other covariates that might be correlated with the selection process, and (5) the reliability of the covariate measures achieved. Only with a very specific causal question at hand can one responsibly explore the likely adequacy of longitudinal data for answering the question at hand with externally warranted support for the accuracy of the answer provided.

Short of that, all one can do is elaborate what seem today to be the best procedures for generating “satisficing” or “adequate” causal answers. We do not argue that the methods we summarize below are “as good as” an RCT from an internal validity perspective. Nor do we argue that they will always meet the key statistical assumption of ignorability. Instead, our criterion for the advice we summarize is that it emanates from research that identifies what usually “works” in quasi-experimental design and analysis to recreate a causal answer that is similar to an RCT’s on the same topic.

We assert that causal identification satisfies when it is based on (1) seeking out “kinks” in distributions where there is a sudden and dramatic change in the probability of treatment, of which the extreme is RD; (2) collecting pretest data or data from non-equivalent comparison groups so as to generate a no-treatment comparison regression function in CRD; (3) collecting data at multiple time points prior to whatever causal agent is under study and doing the same for a comparison group not exposed to the intervention, thus forming a CITS design; (4) selecting comparison cases so that they have maximal overlap in those physical space attributes that are related to the main study outcome, as with schools within the same district or even students within the same school (thus, local matching); (5) thinking through various scenarios whereby some study units are exposed to treatment and others not and then measuring well the attributes so identified (thus, theory-based covariate selection). In addition, but alternatively if need be, we recommend (6) collecting data on a wide range of covariate constructs at different levels that should include a pretest measure of the study outcome and where each covariate should be assessed with multiple items and at more than one time and (7) moving to a hybrid comparison group matching process for those cases where local matching fails to meet prespecified criteria for an acceptable level of matching and a “rich” and selection-theory-informed set of covariate measures is available to be used instead.

The bottom line, though, is that “design rules,” and that different designs prioritize on different kinds of research question. Surveys were designed to describe populations; experiments were designed to test causal hypotheses. Given this fundamental difference, surveys can at best probe causal hypotheses; they will rarely test them as well as an RCT. Of course, consensual standards might change about what constitutes “satisficing” rather than optimal causal tests, or they might change about whether internal validity deserves to be so paramount among all the other validity criteria by which research is judged. The debates needed to change such normative beliefs would cause considerable dissension, not just among scholars but also within federal agencies where some but not all divisions have adopted the current internal validity and RCT priorities. Future debates about shifting validity priorities may come; but since they are not reality, today surveys will continue to be seen as poor substitutes for experiments and experiments as poor substitutes for surveys.

REFERENCES

- Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *The Review of Economics and Statistics*, 86(1), 180–194.
- Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, 22, 207–244.
- Bifulco, R. (2012). Can nonrandomized estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. *Journal of Policy Analysis and Management*, 31(3), 729–751.
- Black, D., Galdo, J., & Smith, J. C. (2005). Evaluating the regression discontinuity design using experimental data (Working paper). Available from <http://www.economics.virginia.edu/sites/www.economics.virginia.edu/files/econometrics/smith.pdf>
- Bloom, H. S., Michalopoulos, C., & Hill, C. J. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In H. S. Bloom (Ed.), *Learning more from social experiments* (pp. 173–235). New York: Russell Sage Foundation.
- Buddelmeyer, H., & Skoufias, E. (2004). *An evaluation of the performance of regression discontinuity design on PROGRESA*. Washington, DC: World Bank.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies often produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27(4), 724–750.
- Cook, T. D., Steiner, P. M., & Pohl, S. (2009). Assessing how bias reduction is influenced by covariate choice, unreliability and data analytic mode: An analysis of different kinds of within-study comparisons in different substantive domains. *Multivariate Behavioral Research*, 44, 828–847.
- Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30(3), 418–446.
- Diaz, J. J., & Handa, S. (2006). As assessment of propensity score matching as a nonexperimental impact estimator: Evidence from Mexico's PROGRESA program. *The Journal of Human Resources*, XLI(2), 319–345.
- Fortson, K., Verbitsky-Savitz, N., Kopa, E., & Gleason, P. (2012). Using an experimental evaluation of charter schools to test whether nonexperimental comparison group methods can replicate experimental impact estimates. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Fretheim, A., Soumerai, S. B., Zhang, F., Oxman, A. D., & 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(8), 883–887.
- Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy*, 589, 63–93.
- Gleason, P. M., Resch, A. M., & Berk, J. A. (2012). *Replicating experimental impact estimates using a regression discontinuity approach*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.
- Green, D. P., Leong, T. Y., Kern, H. L., Gerber, A. S., & Larimer, C. W. (2009). Testing the accuracy of regression discontinuity analysis using experimental benchmarks. *Political Analysis*, 17(4), 400–417.
- Hallberg, K., & Cook, T. D. (2013a). Empirically examining the performance of approaches to multi-level matching to study the effect of school-level interventions. Manuscript in preparation.
- Hallberg, K. & Cook, T. D. (2013b). The role of pretests in education observational studies: Evidence from empirical within study comparisons. Manuscript in preparation.
- Hallberg, K., Wong, V. C., & Cook, T. D. (2013). How close is close enough? An empirical investigation of local focal intact group matching. Manuscript in preparation.
- Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 605–654.
- Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, 27(3), 205–224.
- Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101, 901–910.

- Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, 122(1), 159–208.
- Peikes, D. N., Moreno, L., & Orzol, S. M. (2008). Propensity score matching: A note of caution for evaluators of social programs. *The American Statistician*, 62, 222–231.
- Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, 31(4), 463–479.
- Schneeweiss, S., Maclure, M., Carleton, B., Glynn, R. J., & Avorn, J. (2004). Clinical and economic consequences of a reimbursement restriction of nebulised respiratory therapy in adults: Direct comparison of randomised and observational evaluations. *The BMJ*, 328(7439), 560.
- Seaver, W. B., & Quarton, R. J. (1976). Regression discontinuity analysis of dean's list effects. *Journal of Educational Psychology*, 68(4), 459.
- Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484), 1334–1344.
- Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods*, 16(2), 179.
- Shadish, W. R., Rindskopf, D. M., & Boyajian, J. G. (2013). A within-study comparison of results from single-case designs to a randomized experiment. Manuscript in preparation.
- Somers, M., Zhu, P., Jacob, R., & Bloom, H. (2013). The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation (MDRC working paper in research methodology). New York: MDRC.
- St. Clair, T., Cook, T. D., & 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(3), 311–327.
- Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36(2), 213–236.
- Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15(3), 250–267.
- Stuart, E. A., & Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, 33(3), 279–306.
- Wilde, E. T., & Hollister, R. (2007). How close is close enough? Testing nonexperimental estimates of impact against experimental estimates of impact with education test scores as outcomes. *Journal of Policy Analysis and Management*, 26(3), 455–477.
- Wing, C., & Cook, T. D. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison. *Journal of Policy Analysis and Management*, 32(4), 853–877.
- Wong, M., Cook, T. D., & Steiner, P. M. (in press). 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*.