Estimating Causal Effects Using School-Level Data Sets

by Elizabeth A. Stuart

Education researchers, practitioners, and policymakers alike are committed to identifying interventions that teach students more effectively. Increased emphasis on evaluation and accountability has increased desire for sound evaluations of these interventions; and at the same time, school-level data have become increasingly available. This article shows researchers how to bridge these two trends through careful use of school-level data to estimate the effectiveness of particular interventions. The author provides an overview of common methods for estimating causal effects with school-level data, including randomized experiments, regression analysis, pre–post studies, and nonexperimental comparison group designs. She stresses the importance of careful design of nonexperimental studies, particularly the need to compare units that were similar before treatment assignment. She gives examples of analyses that use school-level data and concludes with advice for researchers.

Keywords: effectiveness study; observational study; propensity score; treatment effect

Education researchers, practitioners, and policymakers alike are committed to identifying interventions that teach students effectively in a variety of areas. Their search involves questions such as: What interventions offer the best reading curricula? Which ones have the potential to improve academic performance in high-poverty schools? How can school libraries be used more effectively? School-level data sets, which are becoming increasingly available, can be used to address these types of questions about the effects of school- or district-level interventions. Especially with increasing emphasis on rigorous evaluation (U.S. Department of Education, 2003; Whitehurst, 2004), it is important for researchers to clearly understand how school-level data can (and cannot) be used in estimating causal effects and the issues involved in doing so. However, to date, there has not been a clear exposition of the assumptions and methods appropriate for this setting. This article discusses those assumptions and methods, providing examples and advice regarding how researchers can make the most of the available data to learn "what works."

One such data set, the National Longitudinal School-Level State Assessment Score Database (NLSLSASD), which is now being assembled by the American Institutes for Research for the U.S. Department of Education, contains assessment data on approximately 80,000 public schools across the country. This public use data set is available at http://www.schooldata.org and currently (as of spring 2007) contains data from 1993 through 2003, with the exact years available varying by state. The data have also been merged with the Common Core of Data to provide other, more basic information on the schools, such as the percentage of students eligible for free or reduced price school lunch. When used carefully, and particularly when supplemented with more information on school characteristics, the NLSLSASD is a rich and useful data set for estimating the causal effects of education interventions at the school level. However, despite the availability of this rich resource, relatively few studies have used the data; and as of spring 2007, only one appears to have been published in a peer-reviewed journal. Some of the studies that have used the NLSLSASD and other potential uses are discussed later.

Estimating Causal Effects Using School-Level Data Sets

This article presents the now common potential outcomes framework for estimating causal effects, emphasizing the assumptions necessary to estimate these effects and how the framework and assumptions relate to school-level data sets. Many of the ideas discussed here are the same as for student-level analyses. The article represents an effort to clarify the differences and the implications for school-level analyses. It also discusses common methods for estimating causal effects and provides some examples of the types of analyses now based on the NLSLSASD. It concludes with advice for researchers.

A Framework for Defining Causal Effects

The framework presented here is based on a formulation by Neyman (1923/1990) and Fisher (1925) for randomized experiments, which was extended to observational studies by Rubin (1974, 1978). Rather than discussing the general concept of causation (e.g., Pearl, 2000), this article follows the tradition of Neyman, Fisher, and Rubin in focusing on estimating the effects of particular interventions. In this framework, the three building blocks for defining causal effects are (a) treatment and control conditions, (b) units, and (c) potential outcomes.

Treatment and Control Conditions. Much of the current research in education is intended to measure the effect of some intervention (the treatment) relative to a different intervention or no

Educational Researcher, Vol. 36, No. 4, pp. 187–198
DOI: 10.3102/0013189X07303396

intervention at all (the control). Although in general the treatment is easily defined because it is the intervention of primary interest, the definition of the control may be more elusive. For instance, in a study estimating the effect of a new reading curriculum, to what do we want to compare the new curriculum? To the previous curriculum? To a different new reading curriculum? To no instruction in reading? Each study may define the control condition slightly differently, and so having a clear understanding of the control condition of interest will assist both in designing a study that estimates causal effects and in interpreting the results.

Units. The units are "entities" to which a treatment is assigned at a particular time. A new reading curriculum for example may be assigned at the beginning of a school year to an entire school, certain classrooms within a school, or individual students. When the units are schools or school districts, school-level data sets such as the NLSLSASD can be used to estimate the causal effects of an intervention at the school or district level. However, without student- or classroom-level data, these data sets cannot be used to estimate the effect of interventions applied at the student or classroom level.

Having a clear understanding of the appropriate unit of analysis prevents us from drawing conclusions at an incorrect level (e.g., at the student level when schools were randomly assigned to treatment conditions). In data sets with only school-level data, the units must be schools or other higher-level units such as districts. These school-level effects must be interpreted carefully in that researchers cannot assume that the relationships observed at the school level also apply at the student level. The tendency to inappropriately apply relationships observed at a higher level (e.g., results for schools) to a lower level (e.g., results for students) is known as the "ecological fallacy" or Simpson's Paradox (see e.g., Freedman, 2001).

Potential Outcomes. The causal effect of an intervention on a particular unit is a comparison of two "potential outcomes": the outcome if the unit receives the treatment and the outcome if the unit receives the control (Rubin, 1974). For instance, assume that the new reading curriculum is assigned at the school level and that the control is the existing reading curriculum used throughout a state. The potential outcome under the treatment condition (the new reading curriculum), often denoted $Y(1)$, is the outcome (e.g., average test scores at the end of the school year) that a school would experience if it were using the new reading curriculum. Likewise, the potential outcome under control, $Y(0)$, is the average test scores a school would experience were it using the existing reading curriculum. The "fundamental problem of causal inference" (Holland, 1986) is that for each unit, we can observe only one of these potential outcomes because each unit receives either treatment or control.[1] Causal inference is thus inherently a missing data problem, where at least half of the values of interest (the potential outcomes) are missing.

Concepts for Learning About Causal Effects

Given that we can never observe both of the potential outcomes for a particular unit and that therefore unit-level causal effects are never observed, how can we learn about causal effects? The previous section presented the framework for defining causal effects; this section discusses the three key concepts required to estimate those effects: replication, stability, and the assignment mechanism.

Replication. Replication means that there are multiple units for which we can observe one of the potential outcomes. If there were only one unit (e.g., one school) that received either treatment or control, we would have no information on the missing potential outcome. However, with some units receiving the treatment and others the control, we can use the treated units to learn about the potential outcomes under treatment and the control units to learn about the potential outcomes under control. In the school-level setting, this means that we must have a set of schools that receives the treatment and a set of schools that receives the control.

Stability. The stability assumption, also known as the stable unit treatment value assumption (SUTVA), has two components. The first is that no unit's potential outcomes are affected by the treatment received by any other units. In other words, there is no interaction between units. This assumption is sometimes difficult to believe in educational settings. For instance, if the units are classrooms and if students in control classrooms interact on the playground with students in treatment classrooms, the treatment may "spill over" to the control students, thus affecting their outcomes. In many cases, this component of SUTVA may in fact be more reasonable in school-level analyses than in student-level analyses. However, we could see a similar spillover effect in teaching methods across schools in the same district if teachers delivering the treatment curriculum interact with teachers delivering the control curriculum. A carefully designed study can preclude the effects of this interaction, for example, by choosing treatment and control schools in different districts or by otherwise preventing communication between treatment and control units. However, the implications of these design choices must be carefully thought through.

The second component of SUTVA is that there are no "versions" of the treatment or of the control: The same treatment is administered to all units in the treatment group (and likewise for the control group). This component of SUTVA may be more questionable in school-level settings, where the implementation of a particular program may vary considerably across schools. In that case, it may help to define the "treatment" as the set of services given to schools to help them implement the program, with some understanding that there will be variation in how schools use those services and actually carry out the program.

When SUTVA holds, we can express the true or ideal data in the form shown in Table 1, where each row is a unit and the columns represent the covariates—characteristics of the schools measured before the treatment was received—and all potential outcomes. Of course, we can never observe both potential outcomes for any unit, and thus this complete set of data is never fully observed; rather, it is the ideal data that we would like to be able to observe. In this example the impact is represented as the difference between the two potential outcomes; in other situations the impact might be the ratio or some other function of the potential outcomes.

The Assignment Mechanism. The assignment mechanism determines how units are assigned to the treatment and control groups and thus which potential outcomes are observed. Random assignment is an example of an assignment mechanism that has particularly desirable properties, as discussed in the following section. In observational studies however, in which the treatment is not

Table 1
The "Truth": The (Unobservable) Ideal Data for Causal Inference

	Covariates	Potential Outcomes		
Unit	*X*	*Y*(0)	*Y*(1)	Impact
1	X_1	$Y_1(0)$	$Y_1(1)$	$Y_1(1) - Y_1(0)$
2	X_2	$Y_2(0)$	$Y_2(1)$	$Y_2(1) - Y_2(0)$
3	X_3	$Y_3(0)$	$Y_3(1)$	$Y_3(1) - Y_3(0)$
4	X_4	$Y_4(0)$	$Y_4(1)$	$Y_4(1) - Y_4(0)$
. . .	. . .	. . .	. . .	. . .
N	X_N	$Y_N(0)$	$Y_N(1)$	$Y_N(1) - Y_N(0)$

Table 2
The Observed Data: The Data Available for Causal Inference

	Covariates	Treatment Assignment	Potential Outcomes		
Unit	*X*	*T*	*Y*(0)	*Y*(1)	Impact
1	X_1	0	$Y_1(0)$	—	—
2	X_2	1	—	$Y_2(1)$	—
3	X_3	0	$Y_3(0)$	—	—
4	X_4	0	$Y_4(0)$	—	—
. . .	. . .	. . .	. . .	. . .	. . .
N	X_N	1	—	$Y_N(1)$	—

Note. Cells containing dashes represent data that cannot be obtained.

assigned randomly and the true assignment mechanism is unknown, we must posit a hypothetical assignment mechanism. In other words, how were some schools selected (either by themselves or by an outside entity) to implement the intervention while other schools were not selected?

A key assumption in observational studies is that of strongly ignorable treatment assignment (Rosenbaum & Rubin, 1983b), which implies that (a) treatment assignment is independent of the potential outcomes given the observed covariates and (b) there is a positive probability of receiving each treatment for all values of the observed covariates. Therefore, under this assumption, conditional on the observed covariates, there are no differences between the treatment and control groups on unobserved covariates that are correlated with the potential outcomes. Analyses of sensitivity of the results to this assumption can and should be performed. For example, there may be concern that schools that choose to implement a particular program differ from those that do not and that the two sets of schools differ not only on observed characteristics, such as test scores and enrollment, but also on unobserved characteristics, such as the underlying organizational health of the school or the motivation level of the principal. Sensitivity analyses, such as that described by Rosenbaum and Rubin (1983a), assess the extent to which such an unobserved variable might affect the conclusions about the effect of the intervention.

As a result of the assignment mechanism, we are left with "the fundamental problem of causal inference" and do not observe the full data as shown in Table 1. Instead, we see only a subset of those data, as shown in Table 2, where at least half of the potential outcomes are missing. It is from data such as these that we must learn about causal effects.

Methods for Learning About Causal Effects

This article considers three common methods for estimating causal effects: (a) randomized experiments, (b) pre–post (or before–after) studies, and (c) comparison group designs (or observational studies).

Randomized Experiments

Although not without their own complications, randomized experiments are considered to be the gold standard of causal inference. A randomized experiment is characterized by a known assignment mechanism, where units are randomly assigned to treatment or control groups on the basis of observed covariates. Randomization yields two clear benefits: (a) The treatment assignment is known to be unrelated to the potential outcomes (and thus ignorable), and (b) the treatment and control groups will be "balanced" on all covariates, at least in large samples. In other words, before treatment assignment there are no systematic differences between the treatment and control groups, and thus any differences observed in the outcomes must be due to the treatment and not due to any (observed or unobserved) preexisting differences between the groups.

Because of these benefits, randomized experiments are becoming increasingly common in education research (U.S. Department of Education, 2003; Whitehurst, 2004). Although students are often the unit of random assignment, a number of studies have also been conducted in which schools or districts are randomized to treatment conditions. These include an evaluation of Comer's School Development Program (Cook, Murphy, & Hunt, 2000) and the National Study of the Effectiveness of Reading Comprehension Interventions, currently being carried out by Mathematica Policy Research, Inc. under contract with the U.S. Department of Education.

There are two ways in which data sets such as the NLSLSASD may be useful in randomized experiments. The first is in the design of the study, as a source of information on schools being considered, to identify schools that meet some entry criteria, or to design a study with matched pair or stratified randomization. The second is in the analysis of the outcomes, as a way to obtain outcome variables, such as state assessment test scores, at relatively low cost. Although the most obvious use is for school-level analyses where schools are randomized, both of these uses may also be appropriate when the randomization is of students rather than schools. For example, in an evaluation of the Moving to Opportunity experiment, which randomly selected families to receive a housing voucher to encourage them to move to a new neighborhood with lower crime rates, Sanbonmatsu, Kling, Duncan, and Brooks-Gunn (2006) used the NLSLSASD to obtain school-level data on the neighborhood schools that the students end up attending.

Randomized experiments are discussed in detail in other sources (e.g., Shadish, Cook, & Campbell, 2002). Because it is likely that most of the analyses based on school-level data sets such as the NLSLSASD will be nonexperimental, the rest of this article focuses on nonrandomized designs.

Pre–Post (or Before–After) Studies: An Illustration

Another common method for estimating causal effects is to use a pre–post or before–after study, where an outcome such as test scores measured after the treatment is compared with the test scores before the treatment and any difference in the scores is attributed to the treatment. For example, in a study of whether increased funding for a school library increased test scores, the school test scores at the end of the school year might be compared with test scores in the previous year. Analyses such as these need to be done very carefully, with clear understanding of the assumptions underlying the method.

To illustrate the pitfalls of simple pre–post comparisons, we use an example originally posed by Frederick Lord in 1967 and clarified by Holland and Rubin in 1983. This discussion has often been invoked in debates on the use of covariance adjustment versus gain scores in education research, although it perhaps more clearly shows the fundamental importance of the assignment mechanism in causal inference. Consider the following observation by Lord:

> A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded. (p. 304)

Lord goes on to say that the distribution of male weights is the same in September and June, and the distribution of female weights is also the same in September and June: The average weights for males and female students do not change and neither do their variances. Lord then posits two statisticians who come to two apparently contradictory conclusions about the differential effects of the dining hall diet.

Statistician 1 observes that there are no differences in the weight distributions between the beginning and end of the school year for men or women and so concludes that there is no differential effect. In particular, Statistician 1 uses the difference in mean weight gains for men and women as the estimated differential causal effect and determines that there is no differential causal effect because neither group gains or loses weight.

Statistician 2, in contrast, uses regression adjustment to compare the average June weights of men and women with the same September weight. Statistician 2 uses the following linear model: $E(Y|X, G=g) = a_g + bX$, where Y is June weight, X is September weight, and G is the gender of the student (1 = male, 0 = female). This model assumes that the regressions of June weight on September weight for men and women are linear and parallel to one another. Statistician 2 estimates the differential causal effect as $a_1 - a_0$, which is the covariance adjusted mean difference in June weights. Statistician 2 thus concludes that because a man will weigh more in June than will a woman of the same initial weight (i.e., $a_1 > a_0$), the diet must have a larger effect on men.

How can this apparent contradiction be resolved? Which statistician is correct? The answer, described in detail in Holland and Rubin (1983), is that either statistician can be correct: It depends. In particular, it depends on what is assumed about the control condition. As Lord (1967) originally posed the question, no control condition is described: Everyone in the school receives the school diet, so we do not observe any individuals under the control condition. In fact, because we do not even know what the control would be, any assumption about the control condition is untestable. Not even model diagnostics or fit tests will help here because we have no data on what happens to students under the control condition and thus no data on which to run the diagnostics.

Both statisticians assume that the potential outcome under control is a linear function of the September weight, $E[Y(0)] = a + BX$, but with different assumptions about a and B. Statistician 1 assumes that $a = 0$ and $B = 1$: that each student's June weight under control is equal to his or her weight in September, $Y(0) = X$. Statistician 2 on the other hand assumes that the value of a depends on a student's gender but that B is the same for men and women. Although both statisticians' analyses are correct as descriptive statements, neither is necessarily correct or incorrect as a statement about causal effects: It depends on the assumptions made. As stated by Holland and Rubin (1983), "Since [both assumptions] cannot be tested with the available data, acceptance or criticism of [them] must be based on intuition and/or subject-matter experience" (p. 11).

Like the approach taken by Statistician 1 in Lord's Paradox, a standard pre–post comparison implicitly assumes that the potential outcome under control is equal to the pretest score: $Y(0) = X$ (or in a more sophisticated analysis, that the trends in test scores would be the same before and after treatment in the absence of the intervention, as in Rand Corporation, American Institutes for Research, & National Opinion Research Center, 2004). Stated another way, a pre–post comparison assumes that in the absence of the treatment, test scores in any given year would be the same as those in the previous year. This assumption should be assessed with care in each particular study, and it should be clear in any discussion of effects that this assumption is driving the results. For example, when the intervention is applied to individual students, is it reasonable to assume that absent the treatment a student's test score at the end of the year will be equal to his or her test score at the beginning of the year? Similarly, student- or school-level test scores may increase from one year to the next just because students become more familiar with the type of test administered. Without a comparison group, it is impossible to determine how much of the change is due to the treatment itself and how much is due to other factors such as changes in time. In other words, without a comparison group, it is impossible to estimate causal effects without making strong, untestable assumptions about the potential outcomes under control.

Comparison Group Designs and the Importance of Careful Design

Although a well-designed random assignment study is the most desirable way to estimate causal effects, it is not always possible in education research, for a variety of reasons. Thus, education researchers often analyze observational data, in which we simply observe that some units received the treatment and others did not. If these two groups of units (e.g., schools) are very different from one another, the potential outcomes under treatment for the control group are predicted by using information on the treated schools, which look very different from the schools in the control group. Likewise, the potential outcomes under control for the treatment group are predicted by using information on the control schools, which look very different from those in the treatment group. Thus, when using comparison group designs, it is important to ensure that the groups being compared are as similar as possible.

Replicating a Randomized Experiment. In analyzing school-level data in an observational study, the goal is to conceptualize a hypothetical randomized experiment and try to replicate its design with the observational data, with the aim of comparing schools that look similar before treatment assignment. The full design of an observational study should thus include not only the prespecification of the outcome models that will be run but also the careful selection of the schools that will be used in the outcome analysis. As discussed by Rubin (2001), observational studies all too often consist simply of running models with a treatment indicator and a set of covariates as predictors. Instead, observational studies should be designed as randomized experiments are designed—without access to the outcome data and with a clear understanding of the treatment and control conditions. This approach involves investigating the extent to which the treatment and comparison groups are similar in terms of background covariates and then using methods such as matching or subclassification to ensure that the treatment effects are estimated by using schools that look similar to each other before the treatment is received.

Dangers of Regression Adjustment on the Full Samples. A relatively common method for estimating causal effects with observational data is to simply run a regression of the outcome on treatment assignment and the covariates, "controlling for" the covariates in this way. For example, using all schools in the state, school-level test scores in the year after a program is implemented in a subset of the schools may be regressed on an indicator for whether each school implemented the program and a set of covariates such as test scores before the program was implemented, school enrollment, school funding level, and so forth. There are two potential problems with this. First, intuitively, a regression analysis based on the full original data sets assumes that the relationship between the covariates and the outcome is linear across the entire space of the covariates, which may not be the case. Second, a regression models just the observed outcome, which for each unit (school) is actually one of two possible variables: For the control group, it is the potential outcome under control, whereas for the treatment group it is the potential outcome under treatment. In mathematical terms, $y_i^{obs} = T_i * Y_i(1) + (1 - T_i) * Y_i(0)$, where y_i^{obs} is the observed value of the outcome for school i, T_i is an indicator of the treatment received by school i (1 = treatment, 0 = control), and $Y_i(1)$ and $Y_i(0)$ are school i's potential outcomes under treatment and control, respectively.

Consider a situation where there is little overlap in the covariate distributions of the treatment and control groups and we would like to impute the missing potential outcomes under control for the treatment group. Although a linear model might be a good fit for the observed potential outcomes under control, we cannot know whether the linearity still holds in the space of the treated units' covariate values if the treatment units have a very different distribution of covariates as compared with the control group. Moreover, because few control units are observed in that space, there are no regression diagnostics that can be done to assess the model fit there. This may happen in a school-level study if for example the schools in the treatment group have higher test scores, lower enrollment, or higher funding levels than those in the control group, even before the program of interest is implemented.

The extrapolation required when the treatment and control groups have very different covariate distributions can lead to large biases if the wrong outcome model is used. For example, Cochran and Rubin (1973) showed that linear regression adjustment (i.e., ordinary least squares [OLS]) can actually increase bias in the outcome when the true relationship between the covariate and outcome is even moderately nonlinear, especially when the initial covariate bias and variance differences between the groups are large, as is often the case with observational data. Rubin (2001) gave three conditions for trustworthiness of a regression analysis and asserted, "If any of these conditions is not satisfied, the differences between the distributions of covariates in the two groups must be regarded as substantial, and regression adjustment will be unreliable and cannot be trusted" (p. 174). The conditions are as follows:

1. The difference in means of the propensity score [defined in the following section] in the two groups being compared must be small (e.g., the means must be less than half a standard deviation apart), unless the situation is benign in the sense that: (a) the distributions of the covariates in both groups are nearly symmetric, (b) the distributions of the covariates in both groups have nearly the same variances, and (c) the sample sizes are approximately the same.
2. The ratio of the variances of the propensity score in the two groups must be close to one (e.g., ½ or 2 are far too extreme).
3. The ratio of the variances of the residuals of the covariates after adjusting for the propensity score must be close to one (e.g., ½ or 2 are far too extreme). (p. 174)

Matching Methods. How can we create situations where the conditions for regression are more appropriate? One way is through matching methods, which ensure that the treatment and control groups being compared are similar on the covariates. Matching methods such as propensity score matching are becoming more and more popular as ways to estimate causal effects by using observational data. These methods, which select subsets of the original treatment and control units that are the most similar on the observed covariates, can be conceived as a way to replicate a randomized experiment by selecting treatment and control units (schools) that look only randomly different from one another on all of the observed covariates. An example of how this can be done with school-level data is given in the following section.

Propensity score matching or subclassification is particularly useful for selecting units that are similar to one another on a large set of covariates. The propensity score (Rosenbaum & Rubin, 1983b) collapses all of the observed covariates into a scalar summary that is the most important for selecting matched samples: the probability of receiving the treatment, conditional on the covariates. The matching can then be done on this scalar summary rather than on all of the covariates directly. The intuition behind this is that if two units have the same propensity score but are in different treatment groups, the determination of which unit received treatment and which received control was random. Thus, within a small range of values of the propensity score, the distribution of covariates should be the same in the treatment and control groups. In the school-level setting, this means that treatment and control schools with similar propensity scores should have similar distributions of the covariates that went into the propensity score, such as baseline test scores, enrollment, and school funding level.

There is not enough space here to go into the details of matching methods; some of the complexities include the choice of covariates to include in the matching, the number of matches to select, whether to match "with replacement," and how to define the distance between two units. Theoretical results regarding the propensity score and reviews and examples of the methods can be found in Rosenbaum and Rubin (1983b), Dehejia and Wahba (1999), and Imbens (2004). Software to implement a wide variety of matching methods can be found in Ho, Imai, King, and Stuart (2007); a practical discussion of matching methods, including advice regarding specific methods, can be found in Stuart and Rubin (in press).

Once matches or subclasses have been formed, the outcome analysis can proceed. Because the covariate distributions are selected to be similar, the impact estimates will be less dependent on the modeling assumptions than are regression estimates based on the full sample. Regression modeling on matched samples relies on an assumption of linearity across a smaller range of the covariates (and in a space where there is overlap in the covariate distributions of the two groups), which is much more likely to be reasonable. In fact, matching methods combined with regression have been shown to yield less biased estimates of treatment effects than does either method alone (e.g., Dehejia & Wahba, 1999; Ho, Imai, King, & Stuart, in press; Rubin, 1973a, 1973b). These results have a parallel in the randomized experiment literature showing the benefits of using a covariate both for prerandomization stratification and in the analysis of the outcome (Maxwell, Delaney, & Dill, 1984).

In addition to reducing dependence of the impact estimates on the model specification, matching methods have two other benefits. First, they highlight data sets or analyses in which units that are very different from each other would be compared. Whereas standard regression diagnostics do not warn the analyst about the extent of extrapolation required for inferences, matching methods can both provide this information and ensure that the analysis is done on comparable units. Even if matching is not used to select the comparison group, the process of attempting to choose matches will highlight the extreme extrapolations that may be required if for example the treatment and control schools are very different from one another in terms of the background covariates. The second benefit of matching is that the outcome variable is not used in the matching process. In standard analyses of observational data, each time multiple outcome models are run, the analyst sees the estimated treatment effect. Setting up the design first and then selecting comparable units to reduce model sensitivity prevents bias, or even the appearance of bias, that might result from selecting a set of matched units to yield a desired result.

Estimating School-Level Causal Effects Using the NLSLSASD

This section reviews the current work in which one school-level data set, the NLSLSASD, has been used to estimate causal effects. Few articles exist that use the NLSLSASD, and as of spring 2007, there appears to be only one in the peer-reviewed literature (Sanbonmatsu et al., 2006). However, the interventions examined in studies that use the NLSLSASD include Comprehensive School Reform (CSR), Reading First (RF), and increasing school choice. Because the NLSLSASD is a school-level data set, the units to be studied must be schools or higher-level units such as districts. In addition, the potential outcomes in any analysis based on the NLSLSASD or other school-level data sets are likely to be school-level test scores under treatment and control.[2]

Selecting a Comparison Group

The NLSLSASD essentially dictates the units and the outcome of interest for any study in which it is used; but because data are observed on such a large number of schools, it can be used to investigate the effects of a variety of interventions. The treated schools are those that received the intervention. Because the interventions of interest were rarely randomly assigned to a set of schools, defining the control condition and identifying comparison schools are sometimes more difficult to do. Indeed, one of the most common problems faced in the studies using the NLSLSASD was finding an appropriate comparison group. Although most of these studies do have some sort of comparison group, the adequacy of the comparison group varies in terms of its comparability with the treated schools. Of the 10 studies found that used the NLSLSASD to estimate school-level causal effects, 2 did not use any comparison group, 3 used all other schools in the state as a comparison group (and did not illustrate that the comparison and treatment groups were similar on any covariates), and 1 did not specify how the comparison schools were selected. Only 4 studies chose a comparison group based on pretreatment similarity to the treatment group; of these, 2 studies chose a comparison group using only one covariate (school poverty level), and only 2 selected a comparison group using more than one covariate (school poverty level, previous test scores, and racial distribution).

The NLSLSASD data raise two main challenges in identifying an appropriate comparison group. The first is finding control schools that look like the treatment schools, and the second is having enough information even to determine whether schools are similar. These two issues are discussed next, followed by an example of how to select well-matched comparison schools.

Selecting Appropriate Comparison Schools. The available studies that use the NLSLSASD indicate that it is difficult to select appropriate comparison schools from the NLSLSASD, either because there are not enough comparison schools available in the data set or more generally, because there are no good matches in the population. For example, the authors of the longitudinal assessment of the CSR program (U.S. Department of Education, 2004b) used a matched pairs analysis, but because schools were chosen to be in the program because they had particularly low test scores, the non–CSR schools had by definition slightly higher baseline achievement than did the CSR schools. The authors therefore selected the "best available matches given the requirements" (p. 11), but Exhibit B-5 in that report illustrates that there were large preexisting differences between the CSR schools and the non–CSR comparison schools.

When designing the Reading First implementation study, Moss et al. (2004) faced a similar problem. Because of the criteria used to determine which schools receive funding, it is difficult to identify untreated schools that are similar to the treatment schools and therefore suitable as a comparison group. The authors addressed this limitation by making it clear that they were not measuring the "effect" of Reading First programs. They stated: "This design will not attempt to explicitly measure differences between RF and non–RF schools" (p. 12). As the authors suggested, it is sometimes necessary to conclude that causal effects cannot be estimated on the basis of the available data because the treatment and control schools are too different on too many variables to permit accurate identification of the effect of the intervention.

One approach to selecting comparison schools is to use all untreated schools as a comparison group; 2 of the 10 studies used this approach. However, many of those schools may not be comparable to the treated schools. For example, if a large set of low-poverty schools was included in a study of a program for high-poverty schools simply because they were "untreated," it would be impossible to say whether an outcome difference between the treatment and the comparison schools was a function of the treatment alone or of other characteristics exhibited by the treatment schools but not by the comparison schools.

One example of a study that uses all untreated schools as a comparison group is an evaluation of the CSR demonstration (U.S. Department of Education, 2004a), in which the authors stated that they considered selecting comparison schools but that "there were concerns about whether the methods used to select these comparison schools actually resulted in a truly comparable comparison group" (U.S. Department of Education, 2004a, p. 21). If the authors did feel that the full group of untreated schools formed a better comparison group than would a smaller sample of matched schools, it would have been helpful to show diagnostics of this, for example, by comparing the means of baseline characteristics of the groups. The authors also stated that they did not choose a comparison group "because the choice of a comparison group could bias results in favor of finding an effect" (U.S. Department of Education, 2004a, p. A-10). However, when matching is used to select a comparison group solely on the basis of pretreatment covariates and when the analyst choosing the matches has no access to the outcome variable, the results will not be biased in one direction or the other. That is, if the comparison schools are very similar to the treated schools before treatment assignment, then it is likely that the resulting impact estimates will be less biased than if the full set of untreated schools was used.

Identifying Appropriate Comparison Schools. The second problem in using the NLSLSASD data is that it is difficult to determine whether two or more schools are similar because of limited covariate data on schools. To attribute differences in the outcome to the intervention, researchers have to assume that the schools were only randomly different from one another on all background covariates before the treatment was assigned. In other words, they have to assume that treatment assignment was random, conditional on the observed covariates (strongly ignorable treatment assignment). As discussed earlier, most current studies using the NLSLSASD match on only zero, one, two, or three covariates; this implies that there may still be large differences in the pretreatment distributions of many covariates. If there is extensive background information on school-level factors that may affect the outcome of interest—test scores, poverty level, school size, and so forth—and if close matches are found on all of the observed covariates, then the assumption of ignorable treatment assignment may be fairly believable. However, if the information on the schools is limited, the assumption may be less credible.

Fortunately, the problem of identifying comparison schools is more easily solved than is the problem of selecting appropriate comparison schools. With regard to the latter, good comparison schools may simply not exist. With regard to the former however, the covariate information (i.e., the data used to identify good matches) can be obtained either from other data sets, which then can be merged with the NLSLSASD, or by collecting more data from the schools. For example, a study by the U.S. Department of Education (2000) showed the feasibility of linking the NLSLSASD with the Schools and Staffing Survey. Census data also can be merged with the NLSLSASD to provide community-level demographic information. In addition, more extensive test score data collected from schools, for example standard scores or percentiles rather than simple percentages of students proficient at various levels, can be used in the matching process to round out the picture of achievement before treatment assignment.

Illustrative Example of Selecting Comparison Schools. This section provides an illustrative example of how well-matched comparison schools can be selected. It examines the effect of magnet schools on school test scores, using elementary-level magnet schools in Virginia as the treatment group. The control condition of interest is the "standard practice" of nonmagnet schools, and so all nonmagnet elementary schools in Virginia are considered as potential comparison schools. The purpose is to design a study that will later compare the longer-term outcomes (e.g., 2005 or 2006 average test scores) of the magnet and well-matched nonmagnet elementary schools.[3]

According to the NLSLSASD data, in the fall of 2002 there were 55 elementary-level magnet schools and 384 nonmagnet elementary schools in Virginia. A naive analysis would compare the outcome test scores of the 55 magnet schools with those of all 384 nonmagnet schools. However, columns 2 and 4 of Table 3 show that in fact the magnet and nonmagnet schools had very different distributions of school characteristics and test scores before the fall of 2002 (selected variables are shown; many more are available on the NLSLSASD data set). Thus, any comparison of outcomes between these full sets of schools would likely yield biased estimates of the effect of being a magnet school.

A more careful analysis would compare the magnet schools with well-matched nonmagnet schools. A first step might be to try finding "exact matches" on school characteristics before the fall of 2002, such as test scores, racial composition, and poverty level. This would involve finding for each magnet school a nonmagnet school with exactly the same values for all of these covariates. However, that quickly runs into the problem found in the studies discussed earlier: not enough good matches. To illustrate, each variable in

Table 3
Covariate Balance in Virginia Magnet and Nonmagnet Elementary Schools

Covariate	Magnet Schools Mean	Nonmagnet Schools Mean		Standardized Bias	
		Before Matching	After Matching	Before Matching	After Matching
Propensity score	0.39	0.09	0.36	1.21	0.14
Student:teacher ratio	12.56	13.73	12.59	–0.43	–0.01
% Title I school	40.00	41.70	43.60	–0.03	–0.07
% Students eligible for free lunch	34.72	31.71	33.98	0.13	0.03
% Students eligible for reduced price lunch	9.51	8.13	9.37	0.33	0.03
% Asian	9.13	4.04	9.37	0.64	–0.03
% Hispanic	18.65	6.91	17.85	0.73	0.05
% Black	32.88	31.24	33.22	0.06	–0.01
% White	39.03	57.57	39.19	–0.87	–0.01
% Passing history/social science (Grade 3, 2002)	71.38	74.24	71.51	–0.18	–0.01
% Passing history/social science (Grade 5, 2002)	66.77	69.75	67.27	–0.18	–0.03
% Passing history/social science (all elementary, 2000)	52.76	54.41	53.23	–0.09	–0.03
% Passing history/social science (all elementary, 2001)	58.41	64.58	58.50	–0.33	0.01
% Passing math (Grade 3, 2002)	71.90	78.09	72.39	–0.41	–0.03
% Passing math (Grade 5, 2002)	62.88	68.23	61.74	–0.31	0.07
% Passing math (all elementary, 2000)	60.07	64.74	60.25	–0.26	–0.01
% Passing math (all elementary, 2001)	63.94	68.80	62.97	–0.29	0.06
% Passing reading (Grade 3, 2002)	65.14	69.59	64.85	–0.29	0.02
% Passing reading (Grade 5, 2002)	72.30	75.12	72.08	–0.20	0.02
% Passing reading (all elementary, 2000)	59.61	61.86	58.84	–0.15	0.05
% Passing reading (all elementary, 2001)	60.84	66.36	61.21	–0.35	–0.02
% Passing science (Grade 3, 2002)	69.72	76.47	71.78	–0.48	–0.15
% Passing science (Grade 5, 2002)	66.56	72.98	66.71	–0.40	–0.01
% Passing science (all elementary, 2000)	61.21	65.01	60.39	–0.22	0.05
% Passing science (all elementary, 2001)	66.18	71.58	66.34	–0.32	–0.01
% Passing writing (Grade 5, 2002)	81.33	82.21	81.51	–0.08	–0.02
Average reading percentile rank (Grade 4, 2002)	50.96	51.32	50.56	–0.02	0.02
Average language percentile rank (Grade 4, 2002)	58.55	59.17	58.89	–0.04	–0.02
Average math percentile rank (Grade 4, 2002)	58.71	59.12	58.84	–0.02	–0.01
Total average percentile rank (Grade 4, 2002)	56.58	56.56	56.58	0.00	0.00
Total % at or above second quartile for all students (Grade 4, 2002)	85.09	85.87	84.04	–0.06	0.08
Total % at or above third quartile for all students (Grade 4, 2002)	58.54	58.03	58.51	0.02	0.00
Total % at or above fourth quartile for all students (Grade 4, 2002)	26.81	26.16	28.29	0.04	–0.08
Sample size	55	384	55	384	55

Note. Standardized bias is defined as the difference in means divided by the standard deviation in the full comparison group.

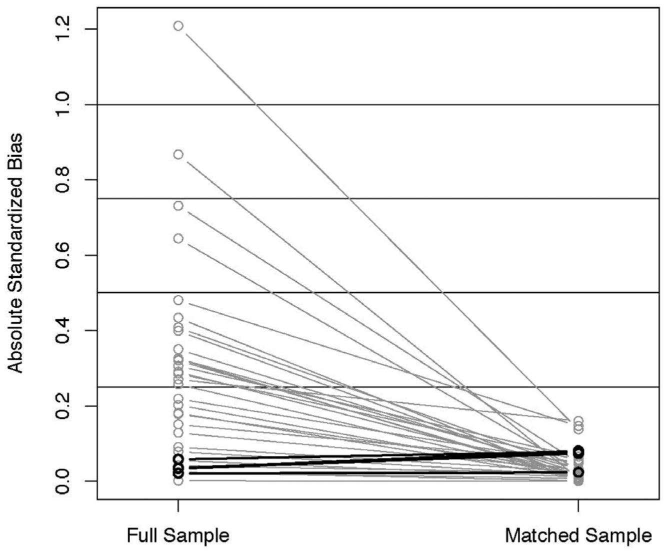

FIGURE 1. *Absolute standardized biases in full and matched samples. Absolute standardized bias is defined as the absolute value of the difference in means divided by the standard deviation in the full comparison group. Values over 0.5 are considered to be particularly problematic; ideally, values should be less than 0.25 to prevent bias. Black lines indicate variables whose absolute standardized bias increased with matching; gray lines indicate variables whose absolute standardized bias decreased.*

Table 3 was redefined as categorical, indicating the quartile of the state distribution in which each school falls. However, even with this simplification of the data, very few comparison schools with the same values as the magnet schools are found. Trying to find matches using just the demographic quartile variables (student:teacher ratio, Title I status, percentage eligible for free lunch, percentage eligible for reduced price lunch, percentage White, and percentage Black) leads to just 35 of the 55 magnet schools having an exactly matched nonmagnet school. Adding in any of the test score variables reduces the number of matches even further. The problem is one of high dimensionality: With many covariates and relatively few units, it is very difficult to find pair matches such that each pair has exactly the same (or even similar) values on each of the covariates.

This is where the propensity score can be of use in selecting groups of treatment and comparison units with similar *distributions* of the observed covariates. A propensity score was estimated by using logistic regression to predict magnet school status as a function of all of the covariates listed in Table 3; the propensity scores are the resulting predicted probabilities of being a magnet school. A fairly simple matching method was then used, selecting for each magnet school the nonmagnet school with the closest propensity score. Columns 3 and 5 of Table 3 show the resulting balance in the matched samples. We see that nearly all of the covariates have very similar means. The propensity score, as a summary of all covariates, has a standardized bias of 1.21 standard deviations in the full samples and only 0.14 standard deviations in the matched samples.[4] The largest standardized bias in the matched sample on any individual covariate is 0.15, which is a much better situation for regression adjustment, as discussed earlier. Figure 1 graphically illustrates the reduction in bias achieved by matching for each of the covariates, showing that the propensity score matching has selected a set of nonmagnet schools with distributions of demographics and test scores before the fall of 2002 very similar to those of the magnet schools.[5]

Models of the Outcome

Once the study has been designed, either through randomization or the careful selection of comparison units, the analysis of the outcome can proceed. This section describes some of the common outcome models used to estimate treatment effects in analyses of education data. Because studies using the NLSLSASD are generally nonrandomized, the discussion emphasizes the underlying hypothetical randomized experiments, with particular attention to what is being assumed regarding the potential outcome under control for the treatment units. The outcome of interest is assumed to be school-level test scores; scores measured after the intervention was implemented are referred to as posttest scores, and test scores measured before the intervention was implemented are referred to as pretest scores. Space does not permit going into the details of each of the models; this section is geared primarily toward providing an overview and references for more information. Fairly simple models are described to make the points clearly; more complex modeling approaches certainly exist. As with all modeling, model checks and diagnostics (e.g., through model comparisons) should be performed to determine the best approach for any particular analysis.

Compare Posttreatment Test Scores. A simple (and perhaps naive) estimate of the treatment effect can be obtained simply by taking the difference in means of the outcomes (e.g., test scores at the end of the school year) between treatment and control schools: $\hat{\tau} = \bar{y}_1 - \bar{y}_0$, where $\hat{\tau}$ is the estimated treatment effect, $\bar{y}_1$ is the average outcome (the average of the average test scores) in the treated group, and $\bar{y}_0$ is the average outcome in the control group. In this case, the potential outcome under control is effectively being imputed for each treated unit as $\bar{y}_0$.

In a randomized experiment where each unit has the same probability of receiving the treatment, this difference in means is an unbiased estimate of the true average treatment effect. However, in the absence of randomization, the difference is a biased estimate of the treatment effect if the treatment and control groups are not comparable (e.g., if smaller schools tend to be in the treated group and larger schools tend to be in the control group). If close matches that are obtained in an observational study are such that the covariate distributions are the same in the two groups, this estimator can provide a good estimate of the treatment effect. However, it is generally better to use one of the regression methods described next, which control for small remaining differences in the covariate distributions between the matched treatment and control groups (see e.g., Imbens, 2004).

Regression Adjustment. Regression adjustment fits a model of the outcome conditional on the covariates. A common procedure is to estimate a linear regression model (OLS) with the treatment indicator and covariates as predictors of the outcome of interest: $E(Y|T, X) = a + \hat{\tau}\ T + BX$, where $\hat{\tau}$ is taken as the estimated treatment effect. The missing potential outcomes under control are effectively being imputed as $E[Y(0)] = a + BX$. As discussed earlier, if the treatment and control groups have dissimilar covariate distributions, estimates based on regression adjustment will rely heavily on the modeling assumptions such as linearity.

The previously described model is a very simple linear model. The use of more complex regression-type models is somewhat limited in this setting because of the lack of student-level data. However, more advanced methods can be used for particular analyses, such as longitudinal analyses, or to estimate multiple regression models simultaneously by using structural equation modeling (Kaplan, 2000). A more complex model that is used frequently in education research is multilevel modeling (also known as hierarchical linear modeling or random effects modeling; Raudenbush & Bryk, 2002), which typically is used with student-level data to account for the clustering of students in classrooms or schools. These models generally have two levels, the first level being a model of the student-level outcome of interest (e.g., student test scores), with coefficients modeled at the second level (the school). In other words, at least some of the parameters of the student-level model depend on the school that a student attends; the school-specific parameters are modeled at the second level. In the setting considered in this article, where there are only school-level data, multilevel models could instead be used to account for the clustering of schools in districts or states or for longitudinal analyses where there are multiple observations on each school (Hedeker, 2004). The longitudinal analyses would enable researchers to answer questions about whether interventions lead to changes in the trajectory of test scores in addition to any changes in the levels of scores in particular years after the intervention (Muthén & Curran, 1997). Of course, all of the concerns regarding the trustworthiness of simple regression adjustments also hold for these more complex modeling approaches.

Comparing Gain Scores. Another common approach for estimating the effect of an education intervention is to compare the gain scores of the treated and control schools. The gain score is defined as the change in scores over time, for example, between the beginning and end of the school year or from one year to the next. These methods are also sometimes called difference-in-difference models because they take the difference (between treated and control groups) of the difference (change) in test scores before and after the intervention. Therefore, they can be thought of as using the ideas behind both comparison group designs and pre–post designs. Analyzing gain scores is very similar in concept to comparing outcome test scores. In fact, if the intervention is assigned randomly, in expectation two methods estimate the same quantity, $\hat{\tau} = E[(\bar{y}_1 - \bar{x}_1) - (\bar{y}_0 - \bar{x}_0)] = E(\bar{y}_1 - \bar{y}_0)$, because $E(\bar{x}_1) = E(\bar{x}_0)$ in a randomized experiment. Therefore, because estimates based on posttest scores or on gain scores provide unbiased estimates of the treatment effect in a randomized experiment, gain scores do not help to reduce bias in randomized experiments as compared with simply analyzing posttest scores. However, if the pretest and posttest scores are correlated, as would be expected, the use of gain scores can yield more efficient estimates. Modeling the gain in test scores as a function of the covariates is also a special case of modeling the posttest score as a function of the covariates, including the pretest score, where the coefficient on the pretest score is set equal to one. Thus, the efficiency gained by analyzing gain scores instead of posttest scores is also a special case of the efficiency achieved by including covariates in regression models in randomized experiments, as discussed by Bloom, Richburg-Hayes, and Black (2005).

The approach of selecting well-matched units and then comparing gain scores is used in the National Longitudinal Study of No Child Left Behind (Rand Corporation et al., 2004), which compares gain scores in a set of treated schools and a set of matched comparison schools. McLaughlin et al. (2002) also compared gain scores in Title I schools to gain scores in non–Title I schools, and the U.S. Department of Education (2004b) compared gain scores between CSR and non–CSR schools.

Interrupted Time Series Design. An interrupted time series design models trends in time and compares the outcome predicted from those trends with the outcome actually observed. For example, if 10 years of pretreatment test scores are available, a model is fit to those 10 years of data, and the outcome at year $t + 1$ is predicted on the basis of that model, where t is the year of treatment assignment. Deviations from that prediction are interpreted as the effect of the treatment administered in year t. Thus, the potential outcome under control is assumed to be the prediction for the year of interest obtained from a model of the trend in scores estimated in the years prior to the treatment.

Although intuitively appealing, these interrupted time series designs are highly dependent on the modeling assumptions. For example, although both a linear and a nonlinear baseline trend may fit the observed data fairly well, the two can result in very different estimates of the treatment effect because of the extrapolation required: The outcome at a future time is predicted only on the basis of data from before that time period. It is therefore very important to assess sensitivity to the model by using several models of the outcome, as in Bloom (2001). In addition, like Lord's Paradox, these designs use as a control the treated schools at an earlier time rather than using a set of comparison schools that have not received the intervention. And without a control group, it is impossible to know whether deviations from the trend are a result of the treatment itself or of other factors that change over time.

Bloom (2001) used an interrupted time series design in the evaluation of Accelerated Schools, providing a nice description of the method and its assumptions. He also mentioned the possibility of combining this method with a comparison group design, an approach that should certainly be pursued. In the Reading Excellence Act (REA) and School Implementation and Impact Study, Moss et al. (2003) did this, using an interrupted time series design in combination with a set of comparison schools in which test scores in schools not receiving REA funds are modeled over time. The values predicted from that model are compared with the outcomes observed in the REA–funded schools. A relevant question not answered in that report is how well the model predicted future scores for non–REA schools. That information could provide a diagnostic for how well the model predicts the score under control for the REA–funded schools.

Regression Discontinuity Designs. Regression discontinuity designs take advantage of the way some programs are administered, relying on a discrete eligibility cutoff in a variable such as test scores or the school poverty rate. Units below the cutoff do (or do not) receive the treatment, whereas units above the cutoff do not (or do). Because units just below and just above the cutoff are assumed to be very similar to one another before the treatment is assigned (at least on the characteristic defining the cutoff), jumps in performance *at the cutoff* are attributed to the treatment. The potential outcome under control for units that received the treatment and are just above (or below) the cutoff is assumed to be similar to the outcome observed for the units that did not receive the treatment and are just below (or above) the cutoff.

As is the case for all of the regression methods discussed, because this method depends heavily on the accuracy of the model assumptions, the control schools used to model the trends must be similar to the treatment schools. In addition, as discussed by Todd, Hahn, and van der Klauww (2001), the treatment effect is identified only for units (schools) with a pretreatment value at the cutoff. Estimating effects for schools with other covariate values requires even more model assumptions.

These designs may hold promise as means of estimating the effects of particular interventions, but it is sometimes difficult to identify an appropriate cutoff and to believe the underlying model assumptions. Ludwig and Miller (2007) provided an example of the use of regression discontinuity, applied to Head Start, and the National Longitudinal Study of No Child Left Behind (Rand Corporation et al., 2004) described a complex use of a regression discontinuity design, identifying some of the complications and assumptions in doing so.

Conclusions and Suggestions for Future Use of School-Level Data Sets

School-level data sets such as the NLSLSASD have a place in both randomized experiments and observational studies. In the former, they can be used in the design phase either as a sampling frame or as a tool for selecting specific schools for use in a study; moreover, because by definition they include school-level data, these data sets can also save resources that would otherwise be devoted to data collection. In the latter, national data sets such as the NLSLSASD provide a large set of potential comparison schools. However, the key to drawing accurate causal inferences from the data is to compare treatment and control groups with similar distributions of covariates so that any difference in the outcome can be attributed to the treatment, not to preexisting differences between the groups. The following suggestions for researchers are intended to maximize the use of these rich and promising databases:

1. When school-level data sets such as the NLSLSASD are used in observational studies to estimate causal effects, researchers should first posit the underlying hypothetical randomized experiment that could have been conducted and then attempt to replicate that experiment with the available data. In designing the analysis and interpreting the results, researchers should be very clear about the control condition to avoid Lord's Paradox, in which a change over time is inappropriately interpreted as a causal effect without clarity about the underlying assumptions.
2. The set of treatment and control schools in any observational study should be very similar on all covariates related to the outcome of interest before the treatment is assigned. Matching methods such as propensity scores can be used to help determine which schools look most alike.
3. In reporting on their findings, researchers should demonstrate to their readers that the treatment and control groups were similar before the treatment was assigned. For example, reports should include a comparison of the means in the two groups on a set of important covariates that are believed to affect the outcome, as illustrated in Table 3 and Figure 1. Few current reports provide this type of information, and those that do use only a small set of covariates.

NOTES

This work was originally sponsored by the U.S. government as part of a National Academy of Sciences Symposium on the Use of the National Longitudinal School-Level State Assessment Score Database (NLSLSASD) held in December 2005. Any opinions, findings, conclusions, or recommendations expressed in the article are those of the author and do not necessarily reflect the view of the U.S. government or the National Academy of Sciences. The author thanks Donald Rubin, David Myers, Sameena Salvucci, and Sonya Vartivarian for very helpful comments and discussion as well as the editors and three anonymous reviewers for their comments and suggestions, which greatly improved the article.

[1] Although a test/retest design may seemingly solve this problem by enabling observation of an individual both without the intervention (the original test) and with the intervention (the retest), in fact an individual at two different points in time constitutes two different units. We cannot be sure that a change from test to retest is due to the intervention and not due to other changes that the individual may have experienced between the two points in time. This is discussed later in the article in the context of pre–post studies and Lord's Paradox.

[2] Other complications also arise in using these types of data sets, such as questions of the comparability of test scores across different states. This article focuses on the complications inherent in defining and estimating causal effects and leaves discussion of other issues to other research.

[3] Of course, there are other issues involved in conducting a study like this, such as the length of time the magnet schools have been in operation. Those issues are not addressed in this article, which focuses instead on the illustration of selecting matched samples.

[4] The standardized bias is defined as the difference in means divided by the standard deviation in the full comparison group.

[5] More complex methods, such as ratio matching or full matching (Hansen, 2004), would also make it possible to retain a larger number of the well-matched comparison schools. Simple 1:1 nearest neighbor matching is used here for expository purposes.

REFERENCES

Bloom, H. S. (2001). *Measuring the impacts of whole-school reforms: Methodological lessons from an evaluation of Accelerated Schools*. New York: Manpower Demonstration Research Corporation.

Bloom, H. S., Richburg-Hayes, L., & Black, R. A. (2005). *Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions* (MDRC Working Articles on Research Methodology). New York: Manpower Demonstration Research Corporation.

Cochran, W., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics*, *35*, 417–446.

Cook, T. D., Murphy, R. F., & Hunt, H. D. (2000). Comer's School Development Program in Chicago: A theory-based evaluation. *American Educational Research Journal*, *37*, 535–597.

Dehejia, W., & Wahba, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053–1062.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver and Boyd.

Freedman, D. A. (2001). Ecological inference and the ecological fallacy. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia for* *the social and behavioral sciences* (Vol. 6, pp. 4027–4030). New York: Elsevier.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*, 609–618.

Hedeker, D. (2004). An introduction to growth modeling. In D. Kaplan (Ed.), *Quantitative methodology for the social sciences* (pp. 215-234). Thousand Oaks, CA: Sage.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). *MatchIt: Nonparametric preprocessing for parametric causal inference* [Computer software and manual]. Retrieved January 23, 2007, from http://gking.harvard.edu/matchit/

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (in press). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, *81*, 945–960.

Holland, P. W., & Rubin, D. B. (1983). On Lord's Paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederick Lord* (pp. 3–25). Hillsdale, NJ: Lawrence Erlbaum.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*, 4–30.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *55*, 304–405.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's outcomes? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, *122*, 159–208.

Maxwell, S. E., Delaney, H. D., & Dill, C. A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin*, *95*, 136–147.

McLaughlin, D. H., Bandeira de Mello, V., Cole, S., Blankenship, C., Hikawa, H., Farr, K., et al. (2002). *National Longitudinal School-Level State Assessment Score Database: Analyses of 2000/2001 school-year scores* (Report submitted to the U.S. Department of Education). Palo Alto, CA: American Institutes for Research.

Moss, M., Gamse, B., Jacob, R., Smith, W. C., Greene, D., & Kupfer, A. (2003). *Reading Excellence Act and school implementation and impact study: Annual Report 2002–2003* (Report submitted to Policy and Program Studies Service, U.S. Department of Education). Cambridge, MA: Abt Associates.

Moss, M., Tao, F., Jacob, R., Boulay, B., Gamse, B., Schimmenti, J., et al. (2004). *Reading First implementation study: Final study design* (Report submitted to Policy and Program Studies Service, U.S. Department of Education). Cambridge, MA: Abt Associates.

Muthén, B., & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*, 371–402.

Neyman, J. (1990). On the application of probability theory to agricultural experiments: Essay on statistical principles, Section 9 (D. M. Dabrowska & T. P. Speed, Trans.). *Statistical Science*, *5*, 465–480. (Original work published 1923)

Pearl, J. (2000). *Causality*. New York: Cambridge University Press.

Rand Corporation, American Institutes for Research, & National Opinion Research Center. (2004). *National Longitudinal Study of No Child Left Behind: Plans for analyses of Wave I survey data, Draft 2* (Report submitted to the U.S. Department of Education). Washington, DC: Rand Corporation.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society*, *45*, 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, *29*, 159–183.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, *29*, 184–203.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, *2*, 169–188.

Sanbonmatsu, L., Kling, J. R., Duncan, G. J., & Brooks-Gunn, J. (2006). Neighborhoods and academic achievement: Results from the Moving to Opportunity experiment. *Journal of Human Resources*, *41*, 649–691.

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

Stuart, E. A., & Rubin, D. B. (in press). Matching methods for causal inference: Designing observational studies. In J. Osborne (Ed.), *Best practices in quantitative methods*. Thousand Oaks, CA: Sage.

Todd, P., Hahn, J., & van der Klauww, W. (2001). Identification of treatment effects by regression discontinuity design. *Econometrica*, *69*, 201–210.

U.S. Department of Education. (2000). *School-level correlates of academic achievement: Student assessment schools in SASS public schools*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

U.S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide* (Prepared by the Coalition for Evidence-Based Policy, sponsored by the Council for Excellence in Government). Retrieved February 28, 2007, from http://www.ed.gov/rschstat/research/pubs/rigorousevid/index.html

U.S. Department of Education. (2004a). *Implementation and early outcomes of the Comprehensive School Reform Demonstration (CSRD) program*. Washington, DC: U.S. Department of Education, Office of the Under Secretary.

U.S. Department of Education. (2004b). *Longitudinal assessment of Comprehensive School Reform program implementation and outcomes: First-year report*. Washington, DC: U.S. Department of Education, Office of the Deputy Secretary, Policy and Program Studies Service.

Whitehurst, G. J. (2004, April 26). *Making education evidence-based: Premises, principles, pragmatics, and politics* (Distinguished Public Policy Lecture, Institute for Policy Research, Northwestern University). Retrieved November 3, 2006, from http://ies.ed.gov/director/doc/2004_04_26.pdf

AUTHOR

ELIZABETH A. STUART is an assistant professor in the Department of Mental Health and the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, 8th Floor, Baltimore, MD 21205; *estuart@jhsph.edu*. While working on this article she was a researcher at Mathematica Policy Research, Inc. Her research focuses on developing statistical methods for education and mental health research, particularly relating to causal inference and missing data.

Manuscript received July 13, 2006
Revisions received November 7, 2006, and January 26, 2007
Accepted February 2, 2007