

### 3. Estimating Causal Effects Using Observational Data

SOME OF THE MOST IMPORTANT theoretical and methodological work in education research has resulted from data analyses using large-scale national datasets such as the Early Childhood Longitudinal Study (ECLS) and the National Education Longitudinal Study of 1988–2000 (NELS). For example, two award-winning books, *Public and Private High Schools: The Impact of Communities* (Coleman & Hoffer, 1987) and *Catholic Schools and the Common Good* (Bryk, Lee, & Holland, 1993), were based on analyses of High School and Beyond (HS&B), a longitudinal study of high school students in the 1980s. The number of dissertations, articles in refereed journals, and other publications that have been written from these national datasets is well over 10,000. Large-scale datasets that are drawn from multistage probability samples allow for predictive analyses and tentative causal inference. Investigators can estimate the probable effects of certain conditions for specific populations over time. In instances where there are data elements about school or pedagogical practices, analytic techniques can estimate the likelihood of what would happen if certain organizational, institutional, or instructional reforms were implemented on a larger scale.

In some cases, such datasets can also be used to approximate randomized controlled experiments. For example, matched sampling has been used to assess the causal effects of interventions when randomized experiments cannot be conducted (Rubin, 2006). Over the past three decades, statisticians (e.g., Rubin, 1974, 1978; Rosenbaum, 1986) and econometricians (e.g., Heckman, 1976, 1979) have developed several methods of analysis for making causal inferences with observational data such as large-scale national datasets.

There are several advantages to using large-scale, nationally representative datasets to study student achievement differences. Large-scale studies, such as NELS, are based on nationally representative samples of U.S. students and their parents, schools, and teachers. In contrast to randomized controlled experiments, which are designed to yield valid causal results but often have limited generalizability, large-scale national educational studies are designed to be generalizable to specific populations of students, such as high school students in the United States. Large-scale datasets thus serve as a rich source of descriptive information on students, teachers, and schools. Because they are based on large, nationally representative samples, such datasets are also useful in studying the characteristics and achievement of subgroups such as minority and low-income students, groups that are often targeted for educational interventions. In addition, such datasets are often longitudinal, making it possible for analysts to measure achievement gains at both the individual and group levels. Large-scale datasets can also be used to develop plausible hypotheses regarding the causes of differences in student achievement gains. For example, analyses of administrative data from Texas public school systems have been useful in developing some promising models for estimating teacher quality (Rivkin, Hanushek, & Kain, 2005).

Analyses of large-scale datasets can also inform the design of randomized controlled trials. Such datasets can be

used to identify promising interventions, to target subgroups that are most likely to benefit from such interventions, and to suggest causal mechanisms that may explain why an innovative program may have positive effects on student achievement relative to a more conventional program. Moreover, when randomized controlled trials are not feasible, large-scale nationally representative studies may provide the best source of data on which to base educational policy decisions.

Despite their strengths, large-scale observational datasets do not typically include the random assignment of individuals or schools to treatment and control groups that is the hallmark of randomized controlled trials.<sup>21</sup> Researchers therefore need to be aware of the tradeoffs involved in choosing experimental versus non-experimental designs when both can be used to address a particular research question and both are financially, logistically, and ethically feasible. For example, “natural experiments” constructed from survey data are sometimes used to investigate the effects of particular educational programs or reforms (Angrist & Krueger, 2001). These methods seek to isolate comparisons that capture causal effects even without the benefit of purposeful random assignment.

When constructing treatment and control groups from observational data, researchers have limited control over the composition of the groups being compared. Those who participate in a program may differ systematically from those who do not, which can bias estimates of program effects, a problem referred to as *sample selection bias*. For example, if a researcher is trying to evaluate the effect of a high school dropout program on high school completion rates and the analysis is based only on students who complete the program, the sample used for analysis may overrepresent students at low or moderate risk of dropping out and underrepresent high-risk students who drop out of school prior to completing or entering the program (Cuddeback, Wilson, Orme, & Combs-Orme, 2004).

Researchers have developed several different procedures to adjust for selection bias. One of the earliest and most well-known techniques was developed by James Heckman (1976, 1979). In this two-step procedure, a multiple regression model is estimated for an outcome of interest (e.g., high school completion rates). A selection model is also estimated comparing those who participate in a program with those who do not participate on selected variables. If differences between participants and nonparticipants are detected, then adjustments are made to the first model to correct for these differences. There are limitations, however, to procedures used to correct for selection bias. The selection model used to detect and correct for selection differences may be misspecified. For example, important variables may be missing from the model. In such cases, attempts to correct for selection bias may actually make estimates more problematic (Stolzenberg & Relles, 1997; Winship & Mare, 1992).

In some cases, selection bias can be corrected by adjusting outcomes for relevant observables that are correlated with the outcome variable and the independent variables of interest. This has been termed *observable selection bias* (Barnow, Cain, & Goldberger, 1980). However, unobserved characteristics can also bias estimates of program effects. For example, in assessing achievement differences between public and charter school students, procedures for reducing observable selection bias may be used to adjust for differences in family characteristics such as income and structure. But there may also be unobserved characteristics that are associated with both charter school attendance and student achievement outcomes. Charter schools may attract students who are having academic difficulties in public schools. Families may enroll their children in a charter school specifically because they are already not doing as well as their public school classmates. Charter schools may also appeal to the most motivated parents, eager to provide opportunities to their children that they feel are

lacking in regular private schools. In the first instance, estimates of the effect of charter schools on student achievement may be biased downward; in the second they may be biased upward. Such selection factors are often “unobservables” or omitted variables. They are correlated with the educational intervention in question and therefore bias estimates of the effect of that intervention on outcomes.<sup>22</sup>

Social scientists have developed several methods to adjust for observed and/or omitted variables when making comparisons across groups using observational data. These methods, which include fixed effects models, instrumental variables, propensity score matching, and regression discontinuity designs, have been used to approximate randomized controlled experiments (see Winship & Morgan, 1999, for a useful overview).<sup>23</sup>

## **Methods for Approximating Randomized Assignment**

### *Fixed Effects Models*

Many large-scale nonrandomized datasets contain multiple observations of individuals over time. Since a major concern is that unobserved characteristics are correlated with both treatment and outcome variables, controlling for such unobservables would reduce the bias in the estimate of the treatment effect. One approach to correcting for selection bias when there are identifiable treatment groups is to adjust for fixed, unobserved characteristics that may be associated with selection into the treatment group.<sup>24</sup> Janet Currie (2003) offers a clear example of this approach. She suggests that in looking at the effect of mother’s employment on child outcomes, the mother’s personality may be related to the likelihood both of being employed and of having good child outcomes. If one assumes that personality is unlikely to change over time, it can be considered a fixed characteristic. For example, women who are more “nurturing” may be more likely to stay at home

with their children and to have good child outcomes. In this example, mother's employment or unemployment can be considered the treatment and control conditions; mother's personality is an unobserved characteristic that may be related both to selection into employment and child outcomes. Excluding this variable from analytic models may therefore bias estimates of the effect of mother's employment on child outcomes (e.g., maternal employment may appear to have a more negative effect on child outcomes than it actually does). Currie notes that in cases where the mother works during the infancy of one child but stays at home during the infancy of another, it is possible to compare the effects of mother's employment status on the outcomes for siblings. Because the children have the same mother, the effect of the mother's personality is assumed to be fixed, though unmeasured. The argument is that by comparing the differences in the outcomes of siblings in the two groups, the analyst can obtain an unbiased estimate of the effects of maternal employment on those outcomes.

Currie and Thomas (1995) use a similar approach in analyzing the effects of Head Start on child outcomes. Results of previous studies had consistently shown a negative relationship between Head Start attendance and student learning outcomes. However, the children served by Head Start are typically from low-income families and have parents with low levels of educational attainment. Compared with children from more advantaged families, these children consistently score lower on measures of cognitive growth. To control for differences in the background characteristics of children selected into Head Start versus those not enrolled in a Head Start program, Currie and Thomas looked at the outcomes of siblings who differed with respect to Head Start enrollment. One sibling had attended Head Start and the other had not; in most other respects, however, the children had similar background characteristics. In this instance, household effects were considered fixed since the siblings were from the same household. Using this fixed effects

model, Currie and Thomas found that siblings who attended Head Start did systematically better than siblings who did not, even though children in Head Start programs had lower-than-average achievement overall. Currie and Thomas were thus able to argue that Head Start does have significant effects on child outcomes.

A study by Bifulco and Ladd (2006) provides another example of the use of fixed effects to adjust for possible bias from unobserved characteristics. The investigators analyzed a large longitudinal sample of North Carolina students that included five third-grade cohorts.<sup>25</sup> In addition to end-of-grade reading and mathematics test scores and data on student background characteristics, the dataset included information on whether the school was a charter or regular public school and a school identifier. Bifulco and Ladd were able to track individual students over time and identify whether they were attending a charter or regular public school in any given year. Approximately 65% of students in the sample had attended both a public school and a charter school; the investigators were thus able to compare the test score gains of students while in charter schools with the test score gains of *these same students* while in traditional public schools. Because the same students were observed in each school setting, the effects of time-invariant student characteristics (both observed and unobserved) were the same across school settings (i.e., they were “fixed” across settings; such fixed effects included the student’s gender, race, and ethnicity). The strength of this method is that it does not rely on comparing charter school students to some other group of students; it therefore substantially reduces the problem of self-selection bias.<sup>26</sup> In comparing the outcomes of students in each setting, Bifulco and Ladd found that students in charter schools scored significantly lower in both reading and mathematics.<sup>27</sup>

As Currie (2003) notes, there are a number of drawbacks to using fixed effects models to correct for selection bias. First,

the assumption that the unobserved characteristic is fixed or time invariant may not be valid. Mother's personality, for example, may actually change over time; or siblings may respond to their mother differently, which may potentially affect the outcome of interest. The analyst thus needs to provide a convincing rationale for why the variable should be considered fixed. Second, fixed effects models may considerably reduce the size of the sample being analyzed, making it difficult to detect treatment effects. In the case of the Head Start study, for example, only children enrolled in Head Start who had a non-enrolled sibling are used to identify the effects of Head Start. Third, the analytic sample may not be representative of the population of interest. For example, in looking at the effects of mother's employment on child outcomes, it would be helpful to know if mothers who changed their employment status between the birth of one child and the next did so for a specific reason. Mothers who worked during the infancy of one child but not another may have stayed home because the second child had health or developmental problems. Consequently, the sample analyzed would not be representative of the larger population of mothers who chose not to work. Fourth, fixed effects estimates tend to be biased in the direction of "no effect." As Currie notes, "Intuitively, we can divide a measured variable into a true 'signal' and a random 'noise' component. The true signal may be very persistent between siblings (e.g., if both children have high IQ), while the noise component may be more random (e.g., one child has a bad day on the day of the test). Hence when we look at the difference between siblings, we can end up differencing out much of the true signal (since it is similar for both siblings) and being left only with the noise" (pp. 5–6). However, when the analyst can provide a convincing rationale for regarding the variable as fixed and is working with a large dataset, fixed effects models can be a powerful means for detecting treatment effects.

*Instrumental Variables*

A second method to correct for omitted variables is to include an “instrumental variable” in the analysis (Angrist, Imbens, & Rubin, 1996; Angrist & Krueger, 2001). An analytic tool used primarily by economists, instrumental variables were first used over 40 years ago, to estimate supply and demand curves and then to counteract bias from measurement error (Angrist & Krueger). This approach has also been used to overcome omitted variable problems in estimating causal relationships, typically problems that are narrowly defined in scope. In estimating the effect of years of schooling on earnings, for example, the observed relationship between earnings and the explanatory variable, years of schooling, is likely to be misleading because it partially reflects omitted factors that are related to both variables, such as cognitive ability. If ability could be accurately measured and held constant in a statistical procedure like regression, then the problem of omitting this variable could be avoided. But researchers typically are unsure what the best controls are for ability, and without more detailed information, we cannot assume the contribution of ability from the relationship between schooling and earnings.

This is where the instrumental variable enters in. A good instrumental variable should be associated with the treatment or endogenous variable (years of schooling) but be uncorrelated with the omitted variable (e.g., ability) and thus have no association with the exogenous or outcome variable (earnings), except through schooling. Because the instrumental variable is correlated with years of schooling but is uncorrelated with other determinants of earnings, such as ability, the causal effect of the instrument on earnings is proportional to the causal effect of schooling on earnings.<sup>28</sup> Instrumental variable estimates can be computed using two-stage least squares (2SLS) regression analysis. In the first stage, the instrumental variable and any covariates are used to predict the endogenous

variable (years of schooling or “treatment” in our example) in a regression equation. In the second stage, the dependent variable is regressed on fitted values from the first stage regression plus any covariates.<sup>29</sup> If the instrumental variable is uncorrelated with the omitted variable (ability), the predicted value of years of schooling is also uncorrelated with the omitted variable. The bias in the estimation of earnings resulting from the exclusion of ability from the model is thus removed.

In an investigation of the effect of years of schooling on earnings, Angrist and Krueger (1991) used birth date and compulsory school laws as instrumental variables, with ability, family background, and any other unobserved determinants of earnings as the omitted variables. Children whose birthdays occur earlier in the year enter school at an earlier age than students whose birthdays occur later in the year. For example, a child whose birthday occurs before the school year starts will begin kindergarten at age 5. However, a child whose birthday is in December will enter kindergarten the next year, when he/she is almost 6 years old. Assuming that the compulsory age that one can leave school is 16, then students whose birthdays fall earlier in the year can leave school before entering 10th grade, whereas those whose birthdays fall later in the year must remain in school for an additional few months. In examining the relationship between years of schooling and earnings for men who are likely to leave school when they reach the compulsory school age, birth date is a good instrument because it determines who starts school in a given year or a year later, but is not correlated with omitted variables. Compulsory schooling laws derived from the states in which individuals were born are also a good instrument because they determine who can leave school in a given year or a year later but are probably uncorrelated with ability or family background. Angrist and Krueger explain this as follows: “The intuition behind instrumental variables in this case is that differences in earnings by quarter of birth are assumed to be accounted for solely by

differences in schooling by quarter of birth. . . . Only a small part of the variability in schooling—the part associated with quarter of birth—is used to identify the return to education” (p. 74). In this example, the estimated earning gain from more time in school applies to those who are likely to leave school at the minimum leaving age. It might not apply to those who are college bound or who are determined to finish high school even if they could leave school at age 16. In order to test the same hypothesis for other groups, another instrument would have to be found, or different survey data collected.

Angrist and Krueger (1995) found that men born in the first quarter of the year have about one tenth of a year less schooling and earned about 0.1% less than men born later in the year, but this difference was negligible. As it turns out, instrumental variables estimates using the Angrist and Krueger quarter of birth instruments are remarkably similar to the corresponding regression estimates that make no adjustment for unobservables. The investigators therefore concluded that there is no omitted variables bias in standard regression models estimating the effects of education on earnings.<sup>30</sup>

As Currie (2003) observes, there are several difficulties with using instrumental variables to correct for bias resulting from omitted variables. From a pragmatic standpoint, it is quite difficult to identify good instruments. Moreover, although the analyst can check to see whether different instrumental variables produce consistent results, it is not possible to check the validity of one’s assumptions about the variables. In addition, instrumental variables may be only weakly related to the endogenous variable; the use of such “weak” instruments can result in biased and misleading estimates (Currie; see also Staiger and Stock, 1997, and Bound, Jaeger, and Baker, 1995, for a discussion of weak instruments; Angrist and Krueger, 1995, show that the estimates in their 1991 article are not affected by this problem).

### *Propensity Scores*

A third method used to correct for selection bias is propensity scores. An important difference between propensity score methods and instrumental variables methods is that the former can correct for omitted variables bias due to unobserved characteristics while the latter corrects only for bias from observed characteristics or covariates. Propensity-score methods essentially are a version of regression or matching that allows researchers to focus on the observed covariates that “matter most.”

Most regression analyses in nonrandomized observational studies are carried out for the full range of a particular sample, without regard for the probability that individuals have of being in the treatment or control groups. Propensity score matching is a technique developed by Rosenbaum and Rubin (1983) to represent the predicted probability that individuals with certain characteristics would be assigned to a treatment group when assignment is nonrandom (see also Rubin, 1997). The advantage of using propensity score matching is that it aggregates a number of characteristics that individually would be difficult to match among those in the treatment and non-treatment groups. Take the example of student performance in private schools compared to public schools. Students from disadvantaged families are much less likely to attend private schools. At the other end of the spectrum, students from well-off families, particularly minority high-income families, have a relatively higher probability of attending a private school. To approach a random assignment trial, we should compare individuals who have a reasonable probability of choosing to be in either the treatment (e.g., private school) or the control group (e.g., public school). Students with similar propensities to be in the treatment group (whether they are in the treatment group or not) can be matched on the basis of their propensity scores. The difference in their achievement scores would be closer to

the difference we would expect in a random assignment of students to the two groups, since it is much more likely that their pretreatment characteristics are similar.

There are a number of ways propensity scores can be used to match students in the treatment and control groups (in this instance, private and public schools). Perhaps the most common way is to sort students from each group into “bins” or strata based on the distribution of propensity scores. Within each bin, the characteristics of students in the two treatment conditions are similar on a weighted composite of observed covariates. If the average characteristics of students within a bin are not equal, a more refined model is developed, using additional bins or strata until a balance in the characteristics of students in each group (e.g., public and private schools) is achieved. In some cases, there may be “bins” in which there is no overlap between the treatment and control groups, indicating that these individuals (e.g., students at the ends of the spectrum mentioned above) have virtually no probability of attending private schools on the low end or public schools on the high end. Because these students have no matches, they are excluded from analyses. This technique approximates randomized assignment since students within each of the remaining bins or strata have a roughly equal probability (based on their aggregate characteristics) of being selected into either the treatment or control condition.<sup>31</sup>

Propensity scores address an important issue in empirical research, namely, estimates of effects for certain groups when randomization is not possible, and where sample elements have self-selected themselves into treatment or control conditions.

All statistical methods, from the simplest regressions to the most complex structural models, have elements of this limitation when used to analyze phenomena with heterogeneous responses. Nevertheless, many interventions and relationships can be fruitfully studied using

estimated effects for specific subsamples, provided the possible limitations to generalizing the results are understood and explored. Indeed, this lack of immediate generality is probably the norm in medical research based on clinical trials, yet much progress has been made in that field. (Angrist & Krueger, 2001, p. 78)

Stephen Morgan (2001) addressed the issue of observable differences in the nonrandom assignment of students to Catholic and public secondary schools. Using data from the National Education Longitudinal Study, he estimated a Catholic school effect on 12th-grade achievement. He also estimated the propensity of Catholic and public school students to attend Catholic schools based on a set of socioeconomic and demographic variables, as well as the student's score on a 10th-grade achievement test; these propensity scores were used to match students attending Catholic and public schools. Morgan then re-estimated the Catholic school effect on mathematics and reading scores for matched groups of students; he also estimated the effect within propensity score strata. Since not all Catholic school students had a match in the sample of public school students at the upper end of the propensity to attend Catholic school (high-socioeconomic-class students), Morgan conducted two sets of estimates: one that included the unmatched students and one that did not. His findings suggest that the estimated Catholic school effect for those currently attending Catholic schools using propensity score matching is larger than the estimate without propensity score matching and is statistically significant even when non-matched students are omitted. Hong and Raudenbush (2005) also used propensity score matching to estimate the effects of kindergarten retention (versus promotion) on students' reading and mathematics achievement at the end of the retention year. This study is described in detail in Section 4.

In contrast to fixed effects and instrumental variables, propensity score matching adjusts only for *observed* characteristics. Because a large number of background characteristics are used in calculating propensity scores, the probability that a relevant variable has been omitted from analysis is reduced, though not eliminated. However, it is possible to test the sensitivity of results to hypothesized omitted variables (Rosenbaum & Rubin, 1983; Rosenbaum, 1986, 2002). Because an aggregate of characteristics is used in computing propensity scores, and analytic samples are restricted to individuals (or schools) that can be matched across treatment conditions, this approach to approximating randomized assignment is more effective when large, nationally representative datasets are used. The samples on which these datasets are based are sufficiently large to allow for analyses of a subsample and contain comprehensive information on the background characteristics of students and schools. If selection into the analysis is unbiased (e.g., exclusions due to missing data do not result in differences between the analysis sample and the larger sample), these results are also generalizable to the larger population of students or schools.

### *Regression Discontinuity*

A fourth method that can be used to approximate random assignment is regression discontinuity.<sup>32</sup> Regression discontinuity also plays on features of certain occurrences in education that have the qualities of a natural experiment; namely, when group members are subject to a treatment because they fall either above or below a certain cutoff score (for recent examples of this approach, see Cook, in press; Hahn, Todd, & Van der Klaauw, 1999, 2001; Van der Klaauw, 2002). The example used in Campbell's (1969) seminal article on regression discontinuity is the effect of National Merit Scholarships on later income. The fact that those just above or below the

cutoff for acceptance into the program are likely to be similar on a set of unobserved variables that predict scores on the test determining National Merit Scholarship awards suggests that the effect of the treatment on a dependent variable (in this case, future income) could be estimated by comparing this restricted group—those just above the cutoff (who received the treatment)—with those just below the cutoff (who did not receive the treatment). Campbell argued that “if the assignment mechanism used to award scholarships is discontinuous, for example, there is a threshold value of past achievement that determines whether an award is made, then one can control for any smooth function of past achievement and still estimate the effect of the award at the point of discontinuity” (Angrist & Lavy, 1999, p. 548). Assuming that individuals in this restricted group approximate a random assignment to the treatment and control groups (at this particular cutoff point), the estimate of regression at the cutoff point yields an unbiased estimate of the treatment. If there are situations in which there are multiple discontinuities, this provides an even better estimate of the treatment effect since it would then be estimated across a broader range of the initial independent variable (in the merit scholarship case, at different levels of test scores).

Another example of a regression discontinuity analysis is a study conducted by Brian Jacob and Lars Lefgren (2004) that compares remedial summer school and grade retention effects on cohorts of third- and sixth-grade students by using data from the Chicago Public Schools. Jacob and Lefgren limit their analysis of the effects over time of summer school and grade retention on relatively low-achieving students. To determine whether attending summer school and having to repeat a grade had a significant effect on reading and mathematics achievement, they compared students who were just above or below the cutoff for promotion to the next grade. The assumption was that low-achieving students who just barely exceeded the cutoff score for promotion would be similar to students

who fell just below the cutoff for promotion; however, one group of students would receive the “treatment” (attending summer school and potentially having to repeat a grade), while the other would not. One year after the initial promotion decision, the third graders who barely failed to meet the promotional standard scored roughly 20% of a year’s worth of learning higher than their peers who barely passed the standard. The effects faded somewhat by Year 2 but were still statistically significant. For sixth graders, the effects were not positive, although the authors note that the results for these students were confounded by the differential incentives that retained and promoted students faced in subsequent years.

Regression discontinuity designs require that samples be restricted to students who fall just above or below the cutoff point; thus analyses based on large-scale datasets have a greater likelihood of detecting treatment effects. In contrast to propensity score matching, where students are matched on the basis of aggregate characteristics, regression discontinuity *assumes* that students in the two groups have similar characteristics; however the validity of this assumption should be checked.

### **Implications of These Results for Causal Inference**

The methods being used by social scientists in analyzing large datasets address key issues in educational policy—for example, the effect of attending a public or Catholic school, the effect of teachers on student achievement, the effect of program participation on earnings—and also address the selection bias that is inherent in nonrandom assignment. The studies reviewed here have carefully analyzed the sources of bias and a series of estimation problems in the datasets. The main point of discussing these different methods and their applications to a number of different policy problems in education is to show how large datasets that are not based on randomized assignment to

treatment and control groups can be used to obtain unbiased estimates of treatment effects.

It should be clear from this discussion that there are important limits to survey analysis even when adjustments for selection bias and multiple levels of analysis are used. Since populations are heterogeneous, estimates of the relationship between an intervention and educational outcomes corrected for selection bias may not be applicable to groups that have a low probability of falling into either the treatment or control group. Even so, almost all of the studies cited were able to deal with the effects on different groups. Currie and Thomas (1999) estimated the effects of Head Start participation on Hispanic students, for example, and Morgan (2001) estimated the effect of attending Catholic school by social class. Some of these studies reinforce other studies that used different methods but reached similar conclusions. After two decades of discussion, for example, the Morgan analysis of the effect of attending Catholic secondary schools concludes that the effect of the treatment (attending a Catholic school) is positive and significant with respect to achievement, but this effect is not necessarily a consequence of a school's religious status as Catholic. This is also the conclusion reached a decade earlier by Bryk, Lee, and Holland (1993), who described a "Catholic school effect" as being associated with shared values, a moral imperative, and school policies rather than religiosity per se.

In the last few years, analyses of large-scale datasets using the methods described above have produced several important findings, some of which have implications for causal inference and for the design of randomized experiments. In the next section, we highlight several NSF-supported research studies that relied on large-scale datasets and were designed either to estimate causal effects or to provide the preliminary evidence necessary in designing randomized controlled experiments of educational interventions and identifying populations most likely to benefit from them.

## Chapter 3 Notes

- 21 Researchers, however, have consistently recommended that randomized control trials can and should be embedded within large-scale observational studies.
- 22 This is an example of the problem of *endogeneity*. This term refers to the fact that an independent variable (e.g., charter school attendance) is potentially a choice variable that is correlated with an unobserved characteristic (e.g., parent motivation), or is itself caused in some way by the outcome (e.g., student achievement). Strictly speaking, endogeneity requires this feedback; otherwise, the problem is one of omitted variables bias. As in the example above, a student's prior achievement may influence parents' decision to send or not send their child to a charter school. Once this initial decision is made, the student's achievement may, in turn, influence the parents' decision to have the child remain in or exit the school. In such cases, the outcome is observed for both public and charter school students. This differs from Heckman's classic example of sample selection bias where the outcome is observed only for those who choose to participate in a particular program. Heckman's two-step procedure is designed to deal with this truncated distribution.
- 23 Propensity score matching controls only for observable characteristics, whereas the other methods control for both observed and unobserved characteristics; however, in the case of propensity score matching, sensitivity analysis can be used to test for the possible effects of unobserved variables (Rosenbaum & Rubin, 1983; Rosenbaum, 1986, 2002).
- 24 Fixed effects models in this and the next example do not refer to fixed effects as opposed to random effects assumptions that are applied in a general linear model.
- 25 Students in the third grade in 1996, 1997, 1998, 1999, and 2002 were followed until they left the North Carolina public school system, completed the eighth grade, or until the 2001–2002 academic year, whichever came first.
- 26 As the authors note, the drawback of this method is that it is not based on the full population of charter school students. The authors conducted additional analyses to determine whether the subsample

of students for whom test scores were available in both charter schools and regular public schools differed from the larger group of all charter school students in the grades observed. Although some differences between samples were found (e.g., the subsample over-represented charter school students who exited charter schools and underrepresented students who entered charter schools), the average impact of charter schools across all charter school students remained negative.

- 27** The authors used multiple approaches in estimating the effects of charter schools on charter school students and compared results across approaches; we have focused only on the individual fixed effects model.
- 28** If the instrumental variable itself is correlated with the omitted variable or the outcome variable, then it will bias the estimated effect of the independent variable (e.g., years of schooling) on the outcome.
- 29** This two-step procedure does not generate the correct standard errors; in practice, 2SLS software packages should be used for instrumental variables estimation so that the resulting statistical inferences are correct.
- 30** Many of the studies described in this report have been the subject of some controversy. The designs and methods that researchers use have both strengths and weaknesses. Limitations exist in nearly all studies, whether experimental or observational, making it incumbent upon the investigator to acknowledge such limitations and explore alternative explanations for their results.
- 31** An alternative method uses propensity scores to weight all observations to reflect the probabilities that individuals could be in the treatment and control groups and then estimates the treatment effect on the basis of observations weighted by their propensity scores.
- 32** Cook (in press) has recently written an article in which he argues that when a regression discontinuity design is perfectly implemented and the selection process is fully observed, an unbiased causal inference can be made from the model that is produced. In the article he reviews the history of regression discontinuity designs and the assumptions that were made in its development. The article outlines when these designs can be used and why this method is superior to other known causal methods, including its strengths and limitations for estimating causal inference.