當代教育研究季刊　第二十三卷　第四期，2015 年 12 月，頁 093-119
DOI : 10.6151/CERQ.2015.2304.04

CHALLENGES AND OPPORTUNITIES FOR ESTIMATING EFFECTS WITH LARGE-SCALE EDUCATION DATA SETS

Guan Kung Saw [1*]　**Barbara Schneider** [2]

ABSTRACT

School leaders and policy makers are often faced with serious challenges when determining the allocation of scarce resources for specific programs and practices. These decisions, typically made at the district, state or federal level, have become increasingly reliant on scientifically-based evidence that can inform what programs work, for whom, and under what conditions. To answer these questions researchers draw on a variety of methodological designs and statistical techniques to make robust inferences regarding the effect of relatively recent or existing programs and/or practices. The science of estimating effects has grown considerably over the past decade, aided in part by the availability of large-scale data sets that make it possible to simulate near-experimental conditions without employing traditional methods that require randomization of units (e.g., students, schools, districts) to treatment and control situations. These methods are particularly useful especially where randomization of subjects is not feasible. This article examines the opportunities and potential statistical problems when estimating effects with large-scale data sets for education policy and research.

Keywords: *large-scale data, estimating effects, educational evaluations*

* Guan Kung Saw (corresponding author), Ph.D. Candidate, Michigan State University.
E-mail: sawguan@msu.edu
Barbara Schneider, Distinguished Professor, Michigan State University.
E-mail: bschneid@msu.edu
Manuscript received: June 2, 2015; Modified: August 10, 2015; Accepted: September 16, 2015

December, 2015, Vol. 23, No. 4, pp. 093-119
DOI : 10.6151/CERQ.2015.2304.04

Introduction

Policy makers, both globally and within the U.S., are increasingly relying on analyses of international and national large-scale data sets to inform school policy and management decisions ranging from the use of student test scores to assess teacher effectiveness (Corcoran & Goldhaber, 2013; Harris, 2011) to the allocation of resources for specific populations and programs (Greenwald, Hedges, & Laine, 1996; Hanushek, 1997; Organization for Economic Cooperation and Development [OECD], 2013). For example, data collected through the Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA) are used to describe how schools are organized within and across countries and how relationships with educational outcomes have changed over time (Martin & Mullis, 2013; OECD, 2013). In the U.S., one example of data-driven education reform over the past two decades has been the growing use of student achievement data by education authorities to identify and intervene in academically struggling schools. Since the enactment of the No Child Left Behind Act of 2001 (NCLB), all states were required to conduct annual assessments and to use the student test scores to label schools as either "made Adequately Yearly Progress (AYP)" or "did not make AYP." Schools failing to meet AYP goals for several consecutive years were subject to a set of reforms, such as replacement of school staff or school restructuring. Whether a school label such as "failing AYP" or other similar school grading system leads to improvements in school performance is still an empirical question which remains unsettled (Hemelt, 2011; Saw et al., 2015).

When evaluating a program, the methodological "gold standard" is the randomized control trial (RCT) which, to avoid issues of sample selection bias and other confounders, assigns students (or other units) to treatment or control conditions (Shadish, Cook, & Campbell, 2002). Among many methodologists, the RCT is the most rigorous method to estimate programmatic effects; however, in many education contexts, it might not be ethical or logistically possible to implement a randomized experiment (Cook, 2003). Researchers often seek alternative analytic strategies (e.g., fixed effects models, instrumental variable approaches, and regression discontinuity designs; Shadish et al., 2002) that can address the problems of selection or confounding bias. Large-scale data sets, with their typical extensive number of

sampling units, variables, and/or data collection time points, are generally seen as very useful for conducting studies in nonexperimental settings, which allow for creating treatment and control conditions, albeit not as pure as if the subjects were randomly assigned initially (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). But as some have pointed out, large-scale data sets are no "silver bullet" for designing impact studies. There are several analytic limitations in the sample design, data collection process, and method of analysis that can potentially create spurious elements that can undermine efforts to conduct rigorous impact evaluations.

This article provides an extensive discussion of related issues on the opportunities and challenges for identifying causal effects with large-scale data in education.[1] We begin by describing what is new about large-scale education data sets from the perspective of quantitative researchers, who have been long-time designers and users of various large-scale data. We then turn to review causal inference methods that address potential estimation biases in nonexperimental studies with observational data. We also provide several examples on how researchers use and analyze large-scale data sets for estimating effects that can potentially influence policy and practice in education. We highlight the value of national and state longitudinal administrative records and its potential for both advancing research and informing policy and practice in education. Finally, we discuss the opportunities and limitations of using large-scale education data sets for estimating programmatic effects.

Large-Scale Data in Education

For the past fifty or so years, policy makers have increasingly relied on large-scale data sets to inform policy decisions. For example, in the U.S., as a response to provisions of the Civil Rights Act of 1964, an extensive survey study of schools and schooling was conducted during the mid-1960s to learn about differences in the quality of education between desegregated and

[1] We define our data of interest broadly to include large-scale survey data sets (e.g., high school longitudinal study, national longitudinal study of youth) as well as administrative data (e.g., statewide student assessment data, educational personnel records) that are particularly designed or collected for research on education or school-to-work transition or for educational management. Although there are other large-scale data sets that contain some type of education information such as US Census Bureau data and Google search database, they are not a part of the scope of this article.

segregated school environments that involved nearly 600,000 students, 60,000 teachers, and 3,100 schools (Coleman et al., 1966). With the then state-of-the-art computer hardware and software, Coleman and colleagues used multiple variables to estimate the predictors and effects of desegregated schools employing multivariate regression models. Results of this study led to the contentious report of "Equality of Educational Opportunity" (EEO), which demonstrated that families were more important than schools in fostering education achievement. Many scholars argued that the EEO study set a precedent for the use of large-scale data sets in empirical education and social science research by demonstrating its value of advancing knowledge and influencing policy and practice (Heckman & Neal, 1996; Schneider, 2000; Wong & Nicotera, 2004).

Over nearly five decades, the National Center for Education Statistics (NCES), within the U.S. Department of Education, has conducted a series of large-scale survey studies designed to provide both cross-sectional nationally representative data and longitudinal cohort data of high school students, including: the National Longitudinal Study of the High School Class of 1972 (NLS: 72); High School and Beyond Study of the Class of 1980 and 1982 (HS & B); National Education Longitudinal Study of 1988 (NELS: 88); Education Longitudinal Study of 2002 (ELS: 2002); and High School Longitudinal Study of 2009 (HSLS: 09). The sample size for each of these cohorts range from about 20,000 to 35,000, with thousands of individual and contextual variables measured at multiple life stages. These data sets can be used to examine the schooling careers, cognitive, and socio-emotional development of high school students within and across cohorts, and their transition to postsecondary education, young adulthood, and the labor market. In addition to these data sets is the Common Core of Data (CCD), the state data collections that also include information on students, teachers, and schools. Most recently several states have harmonized their data systems into longitudinal ones called Statewide Longitudinal Data Systems (SLDS) which, like the national data sets, provide multiple years of education information on students, teachers, and schools including test scores. Another source of data that can be appended to both NCES national and state education data are U.S. census files (e.g., American Community Survey [ACS]) which include social security information and can describe the characteristics of the communities in which students live and where their schools are located.

Across the globe, many countries have carried out large-scale education studies that track young students as they enter school and move into the labor market. Several examples include the Taiwan Education Panel Survey (TEPS) initiated by Academia Sinica in Taiwan, the National Educational Panel Study (NEPS) conducted by the Leibniz Institute for Educational Trajectories at the University of Bamberg in Germany, and the Longitudinal Surveys of Australian Youth (LSAY) managed by the Australian Government Department of Education and Training. While all these education data sets from different countries have their own unique features, they share several similarities such as involving nationally representative student samples, covering a wide range of individual and contextual information, and following study participants for multiple years. They often serve as the major sources of data for policymakers and researchers to monitor how family, school, and community contexts affect student academic achievement, schooling experiences, and the transition from formal education to work.

In addition, several international organizations collect and maintain large-scale cross-national data sets that can be used to better understand educational processes and outcomes and to inform policy reform. Since its establishment in 1958, the International Association for the Evaluation of Educational Achievement (IEA) has conducted more than 30 research studies of cross-national student achievement and has several more in progress. These studies include the First International Mathematics Study (FIMS) in 1964, Second International Mathematics Study (SIMS) in 1980-1982, and a regular cycle of TIMSS (every four years since 1995) and Progress in International Reading Literacy Study (PIRLS) series (every five year since 2001). The IEA is also the sponsor of the 2008 Teacher Education and Development Study in Mathematics (TEDS-M), the first cross-national study of teacher preparation in 16 countries (Tatto et al., 2012). Furthermore, over the past two decades OECD has annually collected and published a comprehensive set of education indicators across countries, which covers all education levels and involves a variety of policy issues regarding student enrollment, educational attainment, school personnel, and education expenditure. The OECD database is built on the basis of national administrative sources, provided by Ministry of Education or National Statistical agencies (OECD, 2004).

Today, with the advancement of computer technology, education data systems which contain national, state, and census data could be characterized

as part of what is being termed the era of "Big Data."[2] Electronic data, including student assessment information, course-taking and transcript records, and other school personnel and administrative records, are being generated and accumulated rapidly. These data sources are seen by some as invaluable for estimating the effects of educational interventions as they can reflect "real-world" complexity, complementing carefully crafted small experimental settings. These data sets often involve large "n" (huge number of samples or the entire population of student/teacher/school), large "k" (extensive individual and contextual measures), and large "t" (multiple points in time).

Large-Scale Data as Large "n"

Even with hundreds or thousands of observations, there were often issues of statistical power when estimating effects of programs and practices for certain subgroups (e.g., minorities, students with special needs). To overcome the statistical power issue many studies oversample targeted minority group participants. The Early Childhood Longitudinal Study-Kindergarten Class of 2010-11 (ECLS-K: 2011), for example, interviewed 18,174 kindergarteners who represented a national cohort of US children attending kindergarten in 2010-2011. Among these participants, a disproportionately high number of Asian American and Pacific Islander (AAPI) students were oversampled (1,672 students) in order to achieve a minimum required sample size for AAPI group that can meet the study's precision goals (Tourangeau et al., 2015).

Nowadays, large "n" data sets with tens of millions of subjects are quite common. For instance, the California Department of Education, through its Longitudinal Education Data Systems (LEDS), makes available information on their K-12 public school system collected since 1993-1994. The data set include nearly 6 million students, 280,000 teachers, and 10,000 schools for each of 20 years. These large "n's" are novel and valuable for researchers in terms of their ability to capture and provide sufficient sample size for certain minority groups, which may be overlooked in traditional survey studies with probability sampling methods. For example, with the California data system, Pang, V., Han, and Pang, J. (2011) were able to disaggregate the academic performance of 13 AAPI subgroups and empirically test the "model minority"

[2] The term "Big Data" has been widely used to describe gigantic digital data sets held by corporations (e.g., Google, Facebook), government, and other institutions (see National Research Council, 2012; Mayer-Schönberger & Cukier, 2013). In this article, we focus on large-scale data that are collected through surveys or administrative systems, not social media websites.

myth, which positions that the Asian community as a highly-achieving minority group. Their study sample included 1,025,205 observations from 2003 to 2008. Among them, there were about 750,000 whites (73.4%) whereas the numbers of AAPI subgroup students ranged from the 63,860 Filipinos (6.2%, the largest subgroup) to 1,169 Guamanians (0.1%, the smallest subgroup), which yielded adequate sample sizes for statistical analyses and meaningful comparisons. Their findings showed that a majority of AAPI ethnic groups scored at significantly lower levels than white peers both in reading (9 out of 13 AAPI groups) and math (7 out of 13 AAPI groups), thus disputing the premise of the model minority myth. They argued that educational policies and statistical analyses in which student performance is measured using AAPI aggregate level samples lead to overgeneralized findings and unidentified achievement gap issues among various subgroups.

Large-Scale Data as Large "k"

Aside from big sample sizes, much of the modern large-scale education data provides an extensive array of variables (large "k") with information on individual characteristics that previously were difficult to observe or measure. In education, an extensive number of learning activities and performance outcomes of students, such as school enrollment, transfer, attendance/ absenteeism, promotion/retention, course-taking, earned credits, and test scores, are "digitized"; i.e. created and transmitted into electronic transcripts or administrative records. These digitized data are invaluable resources for enabling social scientists to characterize and reconstruct individual student's entire schooling experiences, which can serve as novel, powerful variables for testing research hypotheses or predicting certain educational outcomes including postsecondary attendance and completion.

Having large "k" in the data also means there is more information on study subjects, which, when combined, can help increase the predictive power of statistical models. In recent years the use of predictive analytics in schooling system has started to proliferate. Educators analyze comprehensive and multidimensional student data to anticipate which children are likely to drop out of school or experience some other undesirable outcomes, often times for targeting resources to students at risk. One example is the Consortium on Chicago School Research (CCSR) which developed an "on-track" indicator as an early-warning predictor of high school graduation for more than 20,000 ninth graders in Chicago Public Schools each year

(Allensworth, 2013; Allensworth & Easton, 2005). The "on-track" indicator is constructed based on only two 9th grade variables: (a) the number of full credits earned, and (b) the number of core course failures. These two measures together can correctly predict 80% of which students will graduate from secondary school. CCSR researchers found that no other combinations of student characteristics (including demographics, pre-high school test scores, and school enrollment patterns) have higher predictive power than the "on-track" indicator.

Large "k" data sets are advantageous, yet analysts may find them overwhelmed by the massive numbers of variables (in some cases, "k" even larger than "n") and the complexity of data structure. For example, the National Longitudinal Survey of Youth 1997 (NLSY97) is designed to document the transition from school to work and into adulthood. It collects extensive information about their respondents (8,984 individuals) annually and through a variety of means, including survey questionnaires, computer-adaptive test, and both high school and postsecondary transcripts. As of May 2015, about 68,000 distinct variables are available in the NLSY97 publicly available data set. NLSY97 data users are able to merge individual information with a vast number of variables from other national data sets such as the American Community Survey (ACS) and Integrated Postsecondary Education Data System (IPEDS).

Large-Scale Data as Large "t"

One additional important feature of large-scale data are the multiple number of data collection time points (large "t"), in addition to the number of "n" and continually expanding "k's." For example, Covay Minor, Saw, Frank, Schneider, and Obenauf (2015) analyzed eight years (2004-2011) of statewide administrative data from Michigan, supplemented with the Common Core of Data (CCD) and Bureau of Labor Statistics (BLS), to examine how state policy changes and economic conditions affect teacher turnover in schools by urbanicity. They found that both a new statewide curricular reform announced in 2006 and the recent economic recession which occurred in late 2000s had an immediate or lagged effect on school-level teacher turnover rate, varying across urban, suburban, town, and rural areas. These analyses can capture the impact of policy shocks or larger contextual factors on teacher turnover and its variation by school locale. This type of analysis is only possible with the availability of longitudinal statewide information on teachers and schools

before and after the implementation of a new state-level educational reforms or periods of economic recession.

Large "t" data sets are also particularly useful for trend analyses in educational policy and practice. At the macro level, trend analysis can help answer such questions as how per-pupil expenditures, teacher salaries, and pupil-teacher ratio in K-12 schools have changed over time. Heckman and LaFontaine (2010) drew upon two yearly indicators collected and provided by the U.S. Common Core of Data (CCD) that contain information on student enrollment in each grade level and number of high school diplomas issued, to portray both the levels and trends in American high school graduation rates from 1960 to 2005. At the micro level, experts in educational technology use longitudinal data on instructional practices and student responses to identify what changes have occurred in student learning over time (Bienkowski, Feng, & Means, 2012).

Large-scale data sets with a huge number of samples, measures, and time points represent major resources for expanding the areas of inquiry and for increasing the predictive power of models and the precision of estimates. Some of the most important theoretical, methodological, and substantive work in education and social science research has resulted from empirical studies using large-scale data sets (Schneider et al., 2007). The availability of large collections of data, however, does not automatically translate into an improved ability to identify programmatic effects, which are crucial in informing policy debates and theoretical interpretations. With any scale of nonexperimental data sets, researchers need to be aware of several important problems of potential estimation bias, specifically selection bias, where participants in the treatment and comparison groups are not randomly assigned to groups.

Estimating Effects With Large-Scale Data

Randomized controlled trials (RCTs), in which participants are randomly assigned to receive an intervention (or treatment), are considered to be the most credible research designs that can yield unbiased estimates of the effects of a specific program or practice effects. By randomly assigning participants to treatment and control groups the treatment status is independent of the baseline characteristics of participants; therefore, any resulting benefit or harm can be attributed to the impact of an intervention. RCTs are often not feasible,

especially in education or social settings. Thus researchers often employ nonexperimental or quasi-experimental methods with observational data.

Analyzing experimental and nonexperimental sample, even with large-scale data, is challenging as researchers cannot necessarily be aware of all the unobserved conditions that can bias the estimates. It may be the case that individuals who receive an intervention may differ systematically from those who do not. In formal terms, this problem is referred to as sample selection bias. For instance, when evaluating the effects of a teacher induction program on retention, if a researcher uses only participants who complete the program, the sample may over-represent teachers who were at low or moderate risk of turnover and under-represent high-risk teachers who would leave their teaching position prior to entering or completing the program (Ingersoll & Strong, 2011). To adjust for sample selection bias methodologists have formalized several strategies (see overviews by Vella, 1998; Winship & Mare, 1992). One of the well-established techniques is Heckman's two-stage procedure (Heckman, 1976, 1979), which can help adjust for nonrandom exclusion from the sample when certain assumptions (e.g., independence, normality) hold or the selection process is correctly modeled (Briggs, 2004; Stolzenberg & Relles, 1997; Winship & Mare, 1992).

Another approach to address selection bias is to adjust outcomes for relevant observed variables that are associated with both the outcome and independent variables of interest, which has been described as observable selection bias (Barnow, Cain, & Goldberger, 1980). Basically, conventional multiple regression models aim to correct such selection bias by controlling for potential confounding factors. In some cases, adjusting for observable selection bias is also necessary in an RCT where there is an undesirable imbalance in the number of comparison groups or a problem of imperfect compliance (Ivanova, Barrier, & Berger, 2005; Meinert, 1986). Nonetheless, some unobserved selection factors (or omitted variables, resulting in unobserved differences) may still exist and could bias the estimates. For example, in evaluating the effects of Head Start, a public preschool program for disadvantaged children in the U.S., researchers employed procedures for minimizing observable selection bias to adjust for differences in family background between Head Start participants and nonparticipants. Head Start programs may attract parents who have higher educational expectations for their children. Even within a same household parents may choose to enroll the least able child or a more able one in Head Start program, depending on their

parenting strategy (Currie & Thomas, 1995, 1999). Thus it is hard to disentangle parental expectations from subsequent achievement gains for those who participate in Head Start (as not all will have low scores) from those who do not participate in the program.

The existing literature on estimating effects with observational data has been growing.[3] The following presents a brief review of several analytic strategies that are particularly useful for making inferences with large-scale data sets, including fixed effects models, instrumental variables, regression discontinuity designs, and comparative short interrupted time series. These methods have been widely applied to large-scale education data to approximate randomized controlled experiments. All these nonexperimental methods have their limitations and require certain assumptions to be satisfied in order to be a valid research design (for advanced discussion on theoretical foundations and statistical proof of these methods, see Angrist & Pischke, 2009; Wooldridge, 2010, 2012).

Fixed Effects Models

Although data sets with large "n" and "k" can be very useful in controlling for sample and observable selection bias, there could still be important confounders that are unmeasured. One possible strategy to control for unobserved characteristics is fixed effects models, which exploit within-unit variation (Chamberlain, 1982; Heckman & Robb, 1985). The central idea is to use each unit as its own control such that any unobserved effect attributable to the unit is included in the model. Taking advantage of large-scale data in which there are multiple observations nested within units or clusters (e.g., cohorts, classrooms), the unobserved confounders that are constant within units can be controlled for through estimation in deviations from group means (or called the group mean centering approach in multilevel literature, see Raudenbush & Bryk, 2002).[4] There are two conditions for employing fixed effects models: (a) the dependent variable must be measured

[3] We recommend two comprehensive publications that were written to help guide policy makers, educators, and researchers in understanding causal estimation with experimental and observational designs. Schneider et al. (2007) provide non-technical introduction on the causal inference theories and methods, whereas Murnane and Willett (2011) offer technical guidance on analytical methods for evaluating the causal impacts of educational policies/programs/interventions.

[4] Note that the term "fixed effects" means somewhat different things in the econometric and hierarchical linear modeling (HLM) literature. In HLM, fixed effects refer to the non-random parts of the coefficient estimates (e.g., estimators at the level 1), whereas in econometrics it refers to the idea that the explanatory variables are treated as if the observed quantities were non-random.

for at least two individuals within a unit, and (b) the independent variable of interest must change in value across individuals for a considerable portion of the sample.

For instance, drawing upon the administrative data from the New York City (NYC) Department of Education, Boyd, Lankford, Loeb, Ronfeldt, and Wyckoff (2011) exploited the feature of teachers grouped within schools to explore the dynamics between teacher transferring and hiring within NYC schools from 2007 to 2008. They model a teacher's likelihood of applying for transfer as a function of teacher characteristics and school-level fixed effects, which control for unobserved characteristics of schools (e.g., working conditions, school climate) that might be correlated with a teacher's likelihood of requesting a transfer. Their findings suggested that teachers with pre-service qualifications (such as high certification exam scores, attendance at competitive colleges) are more likely to apply for transfer while practice-based quality measures (such as teacher experience) were less likely to request transfers. Boyd et al.'s (2011) study also demonstrated the benefits of large-scale education data sets. Because the actual transferring of teachers across schools is a two-sided choice, teachers' intention to transfer and school preferences in hiring, the investigators empirically examined the supply- and demand-side theories by linking the teacher applications-to-transfer data to work history files of all active teachers in NYC (80,898 teachers). They found that both supply- and demand-side explanations account for the dynamics of the open labor market in education. Specifically, their analyses of work history files showed that the likelihood of actual transfer is not significantly different between black/Hispanic and white teachers. The results might be misinterpreted to suggest that there are no differences in transferring or hiring by teacher race. In fact, the applications data revealed that black/Hispanic teachers are significantly less likely than white peers to request transfer (low supply). Meanwhile, among teachers who request a transfer, black and Hispanic teachers are significantly more likely than white teachers to get hired (high demand).

If the unobserved confounders are constant over time, the time-invariant characteristics can be removed in the models with panel data (large "k") where there are repeated observations on individuals (Angrist & Pischke, 2009; Wooldridge, 2010). One example of fixed effects models in longitudinal settings is a study conducted by Bifulco and Ladd (2006), where the researchers analyzed a large panel sample of North Carolina students (5

cohorts and in total 495,943 students) to estimate the impact of charter schools on student achievement. The primary challenge in evaluating charter school effects emerges from the fact that charter school students are self-selected and are likely to differ in unobserved traits (e.g., motivation, value placed on education) from their counterparts who choose to stay in traditional public schools. With student-level longitudinal data, Bifulco and Ladd were able to track individual students over time (two to five follow-up years) and identify whether they were enrolling in a charter or public school in any given year. They used repeated observations on individual students to control for individual fixed effects, which are technically comparing the test score gains of students while in charter schools with the test score gains made by the same students while in traditional public schools. Results showed that students have significantly smaller test score gains than they would have in public schools. The advantage of such individual fixed effects models is that it essentially is a comparison of within-individual, not between-individual, thus reducing the problem of self-selection bias.

Instrumental Variables

A second analytic approach to address selection bias is to use an "instrumental variable (IV)" in the regression models (Angrist, Imbens, & Rubin, 1996; Angrist & Krueger, 2001). In many cases in education, lawmakers and researchers are interested in a reliable estimate of the impact of one variable (X) on an outcome (Y). For instance, what is the effect of summer school on student achievement? With observational data, standard regression estimators of the summer school enrollment are likely to be biased because of unobserved selection bias. Students who participate in summer school are likely to have lower levels of cognitive growth, which is a difficult-to-measure characteristic that is associated with both attending summer school and outcome scores. In such circumstances IV methods can be used to obtain consistent estimators in the presence of omitted variables. The key idea is to use a third variable (IV) to isolate variation in the variable of interest (X) that is uncorrelated with the selection bias, and to use this variation to identify its causal effect on an outcome measure (Y). This regression analysis is also called two-stage least squares (2SLS or TSLS) method. There are two conditions that must be satisfied for successful IV estimation, in which a good IV should (a) be correlated with the treatment (e.g., summer school participation) but (b) be uncorrelated to the omitted variables (e.g., cognitive

growth), and thus has no association with the outcomes (e.g., test scores), except through the treatment.

In observational studies it is difficult to obtain a strong IV, while in randomized controlled trials treatment assignment can be used as a valid IV, which typically is statistically independent of unobserved factors correlated with the outcome, and possibly influences the outcome only through its impact on treatment receipt. Sondheimer and Green (2010) demonstrated the use of treatment assignment as an IV for identifying the effect of education on voter turnout, using the data from the Perry Preschool Experiment and the Tennessee Student Teacher Achievement Ratio (STAR) Experiment, combined with voter turnout data supplied by commercial vendors gathering voter registration and turnout information. The authors used the treatment assignment to estimate the intent-to-treat effect of being randomly assigned to treatment groups (i.e., preschool program in Perry or smaller class in STAR) on high school graduation, which in turn is used to predict voter turnout. The findings indicate that both treatment assignments in Perry and STAR experiments have a positive impact on education attainment, which positively related to the likelihood of voting. With an IV analysis, the researchers can obtain consistent estimates on the effect of education on voter turnout while taking into account the possibility that high school graduation is associated with unobserved causes of voting behavior.

Regression Discontinuity Designs

A third research method that can be used to estimate causal effects with observational data are regression discontinuity (RD) designs. Although RD designs were first introduced more than half a century ago (Thistlethwaite & Campbell, 1960), it did not attract much attention among social scientists until relatively recently. In an investigation of the effect of remedial education on student achievement, Jacob and Lefgren (2004) drew upon the student data (74,260 third graders and 73,634 sixth graders from 1997 to 1999) from Chicago Public Schools and exploited the discontinuous change in probability of attending summer school and being retained due to levels of student academic performance. They compared individuals who were just above or below the threshold for (a) summer school requirement, and (b) promotion to the next grade. The assumption was that low-performing students who just barely passed the threshold for summer school requirement or promotion would be identical to their counterparts who fell just below the threshold, thus

reducing the threat of selection bias. Their RD evidence showed that both summer school and grade retention have a positive but modest effect on student test scores for 3rd-grade students but no impact for 6th-grade students.

The rationale behind the RD design is that individuals with assignment scores just below the threshold (who did not receive the treatment) were good comparisons to those just above the threshold (who did receive the treatment). For example, it is often the case that student are chosen to be enrolled in a gifted/talent program (Bui, Craig, & Imberman, 2014) or to receive college financial aid (van der Klaauw, 2002) based on their levels of academic performance. Analysts estimate treatment effects by comparing mean outcomes among observations just above and below the cutoff point (who are likely to be similar on a set of observable and unobservable characteristics), conditioned on the assignment score. The difference between these two conditional means can be understood as a discontinuity in the regression function. Therefore, the size of the discontinuity captures the magnitude of the treatment effect. A valid RD design requires a situation where individuals have imprecise or no control over assignment scores, which implies that the variation in treatment status will be "as good as" randomly assigned around the cutoff (Lee & Lemieux, 2010). Thus, the threat of selection bias is reduced to the greatest extent.

Another example of an RD design is a study conducted by Saw et al. (2015), which evaluated the causal impact of being labeled as a Persistently Lowest Achieving school, using the statewide longitudinal data from Michigan. Starting in 2010 the Michigan Department of Education (MDE) annually published a Top-to-Bottom (TTB) school ranking list constructed based on school achievement levels and improvement rates in the past four years. Schools that fall in the bottom 5% are classified as PLA schools. In 2010, there were 56 regular high schools placed on the PLA list. By using school percentile ranking (the assignment variable), Saw and his associates were able to leverage the RD design to obtain unbiased estimates of the impact of the PLA list on school performance. They found that PLA schools increased their student achievement scores in writing, with marginal evidence in mathematics and social studies, and no evidence of increases in reading and science in year one, compared with those schools which were just above the cut-off for being on the PLA list. The study also serves as an example of making use of administrative data that have "universal" coverage of all public schools (large "n") in an entire state (i.e., Michigan). With the statewide data,

they were able to include all PLA schools in Michigan (n = 43, excluding closed schools and ineligible samples) in their RD analysis, together with relevant school measures as covariates, which yield sufficient statistical power to detect an effect of PLA treatment.

Comparative Short Interrupted Time Series

The comparative short interrupted time series (CS-ITS) design is one of the increasingly used non-experimental methods in educational evaluation, which can produce unbiased estimates of treatment effects (Bloom, 2003; Somers, Zhu, Jacob, & Bloom, 2013; St.Clair, Cook, & Hallberg, 2014). CS-ITS is a special case of interrupted times series (ITS) designs (Cook & Campbell, 1979; Shadish et al., 2002). Fundamentally, ITS designs begin from a long series of repeated measurements on an outcome variable. When an intervention occurs, the time series can be divided into preintervention and postintervention segments. The effect of intervention or treatment is then estimated by comparing the means and slopes of the dependent variable in the two periods. One internal validity threat to the ITS designs involves external forces (unobserved factors) that can contaminate the causal relationship between the intervention and outcomes. This threat could be more problematic in "short" ITS designs where time series data are only available for limited time points before and after the treatment. Adding a comparison time series with no treatment provides a new counterfactual estimate which helps control for other local changes that might have confounded postintervention time points (Bloom, 2003). Thus, comparative short ITS (CS-ITS) is a more practical case for ITS designs in education evaluations where important measures on student, teacher, and school are typically collected annually or in a longer time interval.

An empirical example of using CS-ITS designs is Dee and Jacob (2011), which used the state-level panel data (1992-2007) on student achievement scores from the National Assessment of Educational Progress (NAEP) to identify the impact of No Child Left Behind (NCLB), an US federal policy of school accountability launched in 2001. By reviewing consequential school accountability policies adopted at the state level prior to the implementation of the NCLB, the investigators first classified all states into two major groups: (a) comparison group, which implemented NCLB-like accountability prior to NCLB, thus arguably less affected by NCLB, and (b) treatment group, which had no consequential school accountability policies prior to NCLB. Then they

employed the CS-ITS models to compare the deviation from prior achievement trends among the treatment group with the analogous deviation for the comparison group (final sample size = about 250 state-by-year observations). Their results suggested that NCLB has a significant positive impact on math performance of 4th-graders and 8th-graders but has null effect on reading performance for 4th-graders. Dee and Jacob's (2011) study also represents a valuable case of educational evaluations that utilize cross-sectional state-representative data with comparable outcome measures for multiple points in time before and after (large "k") a policy or important event.

Following the empirical strategy demonstrated in Dee and Jacob (2011), two recent studies drew upon large-scale national survey data sets to estimate the impact of NCLB on school resources and practices (Dee, Jacob, and Schwartz, 2013), and teachers' work environments and job attitudes (Grissom, Nicholson-Crotty, & Harrington, 2014). Dee et al. (2013) analyzed the data from Common Core of Data's Local Education Agency (School District) Finance Survey for years 1995-2008 (142,607 district-by-year observations) and from Schools and Staffing Survey (SASS) for years 1994-2008 (36,000 teacher-year observations, 16,500 principal-year observations). The authors found that NCLB led to district-level increases in school spending, teacher compensation, and the share of elementary school teachers with an advanced degree. They also found that NLCB led schools to reallocate instructional time away from science and social studies and toward the tested subjects of math and reading. Grissom et al. (2014) used the same SASS data as in Dee et al.'s (2013) study but focused on the teacher-reported measures on their work environments and job satisfaction; measures which critics argued to be negatively affected by the NCLB. Their analysis showed that NCLB has had small or null effects on those teacher-reported variables.

Limitations of Using Non-Random Assignment Procedures to Estimate Impacts

There are several limitations of using large-scale data to estimate the effects of specific programs. First, in many large-scale data sets, although the information collected is extensive and potentially interesting, the scope of measures may not be nearly as diverse as the range of topics covered in education and social science (Conaway, Keesler, Schwartz, 2015; White &

Breckenridge, 2014). For example, there has been a considerable focus on the impact of Charter schools on student performance. In 2011-2012, roughly 2.1 million students are enrolled in charter schools nationwide, representing 4.2% of student enrollment (NCES, 2015a). A search on Web of Science revealed that 234 peer-reviewed journal articles concerning charter schools were published in the U. S. in 2010-2014, compared with 289 articles on No Child Left Behind (NCLB) and 215 articles on English learners. The NCLB policy affects all students whereas English learners account for 9.1% of student population (about 4.4 million) in the country (NCES, 2015b). Why is there so much academic work on charter schools? One possible explanation is that in many circumstances charter school enrollment is conducted on the basis of a lottery. Researchers may be lured to the possibility of a convincing impact of charter school based on the random assignment process of lottery, combined with detailed administrative data. However, do charters deserve this disproportionate attention?

Unlike large-scale surveys conducted by the NCES or other research institutions, many education data sets that consist of student tracking and school administrative information are not collected primarily to meet the needs of researchers but rather as a response to reporting system required by governments or other agencies (Dynarski & Berends, 2015). Researchers often have little or no control over the data collection process. These administrative data sets are, for the most part, not generated from instruments and methods designed to produce valid and reliable measures. One notable example is the indicator of students receiving free or reduced price lunch, which is frequently used as a proxy measure for children living in poverty. In the U.S., just over half of public school students were eligible for free/reduced price lunches in 2012-2013, according to the statistics released by the Common Core of Data (NCES, 2015c). In contrast, an estimate computed based on the American Community Survey, conducted by the US Census Bureau, suggests that the poverty rate of public school children was only 22.6% in the same year (NCES, 2015d). The poverty status defined by the Census Bureau is constructed by using a set of financial income thresholds that vary by family size and composition, which is arguably a more accurate measure of poverty level.

Undeniably, any data collection may be contaminated by some type of inaccuracy in measurement. The problem can be more severe and complicated in large-scale education data sets, given the massive amounts of information

involved. Administrative data, a major source of modern large-scale data sets, are typically collected through electronic systems that are especially likely to contain erroneous, miscoded, fragmented, and incomplete information, attributable, in part, to staffs' documentation workloads, poor user-interface design, and other factors (Hoffman & Podgurski, 2013). Some automated processing of electronic data at various stage of data collection can also prevent the opportunities for data quality checking or verification by humans. Moreover, contemporary large-scale education data sets are typically assembled from disparate data sources with high dimensionality at multiple time points. The data construction processes oftentimes involve transforming raw data files from different supplying units into a standard form, linking records together, and creating new variables of the data. These activities may introduce unnecessary errors and create noise and poor-quality data that lead to biased and invalid estimates and inferences.

Challenges and Prospects

One of the major concerns with the emergence of big data in education is protection of human subjects' confidentiality. The risk of identifying individuals or institutions in large-scale data sets remains a concern, largely due to the huge amounts of detailed information available even after accounting for researchers efforts to deidentify individuals. For example, if the micro-data of students include information such as first and last names, birthdate, gender, race/ethnicity, and zip code, the reidentification risk is exceptionally high. Although there is a federal law, the Family Educational Rights and Privacy Act (FERPA), to protect the privacy of a student's education record, the regulations are mostly outdated as the law was enacted four decades ago in 1974. Following a call by President Obama in January 2015, a new bill titled the Student Digital Privacy and Parental Rights Act of 2015 has been recently introduced, aiming to provide an up-to-date framework for addressing current privacy regulation issues on student data.

From a data user's viewpoint, a closely related issue (or a trade-off) to privacy and confidentiality concerns is data access. Research on important educational topics such as student achievement gaps, accountability, and teacher effectiveness require large-scale data that offer a large, representative sample (large "n") with detailed information on individuals and institutions (large "k") in longitudinal settings (large "t"); all of which are critical for a

rigorous research design. For many national survey data sets, there are well-established procedures for accessing and analyzing the data. However, for administrative data that are crucial for cutting-edge empirical studies and for credible public policy evaluation, programs or systems to provide access to these data are still underdeveloped (Card, Chetty, Feldstein, & Saez, 2010). Many European countries, such as Denmark, Norway, and Sweden, demonstrate that broader access is possible with proper deidentification processes and application protocols, and that expanding access to administrative can have a significant impact on research output, which would better serve policy making and practice (Einav & Levin, 2014).

Another major challenge to the development of large-scale data sets for educational evaluation and policy analysis is the building of human capacity and data literacy (Mandinach, 2012; National Research Council [NRC], 2002). For the promises and benefits of the large-scale education data sets to be accomplished, both data developers and data users must be well-trained experts in data science or related fields. For data designers or producers, who are typically federal or private research agencies, building high-level scientific human resources means attracting, training, and retaining an adequate number of qualified leaders and staffs who have knowledge of relevant content in education and of rigorous research design, theory, data collection methods, and analysis techniques (National Research Council [NRC], 2002). For the data users or researchers, acquiring cutting-edge knowledge and knowing how to use large-scale data sets means developing and implementing research plans that can produce rigorous empirical evidence for informing educational policy and practice.

The fostering of a scientifically competent community of educational researchers is becoming more imperative to meet the growing needs and challenges of developing and making use of large-scale education data sets. Two particular trends are worth highlighting. One is blending education data from diverse data collection approaches such as administrative system, survey, and randomized controlled experiments that provide different types of information but are complementary to each other (Dynarski, 2014; Grossman, 2014). Chetty et al.‘s (2011) prominent work on studying the long term effects of class size, which combines experimental data (collected in 1986) with administrative data of US tax records (available from 1996 to 2008), serves as an excellent example in this research area. The other trend is linking education data to various databases designed and managed by other statistical agencies

such as those who are in charge of collecting information on health, labor and industry, corrections, and so forth (Card et al., 2010). An ongoing effort in the state of Washington provides a representative case in this manner. The newly established Education Research and Data Center (ERDC) is working on linking their statewide longitudinal education data across eleven state agencies, including Department of Health, Department of Licensing, and Department of Labor and Industries (ERDC, 2013), which will offer tremendous opportunities for educational and interdisciplinary research.

Although limitations and concerns exist, the development and utilization of large-scale data sets offer enormous potential for advancing educational evaluation and policy analysis. New statistical procedures are being developed and refined all the time, allowing for greater precision and consistency with large-scale data when estimating the effects of particular educational programs, especially when randomized trials are not feasible. Additionally, there are some methods that are particularly useful when trying to understand effects over time, although, much like randomized trials, they are subject to the vagaries of fidelity of implementation and other confounders that may affect the outcome measures. Nonetheless, as extensive resources are being allocated to the schooling of children, policy makers and educators are increasingly relying on studies based upon large-scale data sets to obtain more reliable evidence regarding policies or practices that are effective in promoting excellence and equity in education.

Acknowledgement

This work is supported by the U.S. National Science Foundation (NSF), through Grant DRL-1316702, and the U.S. National Institutes of Health (NIH), through Grant 1R01GM102637-01 (Principal Investigator: Barbara Schneider). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the organizations. Please direct all correspondence to: Guan Saw (sawguan@msu.edu); Michigan State University; 516 Erickson Hall, East Lansing, MI 48824.

References

Allensworth, E. M. (2013). The use of ninth grade early warning indicators to improve Chicago schools. *Journal of Education for Students Placed at Risk, 18*(1), 68-83.

Allensworth, E. M., & Easton, J. Q. (2005). *The on-track indicator as a predictor of high school graduation*. Chicago, IL: Consortium on Chicago School Research.

Angrist, J. D., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444-472.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives, 15*(4), 69-85.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Barnow, B., Cain, G., & Goldberger, A. (1980). Issues in the analysis of selectivity bias. In E. Stromsdorfer & G. Farkas (Eds.), *Evaluation studies* (Vol. 5, pp. 43-59). Beverly Hills, CA: Sage.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief.* Washington, DC: US Department of Education, Office of Educational Technology.

Bifulco, R., & Ladd, H. F. (2006). The impacts of charter schools on student achievement: Evidence from North Carolina. *Education Finance and Policy, 1*(1), 50-90.

Bloom, H. S. (2003). Using "short" interrupted time series-analysis to measure the impacts of whole-school reforms. *Evaluation Review, 27*(1), 3-49.

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2011). The role of teacher quality in retention and hiring: Using applications-to-transfer to uncover preferences of teachers and schools. *Journal of Policy Analysis and Management, 30*(1), 88-110.

Briggs, D. C. (2004). Causal inference and the Heckman model. *Journal of Educational and Behavioral Statistics, 29*(4), 397-420.

Bui, Sa, A., Craig, S. G., & Imberman, S. A. (2014). Is gifted education a bright idea? Assessing the impact of gifted and talented programs on students. *American Economic Journal: Economic Policy, 6*(3), 30-62.

Card, D., Chetty, R., Feldstein, M., & Saez, E. (2010). *Expanding access to administrative data for research in the United States* (NSF White Paper ID 112 on Future Research in the Social, Behavioral & Economic Sciences). Arlington, VA: National Science Foundation.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics, 18*(1), 5-46.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics, 126*(4), 1593-1660.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, F., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Conaway, C., Keesler, V., & Schwartz, N. (2015). What research do state education agencies really need? The promise and limitations of state longitudinal data systems. *Educational Evaluation and Policy Analysis, 37*(1), 16S-28S.

Cook, D. (2003). Why have educational evaluators chosen not to do randomized experiments? *Annals of the American Academy of Political and Social Science*, *589*(1), 114-149.

Cook, T. D., & Campbell, D. T. (1979). *Experimental and quasi-experimental designs for research.* Chicago, IL: McNally.

Corcoran, S., & Goldhaber, D. (2013). Value added and its uses: Where you stand depends on where you sit. *Education Finance and Policy, 8*(3), 418-434.

Covay Minor, E., Saw, G. K., Frank, K. A., Schneider, B. L., & Obenauf, K. (2015). *External contextual factors, teacher turnover, and student achievement: The case of Michigan high schools.* Manuscript submitted for publication.

Currie, J., & Thomas, D. (1995). Does Head Start make a difference? *The American Economic Review, 85*(3), 341-364.

Currie, J., & Thomas, D. (1999). Does Head Start help Hispanic children? *Journal of Public Economics, 74*(2), 235-262.

Dee, T. S., & Jacob, B. A. (2011). The impact of the No Child Left Behind Act on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418-446.

Dee, T. S., Jacob, B. A., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis, 35*(2), 252-279.

Dynarski, S. (2014, November). *Building better longitudinal surveys (on the cheap) through links to administrative data*. Prepared for the National Academy of Education's Workshop to Examine Current and Potential Uses of NCES Longitudinal Surveys by the Education Research Community, Washington, DC.

Dynarski, S., & Berends, M. (2015). Introduction to Special Issue: Research using longitudinal student data systems: Findings, lessons, and prospects. *Educational Evaluation and Policy Analysis, 37*(1), 3S-5S.

Education Research and Data Center (2015). *State of Washington Education Research and Data Center: State and national education and workforce data resources*. [Online] http://www.erdc.wa.gov/data/datalinks/default.asp

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science, 346*(6210), 715-723.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*(3), 361-396.

Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the effects of No Child Left Behind on teachers' work environments and job attitudes. *Educational Evaluation and Policy Analysis, 36*(4), 417-436.

Grossman, P. (2014, November). *Collecting evidence of instruction with video and observation data in NCES surveys*. Prepared for the National Academy of Education's Workshop to Examine Current and Potential Uses of NCES Longitudinal Surveys by the Education Research Community, Washington, DC.

Hanushek, E. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141-164

Harris, D. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement, 5*(4), 475-492.

Heckman, J. J. (1979). Sample selection as a specification error. *Econometrica, 47*(1), 153-161.

Heckman, J. J., & LaFontaine, P. A. (2010). The American high school graduation rate: Trends and levels. *Review of Economics and Statistics, 92*(2), 244-262.

Heckman, J. J., & Neal, D. (1996). Coleman's contribution to education: Theory, research styles and empirical research. In J. Clark (Ed.), *James S. Coleman* (pp. 81-102). London, England: Falmer Press.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics, 30*(1-2), 239-267.

Hemelt, S. W. (2011). Performance effects of failure to make adequate yearly progress (AYP): Evidence from a regression discontinuity framework. *Economics of Education. Review, 30*(4), 702-723.

Hoffman, S., & Podgurski, A. (2013). Big bad data: Law, public health, and biomedical databases. *The Journal of Law, Medicine & Ethics, 41*(1), 56S-60S.

Ingersoll, R., & Strong, M. (2011). The impact of induction and mentoring programs for beginning teachers: A critical review of the research. *Review of Education Research, 81*(2), 201-233.

Ivanova, A., Barrier, R. C., & Berger, V. W. (2005). Adjusting for observable selection bias in block randomized trials. *Statistics in Medicine, 24*, 1537-1546.

Jacob, B., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics, 86*(1), 226-244.

van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review, 43*(4), 1249-1287.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281-355.

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist, 47*(2), 71-85.

Martin, M. O., & Mullis, I. V. S. (Eds.) (2013). *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade-Implications for early learning.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and IEA.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. London, England: John Murray.

Meinert, C. L. (1986). *Clinical trials, design, conduct, and analysis*. Oxford, England: Oxford University Press.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford, England: Oxford University Press.

National Center for Education Statistics (2015a). *Table 216.20. Number and enrollment of public elementary and secondary schools, by school level, type, and charter and magnet status: Selected years, 1990-91 through 2012-13*. Retrieved from https://nces.ed.gov/programs/digest/d14/tables/dt14_216.20.asp?current=yes

National Center for Education Statistics (2015b). *Table 204.20. Number and percentage of public school students participating in programs for English language learners, by state: Selected years, 2002-03 through 2012-13*. Retrieved from https://nces.ed.gov/programs/digest/d14/tables/dt14_216.20.asp?current=yes

National Center for Education Statistics (2015c). Table 204.10. *Number and percentage of public school students eligible for free or reduced-price lunch, by state: Selected years, 2000-01 through 2012-13.* Retrieved from https://nces.ed.gov/programs/digest/d14/tables/dt14_204.10.asp?current=yes

National Center for Education Statistics (2015d). *Table 102.70. Number and percentage of students in prekindergarten through grade 12 living in poverty, by control of school: 2000 through 2013.* Retrieved from http://nces.ed.gov/programs/digest/d14/tables/dt14_102.70.asp

National Research Council (2002). *Scientific research in education* (report from the Committee on Scientific Principles for Education Research, in R. J. Shavelson, & L. Towne (Eds.), Center for Education, Division of Behavioral and Social Sciences and Education.) Washington, DC: National Academy Press.

National Research Council (2012). *Big data: A workshop report*. Washington, DC: The National Academies Press.

Organization for Economic Cooperation and Development (2004). *OECD handbook for internationally comparative education statistics: Concepts, standards, definitions and classifications.* Paris, France: OECD.

Organization for Economic Cooperation and Development (2013). *PISA 2012 results: What makes schools successful? Resources, policies and practices* (Vol. IV). Paris, France: OECD.

Pang, V., Han, P., & Pang, J. (2011). Asian American and Pacific Islander students: Equity and the achievement gap. *Educational Researcher, 40*(8), 378-389.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd Ed.)*. Newbury Park, CA: Sage.

Saw, G. K., Schneider, B. L., Frank, K. A., Chen, I. C., Keesler, V., & Martineau, J. (2015). *The impact of being labeled as a persistently lowest achieving school: Regression discontinuity evidence on school labeling*. Manuscript submitted for publication.

Schneider, B. (2000). Social systems and norms: A Coleman approach. In M. Hallinan (Ed.), *Handbook of sociology of education* (pp. 365-385). New York, NY: Kluwer Academic/Plenum Publishers Corporation.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs* (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

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

Sondheimer, R. M., & Green, D. P. (2010). Using experiments to estimate the effects of education on voter turnout. *American Journal of Political Science, 54*(1), 174-189.

Somers, M., Zhu, P., Jacob, R., & Bloom, H. (2013). *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation* (MDRC working paper in research methodology). New York, NY: MDRC.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation, 35*(3), 311-327.

Stolzenberg, R. M., & Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review, 62*(3), 494-507.

Tatto, M. T., Schwille, J., Senk, S. L., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodriguez, M., & Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher education and development study in mathematics (TEDS-M)*. Amsterdam, Netherlands: IEA.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology, 51*(6), 309-317.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Hagedorn, M. C., Leggitt, J., & Najarian, M. (2015). *Early childhood longitudinal study, kindergarten class of 2010–11 (ECLS-K: 2011), user's manual for the ECLS-K: 2011 kindergarten-first grade data file and electronic codebook public version* (NCES 2015-078). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

White, P., & Breckenridge, R. S. (2014). Trade-offs, limitations, and promises of big data in social science research. *Review of Policy Research, 31*(4), 331-338.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *The Journal of Human Resources, 33*(1), 88-126.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18*, 327-350.

Wong, K. K., & Nicotera, A. C. (2004). Brown v. board of education and the Coleman report: Social science research and the debate on educational equality. *Peabody Journal of Education, 79*(2), 122-135.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach (5th Ed.)*. Mason, OH: South-Western Cengage Learning.