

ESTIMATING SCHOOL-LEVEL CAUSAL EFFECTS USING THE SSASD

ELIZABETH A. STUART, PH.D.

MATHEMATICA POLICY RESEARCH, INC.

JANUARY 8, 2006

PAPER PREPARED FOR NATIONAL ACADEMY OF SCIENCES SYMPOSIUM ON THE USE OF THE SCHOOL-LEVEL STATE ASSESSMENT DATABASE (SSASD)

I. INTRODUCTION

Education researchers, practitioners, and policymakers alike are committed to identifying interventions that teach students more effectively in a variety of areas. Their search involves such questions 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, such as the nationwide School-Level Assessment Database (SSASD) now being assembled by the American Institutes for Research for the Department of Education, can be used to address these types of questions about the effects of school- or district-level interventions.

The SSASD contains assessment data on 80,000 public schools in nearly all states in the country. This public-use dataset is available at [http://www.schooldata.org](http://www schooldata.org) and currently 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 SSASD is a rich and useful data set for estimating the causal effects of educational interventions at the school level.

This paper presents the now-common framework of potential outcomes for estimating causal effects, emphasizing the assumptions necessary to estimate these effects and how the framework and assumptions relate to the SSASD. It also provides some examples of the types of analyses now based on the SSASD.

Copyright © 2005 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.

Opinions and statements included in the paper are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Planning Committee for the Symposium on the Use of School-Level Data in Evaluating Federal Education Programs or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.

A. A FRAMEWORK FOR DEFINING CAUSAL EFFECTS

The framework presented here is based on a formulation by Neyman (1923) and Fisher (1925) for randomized experiments, which was extended to observational studies by Rubin (1974; 1978). In this framework, the three building blocks for defining causal effects are treatments, units, and potential outcomes.

1. Treatment and Control Conditions

Like the studies using the SSASD considered in this paper, current education research is largely intended to measure the effect of some intervention (the treatment) relative to a different intervention or no intervention at all (the control). Although the treatment is generally easily defined since that 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 that 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 and in interpreting results from any study estimating causal effects.

2. The Units

The units are “entities” to which a treatment is assigned. A new reading curriculum, for example, may be assigned to an entire school, to certain classrooms within a school, or to individual students. When the units are schools or school districts, the SSASD, because it contains school-level data, can be used to estimate the causal effects of an intervention at the school-level. Without student- or classroom-level data, the SSASD 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 conclusions being drawn at an incorrect level (e.g., at the student level, when schools were randomly assigned to treatment conditions). In the SSASD, which has school-level data, the units must be schools or other higher-level units such as districts. These school-level impacts 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, for example, Freedman 2001).

3. 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 (for example, average test scores) a school would experience if it were using the new reading curriculum. Likewise, the potential outcome under the control condition, often denoted $Y(0)$, 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. Causal inference is thus inherently a missing data problem, where at least half of the values of interest (the potential outcomes) are missing.

B. LEARNING ABOUT CAUSAL EFFECTS

Given that we can never observe both potential outcomes for a particular unit and thus individual-level causal effects are never observed, how can we learn about causal effects? Section A 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.

1. Replication

Replication means that there are multiple units for which we can observe one of the potential outcomes. If there were only one unit (that received either treatment or control), we would have no information on the missing potential outcome. However, with some units assigned to treatment and others to 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.

2. Stability

The “stability” assumption, also known as SUTVA (the “stable unit treatment value assumption”), has two components. The first is that each unit’s potential outcomes are not affected by the treatment assignment of 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 students in control classrooms interact on the playground with students in treatment classrooms, the treatment may “spill over” onto the control students, thus affecting their outcomes. We could see a similar spillover effect in teaching methods across schools in the same districts 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, although 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—i.e., the same treatment is administered to all units in the treatment group (and likewise for the control group).

3. 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 a known assignment mechanism that is particularly useful because it ensures that the covariates are balanced between the treatment and control groups. In other words, in large randomized experiments the treatment and control groups will be only randomly different from one another on all background covariates, observed and unobserved. Thus, any difference in outcomes between the two groups can be attributed to the treatment itself, not to pre-existing differences between the groups.

In observational studies, however, in which the treatment is not assigned randomly and the true assignment mechanism is unknown, we must posit a hypothetical assignment mechanism. A key assumption in observational studies is that of strongly ignorable treatment assignment (Rosenbaum and Rubin 1983a), which implies that (1) treatment assignment is independent of the potential outcomes given the observed covariates, and (2) 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, as described by Rosenbaum and Rubin (1983b).

II. LORD’S PARADOX

A. The Importance of Thinking Clearly About Causal Effects: An Example

To illustrate the pitfalls of making statements about causal effects without considering the fundamental concepts, we use an example originally posed by Frederick Lord in 1967. This discussion has often been invoked in debates on the use of covariance adjustment versus gain scores in educational research, although it perhaps more clearly shows the fundamental importance of the assignment mechanism in causal inference. The discussion here is based closely on Holland and Rubin (1983). Consider the following:

“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.”
(Lord 1976, 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. In other words, the average weight for men is the same in September and June, and the variance of the male weights is the same in September and June (and the same is true for the women). 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:

“...as far as these data are concerned, there is no evidence of any interesting effect of diet (or of anything else) on student weights. In particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change.” (Lord 1967, p. 305)

Statistician 1, using the difference in mean weight gains for men and women as the estimated differential causal effect, thus concludes that there is no differential causal effect since neither group gains nor 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: $E(Y|X,G=i)=a_i + bX$, where Y is June weight, X is September weight, and G is the gender of the student (1=male, 2=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 $a_1 - a_2$ 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., $a_1 > a_2$), the diet must have a larger effect on men.

B. Resolving the Paradox

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: that it depends. In particular, it depends on what is assumed about the control condition. As Lord 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 don't even know what the control would be, any assumption about the control condition is untestable. Either statistician can therefore be right, depending on what is assumed about the control.

Both statisticians assume that the potential outcome under control conditions is a linear function of the September weight, $Y_i(0)=a+BX$, but they make different assumptions about the values of 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 ($Y_i(0)=X_i$). 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. 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). While 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.

C. Pre-post Studies in Educational Research

In the context of the SSASD, Lord's paradox illustrates the importance of having a clear understanding of the control condition, and observing some units under that condition. One approach sometimes used to estimate causal effects is to do a pre-post design, where, for example, test scores before a new curriculum is implemented are compared with test scores in the academic year after the curriculum is implemented. However, as illustrated by Lord's Paradox, changes over time do not necessarily suggest that the curriculum had a causal effect on test scores.

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 pre-test score: $Y(0)=X$ (or in a slightly more sophisticated analysis, as in Rand Corporation et al. 2004, that the trends in test scores would be the same pre- and post-treatment in the absence of the intervention). 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 impacts 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? Will students not learn anything new if they are not in the treatment group? Similarly, 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.

III. DESIGNING OBSERVATIONAL STUDIES TO ESTIMATE CAUSAL EFFECTS

A. Replicating a Randomized Experiment

Although a well-designed random assignment study is the most desirable way to estimate causal effects, it is not, for a variety of reasons, always possible in educational research. Thus, educational 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 are very different from one another, the potential outcomes under treatment for the control group are predicted using information on the treated individuals, who look very different from those in the control group. Likewise, the potential outcomes under control for the treatment group are predicted using information on the control individuals, who look very different from those in the treatment group.

In analyzing data from an observational study, we will conceptualize a hypothetical randomized experiment and try to replicate its design with the observational data, with the aim of comparing units that look similar before random assignment. The full design of an observational study should thus include not only the pre-specification of the outcome

models that will be run, but also the careful selection of the units 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 through the use of samples that look similar to each other before the treatment is received.

B. Matching Methods

Matching methods such as propensity score matching are becoming more and more popular as a way to estimate causal effects using observational data. These methods, which select subsets of the original treatment and control units who 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 that look only randomly different from one another on all of the observed covariates.

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 and Rubin 1983a) collapses the full set of multidimensional 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 the full set of multidimensional covariates. The intuition behind this is that if two units have the same propensity score but are in different treatment groups, which unit received treatment and which received control was determined randomly. 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. We do not have 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 1983 and Imbens 2004.

C. Dangers of Regression Adjustment on the Full Samples and the Benefits of Matching

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. 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. 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. While 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. In contrast, 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.

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) show 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) gives three conditions for regression analysis to be “trustworthy”: “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. These conditions are:

1. The difference in means of the propensity score 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., $\frac{1}{2}$ 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., $\frac{1}{2}$ or 2 are far too extreme).”

When matching has been used to select similar units, these conditions are much more likely to be satisfied. Consistent with this, matching methods combined with regression have been shown to yield less biased estimates of treatment effects, as compared with regression adjustment alone (e.g., Rubin 1973a,b; Dehejia and Wahba 1999; Ho et al. 2005).

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. The second benefit of matching is that the outcome variable is not used in

the matching process. In standard analyses of observational data, however, 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.

IV. ESTIMATING SCHOOL-LEVEL CAUSAL EFFECTS USING THE SSASD

This section reviews some of the current work in which the SSASD has been used to estimate causal effects (see, for example, Moss et al. 2004; U.S. Department of Education 2004a,b). The interventions, or treatments, considered in this work include comprehensive school reform (CSR), Reading First, and increasing school choice. Because the SSASD is a school-level data set, the units are restricted to be schools or other higher-level units such as districts. Finally, the potential outcomes in any analysis based on the SSASD are likely to be school-level test scores under treatment and control.

A. Selecting a Comparison Group

Although the SSASD essentially dictates the units and the outcome of interest for any study in which it is used, because data is 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 is sometimes more difficult to do. Indeed, one of the most common problems faced in the studies based on the SSASD was finding an appropriate comparison group. Although most of these studies do have some sort of comparison group, the adequacy of this group varies in terms of its comparability with the treated schools. Of the ten studies distributed for this symposium that use the SSASD to estimate school-level causal effects, two did not use any comparison group, three 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 one did not specify how the comparison schools were selected. Only four studies chose a comparison group based on pre-treatment similarity to the treatment group, with two studies choosing a comparison group using only one covariate (school poverty level), and only two selected a comparison group using more than one covariate (school poverty level, previous test scores, and racial distribution).

The SSASD data raises 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 to even be able to determine whether schools are similar.

1. Selecting Appropriate Comparison Schools

The studies considered in this paper indicate that it is difficult to select appropriate comparison schools from the SSASD 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, in the Longitudinal Assessment of the Comprehensive School Reform (CSR) Program (U.S. Department of Education 2004b), the authors use 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,” but Exhibit B-5 in that report illustrates that there were large pre-existing differences between the CSR schools and the non-CSR comparison schools.

In the Reading First Implementation Final Study Design, 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 (RF) programs, in particular, “This design will not attempt to explicitly measure differences in between RF and non-RF schools” (p. 12). As the authors suggest, it is sometimes necessary to conclude that it is not possible to estimate causal effects given the available data because the treatment and control schools are too different on too many variables to be able to accurately identify the effect of the intervention.

One approach to selecting comparison schools is to use all untreated schools as a comparison group; two of the ten studies distributed for this symposium 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 were included in a study of a program for high-poverty schools simply because they were “untreated,” it would be impossible to say whether a difference in the outcomes between the treatment and the comparison schools was a function of the treatment alone or of other characteristics exhibited by the comparison schools but not by the treatment schools.

One example of a study that uses all untreated schools as a comparison group is the evaluation of the Comprehensive School Reform Demonstration (U.S. Department of Education 2004a), in which the authors state 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.” If the authors do feel that the full group of untreated schools forms a better comparison group than would a smaller sample of matched schools, it would be helpful to show diagnostics of this, for example, by comparing the means of baseline characteristics of the groups. The authors also state that they did not choose a comparison group “. . . because the choice of a comparison group could bias results in favor of finding an effect...” (p. A-10). However, when matching is used to select a comparison group solely on the basis of pre-treatment 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 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 is used (see, e.g., Dehejia and Wahba 1999).

2. Identifying Appropriate Comparison Schools

The second problem in using the SSASD data is that it is difficult to determine whether two or more schools look alike because the covariate data on schools in the data set are limited. 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, that treatment assignment was random, conditional on the observed covariates (unconfounded treatment assignment; Rubin 1978). As discussed earlier, most current studies using the SSASD match on only 0, 1, 2, or 3 covariates, implying that there may still be large differences in the pre-treatment 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, etc.—and if close matches are found on all of the observed covariates, then the assumption of unconfounded 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. In terms of the former, however, the covariate information (i.e., the data to identify good matches) can be obtained either from other datasets and merging these files with the SSASD or by collecting more data from the schools. For example, McLaughlin et al. (2000) show the feasibility of linking the SSASD with the Schools and Staffing Survey. Census data also can be merged with the SSASD to provide community-level demographic information. In addition, more extensive test score data collected from schools, for example standard scores or percentiles rather than simply the percentage of students proficient at various levels, can round out the picture of achievement before treatment assignment.

B. Models of the Outcome

This section describes some of the common outcome models used to estimate treatment effects in analyses of education data. Because these studies are generally based on observational data, the discussion emphasizes the underlying hypothetical randomized experiments, with particular attention on what is being assumed about or imputed as 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 “post-test” scores, and test scores measured before the intervention was implemented are referred to as “pre-test” scores.

1. Compare Post-Treatment Test Scores

A simple (and perhaps naïve) estimate of the treatment effect can be obtained simply by taking the difference in means of the outcome (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 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 below, which control for small differences in the covariate distributions between the treatment and control groups (see, e.g., Imbens 2004).

2. 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: $E(Y|T,X) = a + \tau T + BX + e$, where $\hat{\tau}$ is taken as the estimated treatment effect. The missing potential outcomes under control are effectively being imputed as $Y_i(0) = a + BX$.

3. Comparing Gain Scores

Another common approach for estimating the effect of an educational intervention is to compare the treated and control schools in terms of gain scores. 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 pre- and post-intervention. This is very similar in concept to comparing outcome test scores. In fact, if the intervention is assigned randomly, in expectation the two methods estimate the same quantity:

$$\tau = E((\bar{y}_1 - \bar{x}_1) - (\bar{y}_0 - \bar{x}_0)) = E(\bar{y}_1 - \bar{y}_0)$$

because in a randomized experiment, $E(\bar{x}_1) = E(\bar{x}_0)$. Therefore, because estimates based on post-test scores or on gain scores both provide unbiased an unbiased estimate of the treatment effect in a randomized experiment, gain scores do not help to reduce bias in randomized experiments as compared with simply analyzing post-test scores. However, if the pre- and post-test scores are correlated, as would be expected, using gain scores can yield more efficient estimates. Regressing the covariates on the gain in test scores is also a special case of regressing the covariates, including the pre-test score, on the post-test score, where the coefficient on the pre-test score is set equal to one. Thus, the efficiency gained by

analyzing gain scores instead of post-test 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 study of improving literacy through school libraries (Paper 10) and 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. That paper also includes a nice description of the methodology and the underlying assumptions. McLaughlin et al. (2002) also compare gain scores in Title I schools to the gain scores in non-Title I schools, and the U.S. Department of Education (2004) compares gain scores between CSR and non-CSR schools.

4. 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 pre-treatment 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, they can also result in very different estimates of the treatment effect because of the extrapolation required: the outcome at a future point in 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 done in Bloom (2001). In addition, like Lord's Paradox, these designs use as a control the treated schools at an earlier point in time, rather than using a set of comparison schools that did not receive the intervention. And without a control group, it is impossible to know if deviations from the trend are a result of the treatment itself or other factors that change over time.

Bloom (2001) uses an interrupted time series design in the evaluation of Accelerated Schools, providing a nice description of the method and its assumptions. He also mentions the possibility of combining this method with a comparison group design, an approach that should certainly be pursued. In the Reading Excellence Act and School Implementation and Impact Study, Moss et al. (2003) do this, using an interrupted time series design in combination with a set of comparison schools, in which test scores in schools not receiving Reading Excellence Act (REA) funds are modeled over time, and 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.

5. Regression Discontinuity Designs

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

As is the case for all of the regression methods discussed, because this method depends heavily on the “correctness” 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 et al. (2001), the treatment effect is only identified for units (schools) with a pretreatment value at the cut-off. Estimating effects for schools with other covariate values requires even more model assumptions.

These designs may hold promise as a means to estimate the effects of particular interventions, but it is sometimes difficult to identify an appropriate cut-off and believe the underlying model assumptions. The National Longitudinal Study of No Child Left Behind (Rand Corporation 2004) describes a complex use of a regression discontinuity design, identifying some of the complications and assumptions in doing so.

V. CONCLUSIONS AND SUGGESTIONS FOR FUTURE USE OF THE SSASD

The SSASD has a place in both randomized experiments and observational studies. In the former, it can be used in the design phase as either a sampling frame or as a tool for selecting specific schools for use in the study. Moreover, because it includes school-level data, it can also save resources that would otherwise be devoted to data collection. In observational studies, the SSASD provides 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 the covariates so that any difference in the outcome can be attributed to the treatment, not to pre-existing differences between the groups. The following suggestions for researchers and developers of the SSASD are intended to maximize the use of this rich and promising database.

Suggestions for Researchers

1. When the SSASD is 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 SSASD 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 being clear about the underlying assumptions.
2. The set of treatment and control schools in an SSASD-based 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 to 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 and for subgroups of interest for comparison. Few current reports provide this type of information, and those that do use only a small set of covariates.

Suggestions for SSASD Developers

1. To enable researchers to assess the comparability of treated and control schools, SSASD developers should continue to include data on school characteristics or allowing a straightforward way to obtain that information from other databases, and should consider collecting more extensive information from schools.
2. Average standard scores, rather than just the percentage of students proficient at various levels, should be collected. Pre-treatment achievement levels are often the most important variable that dictates the choice of comparison schools, but shifting standards make it difficult to use information from several years when the only information available is the percentage of students meeting a particular standard.

REFERENCES

- Bloom, H. (2001). "Measuring the impacts of whole-school reforms: Methodological lessons from an evaluation of accelerated schools." New York, NY: Manpower Demonstration Research Corporation.
- Bloom, H.S., Richburg-Hayes, L, and Rebeck Black, A. (2005). "Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions." MDRC Working Papers on Research Methodology. New York, NY: Manpower Demonstration Research Corporation.
- Cochran, W. and Rubin, D.B. (1973). "Controlling bias in observational studies: A review." *Sankhya: The Indian Journal of Statistics, Series A* 35: 417-446.

- Dehejia, W. and 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.
- D.A. Freedman. (2001). "Ecological inference and the ecological fallacy." *International Encyclopedia for the Social and Behavioral Sciences*. Vol. 6. N.J. Smelser and P.B. Baltes, eds. Pages 4027-4030. Elsevier.
- Ho, D., Imai, K., King, G., and Stuart, E.A. (2005). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Working paper available at <http://gking.harvard.edu/files/matchp.pdf>.
- Holland, P.W. (1986). "Statistics and causal inference (with discussion)." *Journal of the American Statistical Association* 81: 945-960.
- Holland, P.W. and Rubin, D.B. (1983) "On Lord's Paradox." *Principals of Modern Psychological Measurement*, Chapter 1. Howard Wainer and Samuel Messick, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Imbens, G. (2004). "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and Statistics* 86 (1): 4-29.
- Lord, FM. (1967) "A paradox in the interpretation of group comparisons." *Psychological Bulletin* 55: 304-405.
- McLaughlin, D., Bandeira de Mello, V., Cole, S., Blankenship, C., Hikawa, H., Farr, K., and Gonzalez, R. (2002). "National Longitudinal School-Level State Assessment Score Database: Analyses of 2000/2001 School-Year Scores." Submitted to the U.S. Department of Education. Washington, DC: American Institutes for Research.
- McLaughlin, D., Drori, G., and Ross, M. (2000). School-level correlates of academic achievement: Student Assessment Schools in SASS Public Schools. National Center for Education Statistics Research and Development Report, U.S. Department of Education Office of Education Research and Improvement, NCES 2000-303.
- Moss, M., Gamse, B., Jacob, R., Smith, W.C., Greene, D., and 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, Inc.
- Moss, M., Tao, F., Jacob, R., Boulay, B., Gamse, B., Schimmenti, J., and Faddis, B. (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, Inc.

-
- Neyman, J. (1923). "On the application of probability theory to agricultural experiments: Essay on statistical principles," Section 9. Translated in *Statistical Science* 5 (1990): 465-480.
- Rand Corporation, American Institutes for Research, and National Opinion Research Center. (2004). "National Longitudinal Study of No Child Left Behind: Plans for Analyses of Wave 2 Survey Data, Draft 2." Submitted to the U.S. Department of Education.
- Rosenbaum, P.R. and Rubin, D.B. (1983a). "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70: 41-55.
- Rosenbaum, P.R. and Rubin, D.B. (1983b). "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome." *The Journal of the Royal Statistical Society Series B* 45(2): 212-218.
- Rubin, D.B. (1973a). "Matching to remove bias in observational studies." *Biometrics* 29: 159-184.
- Rubin, D.B. (1973b). "The use of matched sampling and regression adjustment to remove bias in observational studies." *Biometrics* 29: 185-203.
- Rubin, D.B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66: 688-701.
- Rubin, D.B. (1976a). "Multivariate matching methods that are equal percent bias reducing, I: Some examples." *Biometrics* 32: 109-120.
- Rubin, D.B. (1976b). "Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction." *Biometrics* 32: 121-132.
- Rubin, D.B. (1978). "Bayesian inference for causal effects: The role of randomization." *The Annals of Statistics* 6: 34-58.
- Rubin, D.B. (1997). "Estimating causal effects from large data sets using propensity scores." *Annals of Internal Medicine* 127: 757-763.
- 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.
- Todd, P., Hahn, J., and van der Klauww, W. (2001). "Identification of Treatment Effects by Regression Discontinuity Design" *Econometrica*.
- U.S. Department of Education, Office of the Under Secretary, (2004a). *Implementation and Early Outcomes of the Comprehensive School Reform Demonstration (CSRD) Program*, Washington, D.C.

U.S. Department of Education, Office of the Deputy Secretary, Policy and Program Studies Service (2004b). *Longitudinal Assessment of Comprehensive School Reform Program Implementation and Outcomes: First-Year Report*, Washington, D.C.