


Chapter 8

Use of Quasi-Experimental Research Designs in Education Research: Growth, Promise, and Challenges

MAITHREYI GOPALAN 

KELLY ROSINGER

JEE BIN AHN 

The Pennsylvania State University

In the past few decades, we have seen a rapid proliferation in the use of quasi-experimental research designs in education research. This trend, stemming in part from the “credibility revolution” in the social sciences, particularly economics, is notable along with the increasing use of randomized controlled trials in the strive toward rigorous causal inference. The overarching purpose of this chapter is to explore and document the growth, applicability, promise, and limitations of quasi-experimental research designs in education research. We first provide an overview of widely used quasi-experimental research methods in this growing literature, with particular emphasis on articles from the top ranked education research journals, including those published by the American Educational Research Association. Next, we demonstrate the applicability and promise of these methods in enhancing our understanding of the causal effects of education policies and interventions using key examples and case studies culled from the extant literature across the pre-K–16 education spectrum. Finally, we explore the limitations of these methods and conclude with thoughts on how education researchers can adapt these innovative, interdisciplinary techniques to further our understanding of some of the most enduring questions in educational policy and practice.

We need rigorous causal inference research to understand what works in the field of education. Following best practices from medical research, randomized controlled trials (RCTs) have become widely regarded as the “gold standard” for establishing causal evidence in education. While RCTs in education have increased dramatically over the years, some educational topics may not be amenable to RCTs either because they are too expensive, especially in larger place-based or policy-driven

Review of Research in Education

March 2020, Vol. 44, pp. 218–243

DOI: 10.3102/0091732X20903302

Chapter reuse guidelines: sagepub.com/journals-permissions

© 2020 AERA. <http://rre.aera.net>

interventions, or unethical in some cases. This has led to a simultaneous increase in the use of “as-if” random experiments or experiments naturally occurring in the world to establish causal evidence.

The increasing use of quasi-experimental research designs (QEDs) in education, brought into focus following the “credibility revolution” (Angrist & Pischke, 2010) in economics, which sought to use data to empirically test theoretical assertions, has indeed improved causal claims in education (Loeb et al., 2017). However, more recently, scholars, practitioners, and policymakers have questioned if the enthusiasm about RCTs and QEDs has narrowed the focus of research to less important topics. Have they crowded out high-quality descriptive analyses that attempt to make sense of often complex real-world topics that are not amenable to the simple exogenous shocks/variations needed for quasi-experiments (Deaton & Cartwright, 2018; Pritchett, 2018)? Furthermore, the external validity of causal claims made from QEDs and RCTs has not been a focus in the literature until more recently (Tipton & Olsen, 2018)—that is, “Are the results of analyses from specific studies generalizable to other populations not exposed to the intervention/policy?”

In this chapter, we provide an integrative review of the growth of QEDs in education research, their applicability and promise in improving causal inference, and ongoing challenges that exist in adapting these innovative methods in education research with an eye toward informing policy and practice.

WHAT ARE QUASI-EXPERIMENTAL RESEARCH DESIGNS?

To understand the causal effect of any policy or intervention, researchers strive to establish an appropriate counterfactual, or what would have happened in the absence of the policy or intervention, to provide a baseline from which causal effects can be estimated. RCTs are considered the “gold standard” of causal inference because of their ability to create a valid counterfactual by withholding treatment on a random set of subjects—known as the control group. In an educational RCT, subjects—students, classrooms, teachers, schools, or districts—are assigned to treatment and control groups based purely on chance. When treatment is randomly assigned, we can confidently claim that the treatment is the most plausible driver of the outcome. Because it is essential to rule out alternative explanations for an observed outcome to make a causal claim, random assignment ensures that treatment is not systematically related to other observable or unobservable factors. As a result, differences in outcomes can be attributed to the treatment rather than other factors systematically related to receipt of treatment.

However, it is often not feasible to conduct an RCT, especially in some educational settings. One of the most prohibitive barriers to conducting an RCT is the cost associated with many educational interventions. For example, changing classroom sizes requires substantial resources: reducing class size by just one student would cost an estimated \$12 billion a year nationwide (Chingos & Whitehurst, 2011). RCTs also raise ethical concerns because some students are denied educational interventions that

may be beneficial, while others receive interventions that have not yet been rigorously evaluated.

Quasi-experimental research designs, as the name suggests, use nonexperimental (or non-researcher-induced) variation in the main independent variable of interest, essentially mimicking experimental conditions in which some subjects are exposed to treatment and others are not on a random basis. Regression discontinuity, instrumental variables, differences-in-differences, two-way fixed effects, and other QEDs exploit nonrandom but plausibly exogenous (or as-if random) variation in key parameters to establish causality. The reliability of causal claims and estimates varies across these designs and depends on how close the study conditions are to an experiment. QEDs improve our understanding of the causal effects of various educational policies and interventions by focusing on internal validity—did the policy or intervention being studied cause a significant change in the observed outcome (and if so by how much)—thereby yielding an unbiased estimate of the average treatment effect (Campbell, 1957).

There are a number of threats to internal validity that QEDs try to eliminate in different ways, as is appropriate for each design. The main threat QEDs aim to eliminate is selection bias: the fact that students, districts, schools, or colleges that select into treatment may be different from those who do not select into treatment. Parents, students, or school administrators who are more informed, motivated, or in need of interventions that can improve outcomes may be more likely to opt into an educational program. The factors that lead them to select into treatment, however, are also likely connected to educational outcomes, making it difficult to isolate causal effects. QEDs attempt to exploit exogenous shocks that assign treatment to some and not to others on an as-if random basis—for example, policies that apply to one district and not the neighboring district, or thresholds that determine scholarship criteria based on a score or index that assigns treatment (i.e., scholarship) to students just above the score cutoff—that arguably can be used to establish causality.

QED studies that attempt to mitigate selection bias and reduce threats to internal validity receive the second-highest rating of causal evidence, behind well-conducted RCTs, according to the Institute of Education Sciences' (IES) What Works Clearinghouse (WWC) design standards (U.S. Department of Education, 2017).¹ The *hierarchy of evidence* in terms of the strengths and reliability of findings from studies using QEDs is still evolving and is an active line of research exploration.

TRENDS IN THE USE OF QUASI-EXPERIMENTAL RESEARCH DESIGNS IN EDUCATION RESEARCH

To examine the patterns and trends in the use of QEDs in education research, we followed a systematic search process to identify previous literature using such designs. To begin, we made the following list of the terms associated with QEDs: *quasi-experiment*, *natural experiment*, *difference-in-difference*, *regression discontinuity*, *instrumental variable*, *fixed effect*, *exogenous variation*, *two-way fixed effect*, *within sample*

*comparison, synthetic control method, propensity score matching, sibling comparison, and comparative interrupted time series.*² We used the ProQuest Education Database³ to search for articles that included any of the keywords in the title, abstract, or subject headings. To narrow our search, we restricted our search to peer-reviewed journals and used the time frame 1995 to 2018. The initial ProQuest Education Database search offered 2,704 search results from 100 scholarly journals.

Because the search results were from a wide range of social science disciplines beyond education, we narrowed the focus to articles published in 15 top education research journals, including all American Educational Research Association (AERA) journals that publish empirical studies. The list includes *AERA Open*, *American Educational Research Journal*, *American Journal of Education*, *Economics of Education Review*, *Educational Evaluation and Policy Analysis*, *Education Finance and Policy*, *Educational Researcher*, *Journal of Educational and Behavioral Statistics*, *Journal of Educational Psychology*, *Journal of Higher Education*, *Journal of Research on Educational Effectiveness*, *Research in Higher Education*, *Review of Higher Education*, *Sociology of Education*, and *Teachers College Record*. We included these journals in our search not only because they are top-tier journals but also because they represent a broad cross section of education research and are venues in which evaluations of policies that use QEDs are likely to be found. This search offered 632 results (see the supplemental appendix in the online version of the journal for extended bibliography of all these articles).

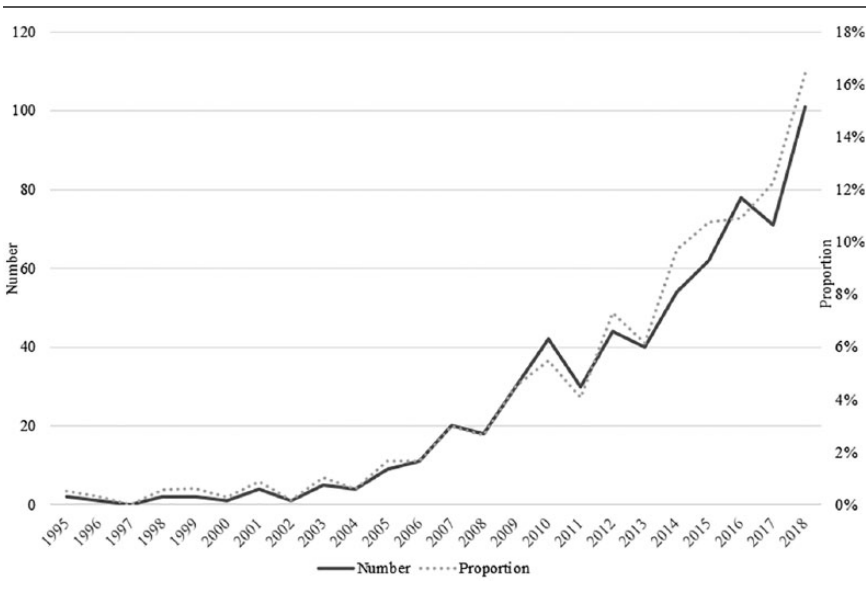
Figure 1 presents the trend in the use of QEDs (as operationalized by our search terms) in our selected set of journals between 1995 and 2018. The primary axis provides the total number of articles that use the relevant search terms described above and the secondary axis provides the proportion of all articles published in the above 15 journals that use those relevant search terms.

Our systematic search results indicate that while education research incorporated QEDs in the 1990s, they did not become popular until the past decade. The use of quasi-experimental terms has increased substantially over time. Until 2008, less than 20 articles using QED methods (as identified by the search terms) were published every year. Beginning 2009, we see an increasing trend with more than 100 articles using QEDs published in 2018. There is an upward trend in the total number of articles published by these journals over the same period as well. Despite this overall growth, we can see that the share of articles using QEDs has clearly gone up over this time period as well (plotted on the secondary axis).

There are several factors that likely contributed to the exponential growth of QEDs in education research. First, increasing accessibility of data, including micro-level data and large-scale longitudinal surveys collected by the federal and state governments, is key. Data accessibility led the growth of quasi-experimental approaches in economics (Angrist & Pischke, 2010; Panhans & Singleton, 2017) and likely also had similar influence on education research.

Second, increasing demand for rigorous policy evaluation has contributed to the increase in the use of QEDs in education research. Strong emphasis on an evidence-based approach to policy and interventions by the government alongside corresponding

FIGURE 1
Number and Proportion of Articles Using Quasi-Experimental Research Designs Between 1995 and 2018 in 15 Education Journals



demand from grant-making agencies have also led to the rapid growth of QEDs in education research. For example, IES’ grant application requirements include the need for rigorous evidence that meets the WWC standards for evidence. Also, the Every Student Succeeds Act in 2015 defined the term *evidence-based* and recommended that state educational agencies, local educational agencies, schools, educators, and partner organizations select and use evidence-based practices. The Every Student Succeeds Act distinguishes between evidence-based practices with strong, moderate, and promising evidence based on the strength of research design, and QEDs were determined to generate moderate evidence.

Finally, the increasing demand for rigor has also been accompanied by an expansion in methodological training, especially, in quantitative methods at schools of education. For example, a quick glance at the number of training grants provided to graduate schools of education by IES since its formation in 2002 illustrates the growth (see <https://ies.ed.gov/funding/grantsearch/index.asp?mode=2&sort=1&order=1&all=1&search=ProgramName&slctProgram=82>). Together, increased data accessibility, improved methodological training, and the growing need for transparent and credible policy evaluations with an emphasis on evidence-based policy in recent years seem to have triggered the growth of QEDs in education research.

To evaluate the relative popularity of various QED methods, we tabulated the number of articles that used each of the search terms that we used in our systematic search within the top education research journals (Table 1).

TABLE 1
Number of Articles Using QEDs by Method Between 1995 and 2018 in 15
Education Journals

Year	Quasi-Experiment	Natural Experiment	Difference in Difference	Regression Discontinuity	Instrumental Variable
1995–1999	2	0	0	0	3
2000–2004	3	1	1	0	4
2005–2009	13	8	8	9	26
2010–2014	34	11	30	41	47
2015–2018	49	21	59	77	39
Total	101	41	98	127	119

Year	Fixed Effect	Exogenous Variation	Two-Way Fixed Effects	Sibling Comparisons	Synthetic Control Methods
1995–1999	2	0	0	0	0
2000–2004	7	1	0	0	0
2005–2009	33	3	0	0	0
2010–2014	42	8	0	1	0
2015–2018	65	12	0	1	0
Total	149	24	0	2	0

Year	Propensity Score Matching	Within-Sample Comparison	Comparative Interrupted Time Series
1995–1999	0	0	0
2000–2004	0	0	0
2005–2009	10	0	0
2010–2014	30	0	3
2015–2018	35	0	8
Total	75	0	11

Note. The categorization of articles across these search terms is not mutually exclusive. For example, a study could include two key words such as quasi-experiment and difference in difference in the title, abstract, or article keywords. QEDs = quasi-experimental research designs.

Regression discontinuity (RD) and difference-in-difference (DID) methods are most commonly used in education research.⁴ The use of fixed effects is also common. Within the context of quasi-experiments, the use of two-way fixed effects is equivalent empirically to conducting DID with more than two time periods.⁵ The popularity of RD design in education research is not surprising. For example, in his review of the history of RD design, Cook (2008) describes how this design first originated in education by psychologists of education—Thistlethwaite and Campbell (1960) before being rediscovered and popularized in statistics and economics around 1995.

TABLE 2
Number of Articles Using QEDs by Topic Between 1995 and 2018 in 15 Education Journals

Year	Early Childhood Education	K–12 Education	Higher Education	Methodology	Other Topics
1995–1999	0	5	0	2	0
2000–2004	0	9	2	4	0
2005–2009	1	51	20	9	8
2010–2014	4	127	55	17	5
2015–2018	13	181	105	7	6
Total	18	372	182	41	19

Note. QEDs = quasi-experimental research designs.

Because several topics of inquiry within education are especially amenable to the use of RD design, it is not surprising that this method is widely used. Indeed, the WWC provides specific standards for evaluating studies using RD design in education research given its increasing use. Second, federalism, especially educational federalism in the United States, where states have tremendous power over education policies makes the use of DID possible and pertinent in education research. Because several educational policies are rolled out by (and across) states over time, this enables researchers to use pre-post designs with nonequivalent comparison groups to understand and evaluate the effects of such policies. Sibling comparison designs and family fixed effects designs seem less common in education research.

While a comprehensive description of the full range of QEDs is beyond the scope of this chapter, in the next section, we include a brief overview of two of the most popular methods—RD and DID (two-way fixed effect designs can be thought of as an extension of the DID). Other QEDs—such as matching techniques and instrumental variables, while also popular—are not described in greater depth in this review due to space constraints (see Stuart, 2010, for a detailed discussion of matching techniques, and Bettinger, 2010, for a detailed discussion of instrumental variables).

Next, we manually coded the articles in terms of the level of education that each article focused on broadly (Table 2).

We find that most of the articles focused on K–12 or higher education: 372 (approximately 59%) articles dealt with topics within K–12 education and 182 (approximately 29%) articles focused on higher education. Relatively few articles covered the issue of early childhood education (ECE); 18 articles, which only accounted for approximately 3% of the total. One reason for this trend in ECE could be that RCTs are still the dominant method used to understand the effects of ECE. Furthermore, several studies using QEDs in ECE have been published in other social science journals. We highlight a few QEDs used in ECE to describe the potential of QEDs to enhance causal evidence research in this domain.

Studies using QEDs in K–12 and higher education have focused on a wide range of topics. For example, within K–12 education, QEDs have been used to analyze the impact of school accountability policies such as the No Child Left Behind Policy Act (NCLB), school finance, class size reductions, school choice, and several other policy changes. Similarly, in higher education, QEDs have been used to assess the impact of financial aid, college selectivity, remedial education, and several other funding/accountability policies on student outcomes. Given the breadth and volume in the use of QEDs in K–12 and higher education, an immersive review of the use of QEDs in these two domains is beyond the scope of this chapter. Therefore, after providing an overview of two commonly used QEDs in the next section, we use specific case studies to highlight how research using QEDs has enhanced our understanding of the impact of certain educational interventions and policies in these domains.

OVERVIEW OF COMMONLY USED QUASI-EXPERIMENTAL RESEARCH DESIGNS IN EDUCATION RESEARCH

Regression Discontinuity Designs

Regression discontinuity designs seem to occupy the top slot when it comes to mimicking an experiment as closely as possible, and the WWC prioritizes evidence from well-conducted RD designs among QEDs. As a result, RD designs have received wide acceptance and use in education research. RD is appropriate for situations in which the eligibility to treatment is defined based on a cutoff on a continuous score or index. For example, a number of states and higher education institutions have enacted merit scholarship programs that provide financial aid to students above some academic eligibility threshold, such as Georgia's HOPE scholarship program that historically covered a portion of college costs for students with a 3.0 high school grade point average (GPA) or higher (see Dynarski, 2002, for more detailed discussion of the HOPE scholarship). Therefore, a student's probability of receiving aid, or "treatment," jumps discontinuously along the variable used in making treatment eligibility decisions. Because students who are just around the eligibility cutoff (e.g., students with a GPA of 3.02 vs. 2.98 when the eligibility cutoff is 3.0) are likely to be similar in observable and unobservable ways, this results in an as-if random experiment in which students just above the cutoff receive aid and students just below the cutoff do not.

The plausibly exogenous variation in who gets treated—those just above the cutoff—can be used to estimate the causal effect of the treatment with those just below the cutoff serving as the counterfactual. In such an RD design, the effect of an intervention is estimated as the difference in the average outcomes between the treatment and the control group members around a narrow threshold of the cutoff, after adjusting statistically for the relationship between the outcomes of interest and the variable used to decide eligibility and thereby the treatment. The estimated effect is known as the "local average treatment effect" and can be generalized only to students at the margin of the cutoff (e.g., students immediately above and below the cutoff value).

For an example of a study that uses an RD design to evaluate the impact of a merit scholarship program on student outcomes, see Curs and Harper (2012).

The variable used to assign subjects to treatment or control groups is commonly referred to as the “forcing,” “assignment,” or “running” variable. To estimate causal effects using RD, the cutoff value of the forcing variable must not be used for assigning other policies or interventions (e.g., remediation or satisfactory academic progress interventions based on GPA). For instance, an RD design will not establish causality if a 2.5 GPA cutoff value is used for both a scholarship program and for maintaining satisfactory academic progress; in this case, researchers will not be able to identify whether any changes in outcomes around the threshold are the result of the scholarship or interventions provided for students who fail to maintain satisfactory academic progress.

Difference-in-Differences Designs

The DID design is becoming increasingly popular in education policy research as well. Conceptually, in the DID design, researchers try to explore the effect of a policy or intervention that affects a group of students, schools, districts, or states at a point in time but not others at the same point in time in a naturally occurring setting. Researchers compare the outcomes of the groups differentially exposed to the treatment across pre- and posttreatment time periods to estimate the causal effects of policies. The posttreatment observations for units that never received treatment serve as the counterfactual, or what would have happened to treatment outcomes in the absence of treatment, and are used to estimate the average treatment effect. The main identifying assumption in a DID design is that the trends in outcomes for treated and control groups are parallel, or common, prior to treatment and that these trends would have remained parallel in the absence of treatment. The post-treatment trend in the control group (the one that does not experience treatment) is the counterfactual in a DID design and is what establishes causality. Disciplines use different terminology for the same general research design: where applied econometrics refers to this design as DID, psychology refers to it as comparative interrupted time series.

While the term *DID* in the past was used to specifically refer to analysis of two groups (treatment vs. control groups) and two time periods (pre- vs. posttreatment), it can be extended to multiple groups and multiple periods. Indeed, the use of group- and time-fixed effects, also known as the two-way fixed effects parameterization, is an extension of the DID that makes a similar common trend assumption to estimate causal effects of interest. This two-way fixed effect extension allows for more flexibility when treatment adoption occurs (i.e., states or districts that adopt policies in different years) and allows researchers to study variation in policy impact over time. Increasingly, event study analyses are used to supplement (or replace) traditional DID designs. In an event study analysis, a researcher adds leads and lags to the treatment variable to examine possible changes in outcomes in 1, 2, 3 (and possibly more) years prior to treatment and 1, 2, 3 (and possibly more) years after treatment. The leads on

the treatment variable allow researchers to examine whether there are anticipatory effects of policies (e.g., whether degree production changes prior to the implementation of differential pricing based on a student's college major), which could indicate that trends in outcomes were not parallel prior to treatment. Lags allow researchers to examine delayed responses to treatment, for instance, if pricing policies take several years to influence degree production in particular majors. See Stange (2015) for an application of this event study analysis.

CASE STUDIES OF QUASI-EXPERIMENTAL RESEARCH IN EDUCATION

We organize the case studies thematically, beginning with ECE research and ending with postsecondary research, to showcase the breadth in the substantive topics in which QEDs have been influential. Through these examples, we illustrate how such methodological advances have challenged and/or enhanced our knowledge regarding educational policy and practice. We draw on several classic studies (published in journals across the social sciences exploring education, broadly defined) that have used QEDs based on our background knowledge in addition to the studies included in our literature search ($N = 632$). We acknowledge that our selection and discussion of specific case study topics, while not exhaustive, nevertheless attempts to strike a balance between breadth and depth for an integrative review.

Early Childhood Education

Recent advances in developmental science aided by a deeper understanding of brain architecture have highlighted early childhood as a particularly sensitive period for promoting children's cognitive and socioemotional development (National Research Council & Institute of Medicine Committee on Integrating the Science of Early Childhood Development, 2000). Such advances combined with economic cost-benefit analyses suggesting high returns to early childhood investment and dynamic complementarities (the notion that "skills beget skills") of early childhood skills (Heckman, 2006) have resulted in a rapid growth of publicly supported ECE programs across the country. According to recent estimates, student enrollment in ECE programs has more than doubled between 2002 and 2018 (Friedman-Krauss et al., 2018). Several observational studies have documented a positive *correlation* between children's attendance in high-quality ECE programs and various outcomes—including academic achievement, behavioral skills, and health (Duncan & Magnuson, 2013; Magnuson et al., 2007).

Despite enthusiasm from researchers and policymakers, questions regarding the efficacy of ECE programs in improving short- and long-term outcomes for children remains. Meta- and re-analysis of studies that used rigorous natural experiments provide mixed evidence (Barnett et al., 2018; McCoy et al., 2017; Morris et al., 2018; van Huizen & Plantenga, 2018). Specifically, the reduction of positive gains (also known as fade-out) made by children who attended ECE in the elementary school

years and the reemergence of positive effects in adolescence and beyond has been widely debated. Because children attend ECE programs on a nonrandom basis—that is, parents choose ECE programs for their children—selection bias remains a concern when comparing ECE participants and nonparticipants. Several experimental methods and QEDs have been adopted to disentangle the causal evidence of ECE. We focus on studies that use QEDs to illustrate the applicability of these methods in this context.

For example, a common QED used to unpack the causal effect of ECE programs is the use of sibling- or family-fixed effects (Currie & Thomas, 1995; Deming, 2009). By comparing the outcomes of siblings within a family (such that one of the siblings was an ECE participant and the other was not), shared family-level characteristics can be controlled for as much as possible to isolate the true effect of ECE participation. Studies that use family-fixed effects find that participation in ECE programs improves college attendance and completion, improves health, and reduces criminal justice involvement (Carneiro & Ginja, 2014; Deming, 2009; Garces et al., 2002).

Second, given that the rollout of publicly funded ECE programs across different counties in the country was mostly random, researchers have exploited this “as-if” random variation in the timing of a child’s exposure to ECE to estimate the causal effect of ECE on longer run outcomes (Johnson & Jackson, 2019; Thompson, 2017). These studies use a DID design to compare differences in children’s outcomes of interest (e.g., achievement, behavior, and other long-term outcomes) between children who were exposed to ECE programs and those who were not, before and after ECE exposure. Similarly, as-if random variation in the timing of state subsidization of kindergarten has also been used in a DID framework to isolate the effect of kindergarten enrollment on student outcomes (Dhuey, 2011).

Finally, studies have also used RD designs to study the effect of ECE programs (Carneiro & Ginja, 2014; Gormley et al., 2005; Jenkins et al., 2016; Weiland & Yoshikawa, 2013). Because early ECE programs such as Head Start had eligibility requirements for participation (based on age, family income, household characteristics, state of residence, year, and others), researchers could compare the outcomes of children who were just above and below the threshold of eligibility using RD to estimate the average treatment effect. However, there have also been concerns raised regarding the use of an age-based RD design (Lipsey et al., 2015) in the ECE literature that later studies have begun to address.

In all, researchers find similar patterns as those observed in well-conducted RCTs. While smaller, local programs such as the Abecadarian and Perry preschool programs were high-quality programs that showed impressive positive results on a wide variety of student outcomes (that nevertheless faded out in elementary school years but reemerged in adolescence), publicly funded, larger ECE programs show smaller but significant effects on long-run student outcomes. Collectively, these studies have improved our understanding of the causal effects of ECE tremendously and encourage continued experimentation and analysis (see Phillips et al., 2017).

K–12 Education

In K–12 education policy, one of the most enduring questions regarding the link between school spending and students' educational outcomes is being increasingly addressed by studies using QEDs. Ever since the Coleman et al. (1966) report published findings suggesting the lack of significant relationship between school spending and student outcomes, there has been a long line of influential literature that was skeptical about the benefits of increased school spending on students' educational outcomes (Hanushek, 1997, 2003). However, most of these past studies suffered from methodological limitations arising from the use of observational data from convenience samples comprising data from few school districts or states and regression models that failed to account for selection bias. While studies observed heterogeneous effects (some positive, some negative, and some found no effect) of spending on key student outcomes even in this past literature (Hedges et al., 1994), none of the observed effects can be treated as causal estimates.

Because historically, local property taxes accounted for a large part of a K–12 school's spending (Howell & Miller, 1997), and the property tax base was typically higher in localities with higher home values, local financing of schools contributed to affluent districts spending more per student compared with poorer localities. Furthermore, given the persistent patterns of residential- and school-segregation by race and socioeconomic status, the link between contemporaneous school spending and student outcomes is likely to be biased in many ways that regression models even with extensive covariate adjustment are unlikely to correct.

Ideally, to reach the “gold standard” of causal inference, one would want to run an RCT where randomly selected school districts received money to spend while others did not. Comparing student outcomes in districts that received the money with student outcomes in similar districts that did not receive the random allocation of money after a set period of time would provide the most rigorous causal estimate of spending on student outcomes of interest. Yet such an experiment would remain a thought experiment at best given the ethical and logistical considerations it entails.

However, in the 1970s, lawsuits challenging the widespread disparity in within-state per-pupil spending across schools and districts resulted in court-ordered school finance reforms across the country. States implementing these court-ordered school finance reforms primarily tweaked their school spending formulas to mitigate inequality in spending. This weakened the positive correlation between per-pupil spending and district-level socioeconomic status/wealth, which was much higher in a school-funding regime that relied on property taxes in the local area. In other words, some school districts' spending changes in response to exogenous events, such as court-ordered school finance reforms, present a natural experiment—school districts that had low per-pupil spending prior to the reforms increased their spending in states that enacted the reforms in comparison with similar districts in states that did

not pass reforms. As a result, these exogenous spending changes could be isolated and used to estimate causal effects of spending using DID designs. This new line of literature comparing a range of short- and long-run student outcomes in those districts where spending increased versus those where spending stayed the same gives us some of the first causal estimates of school spending on student outcomes (Candelaria & Shores, 2019; Jackson et al., 2016; Lafortune et al., 2018).

These studies find that school spending improved student test scores (Lafortune et al., 2018), the overall number of years of completed education, wages, and high school graduation rates (Candelaria & Shores, 2019), and reduced the incidence of adult poverty (Jackson et al., 2016). Cumulatively, these studies also show that the positive effects of increases in per-pupil spending were driven by spending on teacher salaries, longer school days, and reduced student-teacher ratios (Jackson et al., 2016).

However, not all kinds of increases in school spending are related to improved educational outcomes. For example, studies using RD have found that increases in capital spending may not have such similar positive effects (Cellini et al., 2010; Martorell et al., 2016). Capital spending increases in schools and districts come from specific capital campaigns that are initiated by the local districts using referendums. Martorell et al. (2016) analyzed nearly 1,400 capital bond program referenda comparing districts where the referenda resulted in a narrow approval or failure. Because districts that barely passed or failed capital bonds passage are likely to be similar in most respects other than the “treatment,” selection bias into the treatment could be minimized. Studies using RD have also found that noninstructional spending on school counselors in Alabama causally reduced the frequency of disciplinary incidents (Reback, 2010). In all, there is clear, converging evidence that school spending matters. More important, “how” money is spent matters even more, especially for students who were exposed to unequal school resources (also see Jackson, 2018, for a more immersive review on this topic).

Higher Education

Since the mid-2000s, state legislators have increasingly implemented policies that link public funding for public colleges and universities to student outcomes in an effort to improve educational attainment rates by holding higher education institutions more accountable for outcomes. To date, at least 30 states have implemented some version of performance-based funding (PBF) policy, including most recently California, which passed legislation that will link \$2.5 billion in funding for the state’s community colleges to institutional performance (Fain, 2018). By linking a portion of state funding to student outcomes, PBF policies are intended to improve graduation rates—nationwide, just 60% of students who began college in 2009 had completed a degree 6 years later (National Student Clearinghouse, 2018)—and increase the number of degree and certificate holders as an effort to boost economic and workforce development within adopting states.

Descriptive reports that examine student outcomes before and after the implementation of PBF policies have highlighted generally positive impacts on degree completion in the 2- and 4-year college sectors, at least in states that tie a moderate amount of funds to performance (e.g., Callahan et al., 2017; Conklin et al., 2016). While these reports offer insight into trends and relationships between state funding policies and student outcomes, some lack a comparison group of states (ones that do not adopt PBF policies) against which to compare outcomes. With no counterfactual (i.e., no approximation of what would have happened if a PBF policy had not been adopted), the resulting estimates could be biased by broader demographic and labor market trends occurring at the same time.

As PBF policies have grown in popularity over the past 10 to 15 years, the staggered adoption of policies across states offers a natural experiment in which some public institutions are subject to PBF in a given year while others are not. Researchers have used this variation in PBF adoption over time to employ a DID framework, in which outcomes are observed before and after policy implementation in both adopting and nonadopting states, to evaluate their impact and generate causal estimates of PBF policies on a range of outcomes. Since PBF adoption may not be completely exogenous—that is, states with lower college completion rates may adopt PBF policies in an effort to boost degree production—many of these studies estimate DID models with multiple comparison groups, such as all other non-PBF states, states in a regional compact, neighboring states, or matching techniques to test whether findings are sensitive to which states are included in the nontreatment group (i.e., which states are used to construct the counterfactual).

Although Dougherty et al. (2016) note that PBF policies have encouraged institutions to offer additional supports to students, quasi-experimental analyses of PBF policies have largely failed to find clear positive impacts on actual degree production. DID analyses of individual state's PBF policies largely indicate that they have had null and, in some cases, even negative effects on degree production (Hillman et al., 2014; Hillman et al., 2015; Umbricht et al., 2017). When impacts on student outcomes have been positive in DID analyses, it has largely been among certificate production, findings that have been documented in Washington for short-term certificates (Hillman et al., 2015) and in Tennessee (Hillman et al., 2018). This boost in certificate production without concurrent increases in 2- and 4-year degrees raises concerns about the long-term impact of PBF policies since the returns for certificates are lower on average than those for associate's or bachelor's degrees (Belfield & Bailey, 2017). National DID analyses of multiple state PBF policies largely confirm null or negative findings from individual state studies for associate's (Li & Kennedy, 2018; Tandberg et al., 2014) and bachelor's (e.g., Tandberg & Hillman, 2014) degree production and an increase in short-term certificates (Li & Kennedy, 2018). One study found evidence that more recent PBF policies—those that tie performance funds to a college's base funding rather than providing performance funds as an additional bonus—have a modest positive effect on degree production (Rutherford & Rabovsky, 2014).

At the same time, the movement toward tying state funds to colleges' and universities' performance has also raised equity concerns. Primarily, if colleges respond to PBF by altering admissions, recruiting, or financial aid practices to expand enrollment among students with a high likelihood of completing a degree, they may limit access for low-income, adult, and underrepresented minority students, groups that have lower graduation rates on average. DID analyses indicate that the implementation of PBF policies increased selectivity at 4-year colleges in Indiana (Umbricht et al., 2017), which could threaten access for underrepresented students. A multistate study found that public institutions in PBF states receive less Pell Grant revenue after implementation, potentially indicating a preference among colleges that are subject to PBF for higher income students who are more likely to graduate (Kelchen & Stedrak, 2016).

Policy makers have responded to the concerns about limiting college access by including equity premiums or bonuses for colleges that enroll and/or graduate at-risk students. DID analyses demonstrate that such equity premiums can support access and success, at least among some underrepresented groups in higher education (e.g., Gándara & Rutherford, 2017; Kelchen, 2018). A more recent study found no changes in enrollment among historically underrepresented students at community colleges even in states with equity provisions (Kelchen, 2019). Understanding the impact of PBF policies on student outcomes is critical given the amount of funds at stake and the potential unintended consequences.

DID analyses have provided fairly strong and consistent evidence that PBF policies may not be meeting their intended goals, although new research indicates that equity metrics can help offset unintended consequences by supporting access and success for at-risk students. However, institutional, student, and broader state financial and economic factors appear to be more important in shaping educational outcomes than tying institutional funds to performance (Rutherford & Rabovsky, 2014).

In all, the above case studies of various points in the educational pipeline clearly document the power of QEDs in illuminating the causal effects of educational policies and interventions across the pre-K–16 spectrum. While in some cases, they uncover clear causal patterns, in others, they help clarify, extend, and/or add much needed nuance to findings uncovered by other observational research methods to enhance our understanding of policy impacts.

CHALLENGES AND LIMITATIONS IN QUASI-EXPERIMENTAL RESEARCH DESIGN APPLICATION

Methodological Limitations

The allure of using QEDs is clearly driven by the power and promise of these approaches in estimating internally valid, unbiased causal effects of a wide range of educational interventions and policies. Yet the validity of the results from studies based on these designs is intrinsically tied to the underlying assumptions and the

strength of those assumptions. For example, the parallel trends assumption inherent in DID designs are central to the validity of the estimated effects (see Wing et al., 2018, for a review of the robustness checks necessary to rigorously defend these assumptions). New research also indicates additional analyses, and sensitivity checks are needed for DID designs when treatment varies over time (as often happens when states and districts implement new policies; Goodman-Bacon, 2018). Additionally, other policy changes or interventions that occur at the same time as the treatment under study can confound the estimated effect in a DID design. Because researchers rely on observational data for DID (and other quasi-experimental analyses), there is always the possibility that other unobserved changes that occur simultaneously could drive changes in outcomes rather than the treatment of interest. In RD designs, the practice of including higher order polynomials of the running variable to capture nonlinearity has been a common practice. Recently, however, statisticians have shown how those terms might introduce other biases to the estimates (Gelman & Imbens, 2019).

While these shortcomings are not direct criticisms of QED methods per se, it is important to consider these carefully as we continue to adapt QEDs in educational policy and practice. Given the relatively short period of time over which QEDs have been developed and applied across the social sciences, methodologists and empirical researchers continue to expand our understanding of the underlying assumptions in these designs and the impact those assumptions have on estimating valid causal effects. Education researchers aspiring to use these methods must stay abreast of methodological advances if we hope to build a high-quality evidence base of “what works” and understand the intended and unintended consequences of educational policy and interventions.

Second, a methodological limitation often raised against QEDs (as well as RCTs) is that while the studies are able to estimate valid causal effects when identifying assumptions are met, they are unable to unpack the mechanisms underlying the overall treatment effects—that is, *how* the changes in outcomes occur. Good descriptive studies and in-depth qualitative studies have a clear role to play in this regard (Loeb et al., 2017). We discuss the need for good quality descriptive studies in the subsequent section.

External Validity and Generalizability

Despite the promise of QEDs in generating internally valid estimates of causal effects, one of the major concerns underlying these approaches is whether the estimated effects from the analytic samples using these designs are generalizable to other populations of interest—both across space and time. This generalizability concern is referred to as external validity (Campbell, 1957). For example, in evaluating the effectiveness of charter schools, several researchers have relied on “lottery studies.” Given that many charter schools experience oversubscription (more interest than available seats), some students are admitted to charter schools based on a random lottery. By comparing the learning outcomes of students who win the lottery with

those students who lose the lottery and thereby attend traditional public schools, such studies attempt to reduce selection bias given that both lottery winners and losers “selected into” charter schools. These studies have shown null effects overall in terms of learning outcomes in math and reading for charter school students, but significantly high positive effects for some subgroups of students attending high-quality charter schools (see Cohodes, 2018, for a review). However, the lottery approach suffers from a generalizability concern. It is possible to estimate a lottery effect only when a school is oversubscribed and has good administrative data on all lottery applicants available. Such schools may be quite different from an average charter school on several dimensions—such as size, location, student composition, or quality. Thus, findings based on lottery results may not generalize to the entire population of charter schools.

Similarly, studies using RD estimate what is known as the local average treatment effect based on an analytical sample that includes only a small number of observations on either side of the running variable cutoff used to determine treatment eligibility. Because the analytical sample right around the cutoff is not generalizable to the overall population of interest in many ways, the results from RD studies have limited external validity. Active methodological and applied empirical research to expand the external validity of RD designs is currently under way (Wing & Bello-Gomez, 2018). Yet external validity is an important limitation that we need to wrestle with as we continue to adapt these methods in education research.

At the same time, it is important to recognize that concerns regarding external validity are not limited to studies using QEDs alone. For example, similar critiques are leveled against the use of RCTs and experimental research more generally (Schanzenbach, 2012; Stuart et al., 2017). Alternatively, in some instances, studies using QEDs may have *higher* external validity as compared with those using RCTs because QED studies often test interventions or policies in real-world settings. Studies using QEDs can also be conducted in diverse settings where policy variation is observed, which provides opportunities to improve external validity. Therefore, the costs, benefits, and trade-offs between internal and external validity when using these approaches must be evaluated more rigorously to move research forward in this area. Furthermore, internal and external validity require a sharper focus with increased skepticism raised by the “replicability crisis”—instances where results from several classic studies in social psychology and other social sciences have not been replicated on newer, larger samples (Open Science Collaboration, 2015). Critiques such as publication bias (publication of studies that find significant effects as opposed to null effects), underpowered analytical samples, and a narrowed focus on null-hypothesis testing must be grappled with in research that uses QEDs as well.

Crowding Out Descriptive Studies

With the proliferation of the use of QEDs across the social sciences, scholars are also raising concerns regarding the trade-offs inherent in the overreliance on the use of these methods for research and practice more broadly (Deaton & Cartwright,

2018; Pritchett, 2018). First, a concern stemming from one side of this debate is if the search for clever identification strategies that rely on discontinuities and policy variations is limiting the kinds of research questions researchers ask and answer (Ruhm, 2018). Rather than letting the quest to answer policy-relevant research questions of interest using the most appropriate method drive the mode of inquiry, has the chase for exogenous shocks (i.e., search for “as-if” random/natural experiments) driven the questions that are being asked, answered, and published? In other words, similar to the axiom—form follows function—in research, should not the research methods and designs follow the research question and topic of inquiry rather than the other way around?

Education researchers must seriously engage in these discussions just as several social science disciplines are grappling with these issues. For example, a sharp focus on empirical methods to improve causal social science research may at times ignore the theoretical foundations on which research is built. Education research that uses QEDs must be encouraged to adopt practices to mitigate such trade-offs. To build cumulative knowledge, empirical research must help falsify, extend, and/or develop new theoretical predictions, especially in an applied social science field such as education.

Simultaneously, some experts have also raised concerns regarding whether the implicit incentive structures of academic publishing that favored RCTs and QEDs in recent years have crowded out the publication of good descriptive studies (McKenzie, 2018). High-quality descriptive studies are often required to describe key aspects of social phenomena. Recent guidance on how descriptive studies can be planned and executed with high empirical rigor hopefully encourages the publication of both stand-alone descriptive research studies and studies that use descriptive analyses in conjunction with RCTs and QEDs (Loeb et al., 2017).

Descriptive and qualitative studies help generate hypotheses and paint an important contextual picture that is integral for quantitative scholars who hope to identify natural experiments and apply QEDs to test and evaluate causal hypotheses. Therefore, we underscore the need to embrace methodological plurality, especially in education research where the “how” and “why” are just as important as “does the policy/intervention have a positive/negative effect?” and the magnitudes of the effects. Furthermore, advanced machine-learning methods and other data science approaches that favor predictive validity more than estimation of partial causal estimates must also be embraced by education researchers.

Limited Guidance on Policy Implementation and Design

One additional concern over the growing use of QEDs in education research is that they frequently offer limited guidance on policy implementation and design. QEDs lend themselves to binary measures—whether a policy or intervention existed in a given place at a given time. But policy is often much more nuanced than that. For instance, NCLB was implemented differently in each state, essentially resulting in 50 versions of NCLB (Wong et al., 2018), but many studies treat it as a single

homogeneous policy across states. Similarly, PBF policies for public colleges and universities vary from state to state, with some states tying as much as 85% of state appropriations for public colleges to institutional performance while others tie less than 5% to similar metrics. Analyses that treat policies as a homogeneous group when in fact substantial differences exist in implementation may fail to offer helpful guidance for policymakers (see Kelchen et al., 2019, for a more detailed discussion of these issues).

Relatedly, another criticism often raised against an overreliance on QEDs and, more generally, the causal inference research movement has been the limited guidance such research offers educational policy design and implementation (Polikoff & Conaway, 2018). Polikoff and Conaway (2018) argue that even when causal evidence exists, such research by itself may not provide sufficient guidance to policymakers that can inform specific policy design or implementation tweaks. Their recommendation is to increase the dissemination of expert-led, nontechnical, research syntheses that will invariably involve some subjective judgments on the cumulative evidence available to guide policymakers. A recent consensus statement authored by some of the most prominent ECE researchers is a classic example of such a research syntheses (Phillips et al., 2017). In that report, the authors evaluate the “state of scientific knowledge” available from RCTs, QEDs, and other research approaches when analyzing the effect of early childhood education on a variety of student outcomes.

OVERCOMING LIMITATIONS: A PATH FORWARD

It is increasingly clear from our trend analysis as well as the deeper case study reviews that QEDs have been embraced by the education research community. Yet concerns regarding how education research that incorporates such innovative techniques can be effectively used to inform future research, policy, and practice linger. A path forward necessarily involves a thoughtful integration of these approaches within the larger toolkit of education research. We offer some specific recommendations for such integration to flourish.

First, as our trend analysis reveals, despite substantial growth over time, only a small percentage of studies published in the top education research journals use QEDs. This leaves a lot of room for the increased use of these methods. Encouraging the use of these methods must, however, be accompanied with a call for an increase in the use of additional within-study robustness checks of results (Duncan et al., 2014). Using alternative specification checks, providing valid arguments defending the specific assumptions on which these designs’ ability to tease causal inference rests, and including tests that rule out plausible alternative explanations for observed causal effects become ever more important if these methods are increasingly adopted. Furthermore, quantitative scholars must be encouraged to use and refine theoretical frameworks as they continue to test specific hypotheses using QEDs.

Encouraging greater dialogue between qualitative and quantitative scholarship in education research has the potential to improve both types of research in education (Schudde, 2018). The notion of counterfactual thinking encouraged by the causal

inference movement could be adapted in comparative analysis and inform case selections by qualitative researchers (Plümper et al., 2019). On the other hand, in education, more nuanced counterfactual models that explore how control groups' behavior and performance evolves over time using qualitative and quantitative methods must be incorporated in evaluations (Lemons et al., 2014). In-depth qualitative research exploring the mechanisms and theories of change underlying the policy and interventions under study can enhance the quality of studies that use QEDs. Mixed-methods research can also help promote such integration. Studies using QEDs can therefore effectively complement the education research enterprise in several ways with an eye toward improving evidence-based practice and policy.

Finally, most of the challenges we identified above are not limited to just QEDs. Rather, these are challenges that the education sciences face in an effort to ensure that research does not get increasingly disconnected from actual policy design and implementation. It is therefore important to promote the publication of research syntheses using empirical meta-analytical techniques as well as integrative systematic reviews that can evaluate the "state of knowledge" that combines research evidence from a larger body of evidence that includes studies that use QEDs as well as other research methodologies. Developing and enhancing systematic frameworks for assessing the quality of evidence in education that includes these new techniques must be a priority for education research.

CONCLUSION

The rapid growth in the use of QEDs across the social sciences and more specifically in education research is undeniable. In this chapter, we synthesized literature that uses QEDs across the pre-K–16 education spectrum to examine how the use of these methodologies has improved our knowledge and understanding of educational policy issues. In doing so, it seems rather clear that studies using QEDs have significantly improved our understanding of causal relationships in education. Specifically, these methods have been integral in highlighting the strength and magnitude of causal effects of key educational policies and interventions on well-defined student outcomes and, in some cases, have clearly demonstrated unintended consequences of certain education policies and interventions.

However, as is common in any burgeoning literature, there is tension between an overreliance in the use of these methods and the trade-offs that their use entails. Program evaluation literature has long emphasized the trade-off between internal and external validity in social science research (Campbell, 1957). Studies using QEDs (and RCTs) have privileged internal over external validity in education research as well. While scholars, practitioners, and funding agencies grapple with this trade-off, methodologists are increasingly working to improve these methods to provide insights for generalizability (Tipton, 2014; Tipton & Olsen, 2018). Furthermore, the growth of machine learning and data science in education research is another trend that scholars need to engage with when adapting these techniques to inform policy and practice. It is, however, clear that these innovative approaches can be embraced

enthusiastically (but also thoughtfully) to extend our knowledge of educational policies and interventions that strive to go beyond correlational approximations.


Indeed, leaders in the field have noted how far researchers have come in improving rigor in education research (Hedges, 2018). But the increasing shift toward QEDs has some challenges: “Purist devotion to experimental and strong quasi-experimental designs made good sense when we were fighting an uphill battle for increased rigor” (Singer, 2018, p. 23). However, now that QEDs are well-established in education research, it seems ever more important to adopt methodological plurality. Simultaneously, at this time of “replication crisis” in the social sciences, we need to be increasingly cautious and humble about the limits of “evidence and how certain that evidence might be” (Hedges, 2018, p. 18).


Ongoing methodological innovations and the increasing availability of large-scale education data make the use of QEDs and other rigorous observational methods ever more possible. This chapter aimed to illustrate the promise as well as challenges and limitations inherent in adapting QEDs to inform education research, policy, and practice.

ACKNOWLEDGMENTS

We thank John Cheslock, two anonymous reviewers, and the editors for helpful feedback on earlier drafts of this chapter.

ORCID iDs

Maithreyi Gopalan  <https://orcid.org/0000-0002-1013-0672>

Jee Bin Ahn  <https://orcid.org/0000-0001-9087-4892>

NOTES

¹Only well-conducted RCTs with low levels of sample attrition receive the highest rating for evidence by the WWC.

²These search terms include a broadly accepted set of terms to identify QEDs used in social science. While some of these terms have gained more popularity over time, and are more common in some social science fields such as economics, we believe that these search terms are fairly expansive and capture most of the studies published in the top journals in education that use QEDs.

³Information on three of the target journals: *AERA Open*, *Economics of Education Review*, and *Education Finance and Policy*, which were not included in ProQuest Education Database was collected from SAGE Open Access Journals, ScienceDirect Journals, and ERIC, respectively. Also, note that *AERA Open* is a relatively new journal that began publishing in 2015.

⁴Instrumental variable approaches are also used in education research but mostly by economists. More than half of the 119 articles using these approaches were published in a single journal—*Economics of Education Review*.

⁵Studies identified by the search include those that use fixed effects just as a covariate in their analysis. While this method reduces omitted variable bias to some extent, these studies cannot be considered quasi-experimental in a strict sense. Yet we included “fixed effects” in our search term to be as expansive as possible to ensure our review erred on the side of inclusion rather than exclusion.

REFERENCES

- Angrist, J., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2), 3–30. <https://doi.org/10.3386/w15794>
- Barnett, W. S., Jung, K., Friedman-Krauss, A., Frede, E. C., Nores, M., Hustedt, J. T., Howes, C., & Daniel-Echols, M. (2018). State prekindergarten effects on early learning at kindergarten entry: An analysis of eight state programs. *AERA Open*, 4(2), Article 2332858418766291. <https://doi.org/10.1177/2332858418766291>
- Belfield, C., & Bailey, T. (2017). *The labor market returns to sub-baccalaureate college: A review* (Center for Analysis of Postsecondary Education and Employment Working Paper). <https://ccrc.tc.columbia.edu/media/k2/attachments/labor-market-returns-sub-baccalaureate-college-review.pdf>
- Bettinger, E. (2010). Instrumental variables. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International encyclopedia of education* (3rd ed., Vol. 8). Elsevier.
- Callahan, M. K., Meehan, K., & Shaw, K. M. (2017). *Impact of OBF on student outcomes: Tennessee and Indiana*. Research for Action.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297–312. <http://dx.doi.org/10.1037/h0040950>
- Candelaria, C. A., & Shores, K. A. (2019). Court-ordered finance reforms in the adequacy era: Heterogeneous causal effects and sensitivity. *Education Finance and Policy*, 14(1), 31–60. https://doi.org/10.1162/EDFP_a_00236
- Carneiro, P., & Ginja, R. (2014). Long-term impacts of compensatory preschool on health and behavior: Evidence from Head Start. *American Economic Journal: Economic Policy*, 6(4), 135–173. <https://doi.org/10.1257/pol.6.4.135>
- Cellini, S. R., Ferreira, F., & Rothstein, J. (2010). The value of school facility investments: Evidence from a dynamic regression discontinuity design. *Quarterly Journal of Economics*, 125(1), 215–261. <https://doi.org/10.1162/qjec.2010.125.1.215>
- Chingos, M., & Whitehurst, G. (2011). *Class size: What research says and what it means for state policy*. Brookings Institution. <https://www.brookings.edu/research/class-size-what-research-says-and-what-it-means-for-state-policy/>
- Cohodes, S. (2018, February 1). *Charter schools and the achievement gap*. <https://futureofchildren.princeton.edu/news/charter-schools-and-achievement-gap>
- Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F., & York, R. (1966). *Equality of educational opportunity*. U.S. Department of Health, Education, and Welfare, Office of Education. <https://eric.ed.gov/?id=Ed012275>
- Conklin, K. D., Snyder, M., Stanley, J., & Boelscher, S. (2016). *Rowing together: Aligning state and federal investments in talent to common outcomes*. https://www.ecs.org/wp-content/uploads/ECS_FundingReports_HCM_F.pdf
- Cook, T. D. (2008). “Waiting for life to arrive”: A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2), 636–654. <https://doi.org/10.1016/j.jeconom.2007.05.002>
- Currie, J., & Thomas, D. (1995). Does Head Start make a difference? *American Economic Review*, 85(3), 341–364. <https://www.jstor.org/stable/2118178>
- Curs, B. R., & Harper, C. E. (2012). Financial aid and first-year collegiate GPA: A regression discontinuity approach. *The Review of Higher Education*, 35(4), 627–649. <https://doi.org/10.1353/rhe.2012.0040>
- Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21. <https://doi.org/10.1016/j.socscimed.2017.12.005>
- Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, 1(3), 111–134. <https://doi.org/10.1257/app.1.3.111>

- Dhuey, E. (2011). Who benefits from kindergarten? Evidence from the introduction of state subsidization. *Educational Evaluation and Policy Analysis, 33*(1), 3–22. <https://doi.org/10.3102/0162373711398125>
- Dougherty, K. J., Jones, S. M., Lahr, H., Pheatt, L., Natow, R. S., & Reddy, V. (2016). *Performance funding for higher education*. Johns Hopkins University Press.
- Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*(11), 2417–2425. <https://doi.org/10.1037/a0037996>
- Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives, 27*(2), 109–132. <https://doi.org/10.1257/jep.27.2.109>
- Dynarski, S. (2002). The behavioral and distributional implications of aid for college. *American Economic Review, 92*(2), 279–285. <https://doi.org/10.1257/000282802320189401>
- Fain, P. (2018). *As California goes?* <https://www.insidehighered.com/news/2018/06/12/calif-finalizes-performance-funding-formula-its-community-colleges>
- Friedman-Krauss, A. H., Barnett, W. S., Weisenfeld, G. G., Kasmin, R., DiCrecchio, N., & Horowitz, M. (2018). *The state of preschool 2017*. National Institute for Early Education Research.
- Gándara, D., & Rutherford, A. (2017). Mitigating unintended impacts? The effects of premiums for underserved populations in performance-funding policies for higher education. *Research in Higher Education, 59*(6), 681–703. <https://doi.org/10.1007/s11162-017-9483-x>
- Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review, 92*(4), 999–1012. <https://doi.org/10.1257/00028280260344560>
- Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics, 37*(3), 447–456. <https://doi.org/10.1080/07350015.2017.1366909>
- Goodman-Bacon, A. (2018). *Difference-in-differences with variation in treatment timing* (Working Paper No. 25018). <https://doi.org/10.3386/w25018>
- Gormley, W. T. Jr., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41*(6), 872–884. <https://doi.org/10.1037/0012-1649.41.6.872>
- Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141–164. <https://doi.org/10.3102/01623737019002141>
- Hanushek, E. A. (2003). The failure of input-based schooling policies. *Economic Journal, 113*(485), F64–F98. <https://doi.org/10.1111/1468-0297.00099>
- Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science, 312*(5782), 1900–1902. <https://doi.org/10.1126/science.1128898>
- Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness, 11*(1), 1–21. <https://doi.org/10.1080/19345747.2017.1375583>
- Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). An exchange: Part I. Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher, 23*(3), 5–14. <https://doi.org/10.3102/0013189X023003005>
- Hillman, N. W., Hicklin Fryar, A., & Crespín-Trujillo, V. (2018). Evaluating the impact of performance funding in Ohio and Tennessee. *American Educational Research Journal, 55*(1), 144–170.
- Hillman, N. W., Tandberg, D. A., & Fryar, A. H. (2015). Evaluating the impacts of “new” performance funding in higher education. *Educational Evaluation and Policy Analysis, 37*(4), 501–519. <https://doi.org/10.3102/0162373714560224>
- Hillman, N. W., Tandberg, D. A., & Gross, J. P. (2014). Performance funding in higher education: Do financial incentives impact college completions? *Journal of Higher Education, 85*(6), 826–857. <https://doi.org/10.1353/jhe.2014.0031>

- Howell, P. L., & Miller, B. B. (1997). Sources of funding for schools. *Future of Children*, 7(3), 39–50. <https://doi.org/10.2307/1602444>
- Jackson, C. K. (2018). *Does school spending matter? The new literature on an old question*. Paper presented at the Fall 2018 Bronfenbrenner Center for Translational Research Conference. https://works.bepress.com/c_kirabo_jackson/38/
- Jackson, C. K., Johnson, R. C., & Persico, C. (2016). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *Quarterly Journal of Economics*, 131(1), 157–218. <https://doi.org/10.1093/qje/qjv036>
- Jenkins, J. M., Farkas, G., Duncan, G. J., Burchinal, M., & Vandell, D. L. (2016). Head Start at Ages 3 and 4 versus Head Start followed by state pre-K: Which is more effective? *Educational Evaluation and Policy Analysis*, 38(1), 88–112. <https://doi.org/10.3102/0162373715587965>
- Johnson, R., & Jackson, C. K. (2019). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. *American Economic Journal: Economic Policy*, 11(4), 310–349. <https://doi.org/10.3386/w23489>
- Kelchen, R. (2018). Do performance-based funding policies affect underrepresented student enrollment? *Journal of Higher Education*, 89(5), 702–727. <https://doi.org/10.1080/00221546.2018.1434282>
- Kelchen, R. (2019). Exploring the relationship between performance-based funding design and underrepresented student enrollment at community colleges. *Community College Review*, 47(4), 382–405. <https://doi.org/10.1177/0091552119865611>
- Kelchen, R., Rosinger, K. O., & Ortagus, J. C. (2019). How to create and use state-level policy data sets in education research. *AERA Open*, 5(3), Article 2332858419873619. <https://doi.org/10.1177/2332858419873619>
- Kelchen, R., & Stedrak, L. J. (2016). Does performance-based funding affect colleges' financial priorities? *Journal of Education Finance*, 41(3), 302–321. <https://doi.org/10.1353/jef.2016.0006>
- Lafortune, J., Rothstein, J., & Schanzenbach, D. W. (2018). School finance reform and the distribution of student achievement. *American Economic Journal: Applied Economics*, 10(2), 1–26. <https://doi.org/10.1257/app.20160567>
- Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, 43(5), 242–252. <https://doi.org/10.3102/0013189X14539189>
- Li, A. Y., & Kennedy, A. I. (2018). Performance funding policy effects on community college outcomes: Are short-term certificates on the rise? *Community College Review*, 46(1), 3–39.
- Lipsey, M. W., Weiland, C., Yoshikawa, H., Wilson, S. J., & Hofer, K. G. (2015). The prekindergarten age-cutoff regression-discontinuity design: Methodological issues and implications for application. *Educational Evaluation and Policy Analysis*, 37(3), 296–313. <https://doi.org/10.3102/0162373714547266>
- Loeb, S., Dynarski, S., McFarland, D., Morris, P., Reardon, S., & Reber, S. (2017). *Descriptive analysis in education: A guide for researchers* (No. NCEE 2017-4023; p. 53). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.
- Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review*, 26(1), 33–51. <https://doi.org/10.1016/j.econedurev.2005.09.008>
- Martorell, P., Stange, K., & McFarlin, I. (2016). Investing in schools: Capital spending, facility conditions, and student achievement. *Journal of Public Economics*, 140, 13–29. <https://doi.org/10.1016/j.jpubeco.2016.05.002>

- McCoy, D. C., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson, K., Yang, R., Koepp, A., & Shonkoff, J. P. (2017). Impacts of early childhood education on medium- and long-term educational outcomes. *Educational Researcher*, 46(8), 474–487. <https://doi.org/10.3102/0013189X17737739>
- McKenzie, D. (2018, September 4). Have descriptive development papers been crowded out by impact evaluations? *World Bank Blogs*. <https://blogs.worldbank.org/impactevaluations/have-descriptive-development-papers-been-crowded-out-impact-evaluations>
- Morris, P. A., Connors, M., Friedman-Krauss, A., McCoy, D. C., Weiland, C., Feller, A., Page, L., Bloom, H., & Yoshikawa, H. (2018). New findings on impact variation from the Head Start impact study: Informing the scale-up of early childhood programs. *AERA Open*, 4(2), Article 2332858418769287. <https://doi.org/10.1177/2332858418769287>
- National Research Council & Institute of Medicine Committee on Integrating the Science of Early Childhood Development. (2000). *From neurons to neighborhoods: The science of early childhood development* (J. P. Shonkoff & D. A. Phillips, Eds.). National Academies Press. <http://www.ncbi.nlm.nih.gov/books/NBK225557/>
- National Student Clearinghouse. (2018). *Completing college*. <https://nscresearchcenter.org/signaturereport16/>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716–aac4716. <https://doi.org/10.1126/science.aac4716>
- Panhans, M. T., & Singleton, J. D. (2017). The empirical economist's toolkit from models to methods. *History of Political Economy*, 49(Supplement), 127–157. <https://doi.org/10.1215/00182702-4166299>
- Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., Duncan, G. J., Dynarski, M., Magnuson, K. A., & Weiland, C. (2017). *Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects: A consensus statement* (p. 16). Brookings Institution.
- Plümper, T., Troeger, V. E., & Neumayer, E. (2019). Case selection and causal inferences in qualitative comparative research. *PLOS ONE*, 14(7). <https://doi.org/10.1371/journal.pone.0219727>
- Polikoff, M., & Conaway, C. (2018). Getting beyond 'did it work?': Proposing a new approach to integrate research and policy [The Brookings Institution]. *Brookings Brown Center Chalkboard*. <https://www.brookings.edu/blog/brown-center-chalkboard/2018/09/25/getting-beyond-did-it-work-proposing-a-new-approach-to-integrate-research-and-policy/>
- Pritchett, L. (2018). *Lant Pritchett Talk: "The debate about RCTs in development is over. We won. They lost."* Retrieved September 11, 2018, from <http://www.nyudri.org/events-index/2018/2/22/lant-pritchett-talk-the-debate-about-rcts-in-development-is-over-we-won-they-lost>
- Reback, R. (2010). Noninstructional spending improves noncognitive outcomes: Discontinuity evidence from a unique elementary school counselor financing system. *Education Finance and Policy*, 5(2), 105–137. <https://doi.org/10.1162/edfp.2010.5.2.5201>
- Ruhm, C. J. (2018). *Shackling the identification police?* (Working Paper No. 25320). <https://doi.org/10.3386/w25320>
- Rutherford, A., & Rabovsky, T. (2014). Evaluating impacts of performance funding policies on student outcomes in higher education. *Annals of the American Academy of Political and Social Science*, 655(1), 185–208. <https://doi.org/10.1177/0002716214541048>
- Schanzenbach, D. W. (2012). Limitations of experiments in education research. *Education Finance and Policy*, 7(2), 219–232. https://doi.org/10.1162/EDFP_a_00063
- Schudde, L. (2018). Heterogeneous effects in education: The promise and challenge of incorporating intersectionality into quantitative methodological approaches. *Review of Research in Education*, 42(1), 72–92. <https://doi.org/10.3102/0091732X18759040>

- Singer, J. D. (2018). Even more challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, 11(1), 22–24. <https://doi.org/10.1080/19345747.2017.1402397>
- Stange, K. (2015). Differential pricing in undergraduate education: Effects on degree production by field. *Journal of Policy Analysis and Management*, 34(1), 107–135. <https://doi.org/10.1002/pam.21803>
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1–21. <https://doi.org/10.1214/09-STS313>
- Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of school districts that participate in rigorous national educational evaluations. *Journal of Research on Educational Effectiveness*, 10(1), 168–206. <https://doi.org/10.1080/19345747.2016.1205160>
- Tandberg, D. A., & Hillman, N. W. (2014). State higher education performance funding: Data, outcomes, and policy implications. *Journal of Education Finance*, 39(3), 222–243.
- Tandberg, D. A., Hillman, N., & Barakat, M. (2014). State higher education performance funding for community colleges: Diverse effects and policy implications. *Teachers College Record*, 116(12), 1–31.
- 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. <https://doi.org/10.1037/h0044319>
- Thompson, O. (2017). Head Start's long-run impact: Evidence from the program's introduction. *Journal of Human Resources*, 0216-7735r1. <https://doi.org/10.3368/jhr.53.4.0216.7735R1>
- Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, 39(6), 478–501. <https://doi.org/10.3102/1076998614558486>
- Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, 47(8), 516–524. <https://doi.org/10.3102/0013189X18781522>
- Umbrecht, M. R., Fernandez, F., & Ortagus, J. C. (2017). An examination of the (un)intended consequences of performance funding in higher education. *Educational Policy*, 31(5), 643–673. <https://doi.org/10.1177/0895904815614398>
- U.S. Department of Education. (2017). *Standards handbook* (Version 4.0). What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf
- van Huizen, T., & Plantenga, J. (2018). Do children benefit from universal early childhood education and care? A meta-analysis of evidence from natural experiments. *Economics of Education Review*, 66, 206–222. <https://doi.org/10.1016/j.econedurev.2018.08.001>
- Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, 84(6), 2112–2130. <https://doi.org/10.1111/cdev.12099>
- Wing, C., & Bello-Gomez, R. A. (2018). Regression discontinuity and beyond: Options for studying external validity in an internally valid design. *American Journal of Evaluation*, 39(1), 91–108. <https://doi.org/10.1177/1098214017736155>
- Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health*, 39(1), 453–469. <https://doi.org/10.1146/annurev-publhealth-040617-013507>
- Wong, V. C., Wing, C., Martin, D., & Krishnamachari, A. (2018). Did states use implementation discretion to reduce the stringency of NCLB? Evidence from a database of state regulations. *Educational Researcher*, 47(1), 9–33. <https://doi.org/10.3102/0013189X17743230>