1

The Challenge for Educational Research

Throughout the world, education is viewed as a mechanism for expanding economic opportunity, enhancing social mobility, developing a skilled workforce, and preparing young people to participate in civic life. Thus, it is no surprise that almost every government wants to improve the quality of its country's educational system. Public resources are scarce, however, and education must compete with demands for improved health care, housing, and nutrition. When resources devoted to educational activities do not improve student achievement, it is difficult for educational policymakers to lay claim to additional resources. For this reason, policymakers need to use available resources wisely and be able to demonstrate that they have done so. To accomplish these objectives, governments need good information about the impacts that particular policy decisions are likely to have on student achievement. Unfortunately, in the past, this kind of information has not been available.

The Long Quest

The call for better empirical evidence upon which to base sound educational policy decisions has a long history, one that is particularly well documented in the United States. In a speech given to the National Education Association (NEA) in 1913, Paul Hanus—a Harvard professor, and later the first dean of the Harvard Graduate School of Education—argued that "the only way to combat successfully mistaken common-sense as applied to educational affairs is to meet it with uncommon-sense in the same field—with technical information the validity of which is indisputable"

(Hanus, 1920, p. 12). For Hanus, this meant that systematic research must be conducted and its findings applied. In his words, "We are no longer disputing whether education has a scientific basis; we are trying to find that basis." In his lengthy speech to the NEA, Hanus identified a number of school policy decisions that he believed should be based on scientific evidence. These included finding an "adequate and appropriate means of determining the qualifications of well-trained and otherwise satisfactory workers for the educational staff . . .," and formulating "courses of study . . . together with suggestions as to methods of teaching." Although educational policymakers today would use somewhat different terms in framing such questions, these same substantive concerns remain pressing in countries around the world: How do we attract and retain skilled teachers? What are the most important skills for students to acquire? What are the most effective pedagogies for teaching these skills?

For educational researchers in Hanus's time, and for many years thereafter, "carrying out scientific research" meant implementing the ideas of scientific management that had been developed by Frederick W. Taylor and laid out in his 1911 book *Principles of Scientific Management*. Taylor's central thesis was that experts could uncover the single "best" way to do a particular job by conducting "time and motion" studies. Then, the task of management was to provide the appropriate tools, and create training, incentives, and monitoring systems to ensure that workers adopted and followed the prescribed methods.

Although Taylor was careful not to apply his methods to any process as complicated as education, many educational researchers were less cautious. One of the most influential proponents of applying Taylor's system of scientific management to education was Frank Spaulding, who earned a doctorate from the University of Leipzig, Germany, in 1894, served as superintendent of several U.S. school districts during the first two decades of the twentieth century, and, in 1920, became head of Yale University's newly formed Department of Education. Speaking at the same meeting of the NEA at which Hanus gave his address, Spaulding described three essentials for applying scientific management to education. He stipulated that we must: (a) measure results, (b) compare the conditions and methods under which results are secured, and (c) adopt consistently the conditions and methods that produce the best results (Callahan, 1962, pp. 65–68).

Most educators today would agree with these essentials, even though many would object to the "Taylorist" ideas that underlie them. However, it was in applying these essentials to education that controversy arose. The "results" that Spaulding used in his research included "the percentage of children of each year of age in the school district that the school

enrolls; the average number of day's attendance secured annually from each child; [and] the average length of time required for each child to do a given definite unit of work" (Callahan, 1962, p. 69). The attraction of these indicators for Spaulding was that they could be measured relatively easily and precisely. However, they were not very good indicators of the quality of education that schools provided to children. As a result, despite its popularity among educators anxious to build support in the business community, many thoughtful educators found Spaulding's research to be of dubious value for improving the quality of the education actually provided to children.

In the decades following Spaulding's speech, the creation of standardized multiple-choice tests and the development of item-response theory made it possible increasingly to measure students' skills and knowledge in academic domains like reading and mathematics at relatively low cost. Advances in information technology that included the creation and development of digital computers and user-friendly data-processing software made it possible to manipulate large amounts of quantitative information. Advances in statistical methods, including the development of multiple-regression techniques, made it possible to better identify and summarize important patterns in data and to test specific hypotheses.

All of these advances contributed to a major milestone in educational research in the United States, the publication of a study entitled *Equality of Educational Opportunity*, in 1966. The study that led to this report was commissioned by the U.S. Congress as part of the Civil Rights Act of 1964. This legislation ordered the U.S. Commissioner of Education to conduct a study of "the lack of availability of equal educational opportunities for individuals by reason of race, color, religion, or natural origin in public educational institutions at all levels" (Coleman et al., 1966, p. iii). The order of the wording in this quotation—race, color, religion, national origin—suggests that the Congress had little doubt that children who were disadvantaged minorities received fewer and lower-quality educational resources than did white children, and that differences in educational resources were the probable cause of differences in academic achievement.

The task of organizing and conducting the congressionally mandated study fell to the eminent sociologist James Coleman. Coleman and his team went well beyond their charge and conducted a quantitative study that sought to account for variation in academic achievement among American children, incorporating information on both their schooling and their family backgrounds. Coleman's research design borrowed heavily from research that had been conducted previously in agriculture to estimate the impact of different resource combinations on output levels. Coleman applied this so-called production function methodology to

investigate what combinations of educational inputs "produced" particular levels of educational output.

Despite being released on the Friday before the July 4 Independence Day holiday in 1966, the Coleman Report (as it came to be known) did not go unnoticed. As anticipated, it documented that black children had much lower academic achievement, on average, than white children. The surprise was that differences in school resources, such as class size and the educational credentials of their teachers, accounted for almost none of this achievement gap. U.S. Commissioner of Education Harold Howe summarized the main finding of the Report as "family background is more important than schools."[1] Since this interpretation seemed to undercut the initiatives of President Lyndon Johnson's Great Society, which were aimed at reducing race-based economic inequality by improving the schools that served children of color, some policymakers called for an additional, harder look at the data.

This led two prominent Harvard professors, eminent statistician Frederick Mosteller and then political-science professor and later U.S. Senator Daniel Patrick Moynihan, to organize a working group to reanalyze Coleman's data. Their resulting volume, *On Equality of Educational Opportunity*, contained a compendium of papers that revealed a great deal about the academic achievement of American children and the schools they attended (Mosteller & Moynihan, 1972). However, the researchers concluded that the intrinsic limitations of the cross-sectional observational data collected by the investigators in the Equality of Educational Opportunity Survey made it impossible to answer critical questions about whether school resources had *causal* impacts on children's achievement. Consequently, the working group called for better research designs and the collection of representative prospective longitudinal data in order to understand more comprehensively the impacts on children of investments in schooling.

Economist Eric Hanushek was one of the first social scientists to respond to this call. A member of the Moynihan and Mosteller seminar group, Hanushek was well aware of the limitations of the Coleman Report and of the need for the collection of longitudinal data on the achievement of individual children. He collected such data for several hundred children who were attending elementary schools in one California school district. He also collected data describing important attributes of the teacher of each student's class and of the classroom setting, such as the number of students present in the class. Using the same multiple-regression

1. As reported in Herbers (1966).

methods that Coleman had employed, Hanushek sought to answer two questions. The first was whether children in some third-grade classrooms ended the school year with higher achievement than children in other third-grade classrooms, on average, after taking into account the achievement levels with which they started the school year. Hanushek (1971) confirmed that this was indeed the case, and that the differences were large enough to be educationally meaningful. This finding was important in the wake of the Coleman Report because it verified what parents and educators already knew, but that the limited nature of the Coleman Report data could not verify—namely, that school quality did indeed matter.

The second question that Hanushek addressed was whether the budgeted resources that school districts used to purchase actually accounted for why children in some classrooms had higher average achievement at the end of the school year than did children in other classrooms. He focused his attention particularly on the roles of teacher experience and teacher qualifications, because these teacher characteristics are rewarded in almost all public-school salary scales in the United States and other countries. Hanushek found that neither the number of years that a teacher had taught nor whether the teacher had earned a master's degree accounted for much of the classroom-to-classroom variation in children's achievement. Neither did class size, the other large source of difference in cost among classrooms. Hanushek's conclusion was that schools were spending money on things such as teaching experience, higher degrees for teachers, and smaller class sizes that did not result in improved student achievement.

Although researchers applauded Hanushek for confirming that school quality did indeed matter, many questioned his conclusions about the inefficiency of schools (Hedges et al., 1994). They pointed out, for example, that in many schools, children with the greatest learning needs were actually being assigned to the smallest classes. Consequently, the low academic achievement of students in small classes may not have meant that class size did not matter. On the contrary, if schools were attempting to equalize student outcomes by providing additional resources to children with the greatest learning needs, then Hanushek's research strategy could not provide unbiased answers to causal questions such as whether there was an academic payoff to reducing class sizes or to paying teachers with greater professional experience more than novices. Progress in answering such questions would await the development of the research designs and analytic methods that are described in this book.

Another impetus to educational research in the 1960s was the increasing involvement of the federal government in American K-12 education.

The passage of the Elementary and Secondary Education Act (ESEA), in 1965, marked the first time that the federal government had provided significant funding for public K-12 education in the United States. Title I of the Act committed federal funds to improving the schooling of economically disadvantaged children. Fearful that the money would not make a difference to poor children, Senator Robert Kennedy insisted that the ESEA require periodic evaluations of whether the program was producing requisite gains in student achievement (McLaughlin, 1975, p. 3). In essence, Senator Kennedy wanted evidence of a causal impact.

Partly in response to the demands for the implementation of more systematic educational research expressed by Daniel Moynihan, who was by now head of President Richard Nixon's Domestic Council, President Nixon announced, in 1970, that the federal government would create the National Institute of Education (NIE). Touted as the vehicle for fostering systematic scholarship that would solve the nation's educational problems, NIE began operation in 1972, with an annual budget of $110 million. Secretary of Health, Education, and Welfare (HEW) Elliott Richardson informed Congress that the administration would request a $400 million budget for NIE within five years (Sproull, Wolf, & Weiner, 1978, p. 65).

By the end of the 1970s, optimism about the ability of empirical research to resolve the important arguments in education that had marked the inauguration of the NIE had turned to pessimism. Soon after the inauguration of Ronald Reagan in 1980, the NIE was dissolved. In part, the demise of the NIE stemmed from a change in the mood of the country. The rapid economic growth of the mid-1960s, which had increased federal tax revenues and fueled the Great Society programs, ended in 1973, and was followed by a decade of very slow growth. Optimism had initially accompanied the sending of U.S. troops to Vietnam in the early 1960s. By the early 1970s, more than 50,000 American deaths and the accompanying failed foreign-policy objectives had changed the country's mood. As Henry Aaron described in his book, *Politics and the Professors* (1978), many citizens stopped believing that government itself could be instrumental in improving the lives of Americans and came to believe that government was a principal cause of the economic malaise in which the country found itself.

The demise of the NIE was not completely the result of the change in the country's mood, however. Another part of the problem had been the unrealistic expectations that had accompanied its birth. Many of the original proponents of the creation of the NIE had invoked comparisons with research conducted by the National Institutes of Health, which had provided radical new medicines such as the Salk vaccine for polio, and the agricultural research that had resulted in the green revolution.

When the NIE's research programs did not produce analogous visible successes for education, it was deemed a failure. Few of its advocates had appreciated how difficult it would be to answer questions posed by policymakers and parents about the effective use of educational resources.

Yet another part of the explanation for the demise of the NIE, and the low funding levels of its successor, the U.S. Department of Education's Office of Educational Research and Improvement (OERI), was the widespread perception that educational research was of relatively low quality. A common indictment was that educational researchers did not take advantage of new methodological advances in the social sciences, particularly in the application of innovative strategies for making causal inferences.

In an attempt to respond to the concern about the low quality of educational research, the U.S. Congress established the Institute of Education Sciences (IES) in 2002, with a mandate to pursue rigorous "scientific research" in education. One indication of the energy with which the IES has pursued this mandate is that, in its first six years of operation, it funded more than 100 randomized field trials of the effectiveness of educational interventions.[2] As we explain in Chapter 4, the randomized experiment is the "gold-standard" design for research that aims to make unbiased causal inferences.

The Quest Is Worldwide

Although the quest for causal evidence about the consequences of particular educational policies is particularly well documented in the United States, researchers in many countries have conducted important studies that have both broken new ground methodologically and raised new substantive questions. We illustrate with two examples. Ernesto Schiefelbein and Joseph Farrell (1982) conducted a remarkable longitudinal study during the 1970s of the transition of Chilean adolescents through school and into early adulthood. The authors collected data periodically on a cohort of students as they moved from grade 8 (the end of primary school) through their subsequent schooling (which, of course, differed among individuals) and into the labor market or into the university. This study, *Eight Years of Their Lives*, was a remarkable tour de force for its time. It demonstrated that it was possible, even in a developing country that was experiencing extraordinary political turmoil, to collect data on the same

2. See Whitehurst (2008a; 2008b). We would like to thank Russ Whitehurst for explaining to us which IES-funded research projects were designed as randomized field trials.

individuals over an extended period of time, and that these data could provide insights not possible from analyses of cross-sectional data. Substantively, the study documented the important role that the formal education system in Chile played in sorting students on the basis of their socioeconomic status. This evidence provided the basis for considerable debate in Chile about the design of publicly funded education in the years after democracy returned to the country in 1989 (McEwan, Urquiola, & Vegas, 2008).

The book *Fifteen Thousand Hours*, by Michael Rutter (1979), describes another pioneering longitudinal study. The research team followed students in 12 secondary schools in inner-city London over a three-year period from 1971 through 1974, and documented that students attending some secondary schools achieved better outcomes, on average, than those attending other schools. One methodological contribution of the study was that it measured several different types of student outcomes, including delinquency, performance on curriculum-based examinations, and employment one year after leaving school. A second was the attention paid to collecting information on variables other than resource levels. In particular, the researchers documented that characteristics of schools as social organizations—including the use of rewards and penalties, the ways teachers taught particular material, and expectations that faculty had of students for active participation—were associated with differences in average student outcomes.

A close reading of the studies by Schiefelbein and Farrell, and by Rutter and his colleagues, shows that both sets of researchers were aware acutely of the difficulty of making causal inferences, even with the rich, longitudinal data they had collected. For example, Schiefelbein and Farrell wrote: "It is important to reemphasize that this study has not been designed as a hypothesis-testing exercise. Our approach has consistently been exploratory and heuristic. And necessarily so" (p. 35). In their concluding chapter, Rutter and his colleagues wrote: "The total pattern of findings indicates the strong probability that the associations between school processes and outcome reflect in part a causal process" (p. 179). Why were these talented researchers, working with such rich data, not able to make definitive causal statements about the answers to critical questions of educational policy? What does it take to make defensible causal inferences? We address these questions in the chapters that follow.

What This Book Is About

In recent decades, tremendous advances have been made in data availability, empirical research design, and statistical methods for making

causal inferences. This has created new opportunities for investigators to conduct research that addresses policymakers' concerns about the consequences of actions aimed at improving educational outcomes for students. But how can these new methods and data be applied most effectively in educational and social-science research? What kinds of research designs are most appropriate? What kinds of data are needed? What statistical methods are best used to process these data, and how can their results be interpreted so that policymakers are best informed? These are the questions that we address in this book.

The particular designs and methods that we have chosen to describe are sophisticated and innovative, often relatively new, and most have their origins in disciplines other than education. We have sought to present them in a way that is sensitive to the practical realities of the educational context, hoping not only to make you receptive to their incorporation into educational research, but also to persuade you to incorporate them into your own work.

An innovative aspect of our book is that we illustrate all of our technical discussions of new research design and innovative statistical methods with examples from recent, exemplary research studies that address questions that educational policymakers around the world have asked. We explain how these studies were designed and conducted and, where appropriate, we use data from them to illustrate the application of new methods. We also use these same studies to illustrate the challenges of interpreting findings even from exemplary studies and to demonstrate why care in interpretation is critical to informing the policy process.

The studies that we highlight examine a variety of causal questions, examples of which include:

- Does financial aid affect students' and families' educational decisions?
- Does providing students with subsidized access to private schools result in improved educational outcomes?
- Do early childhood programs have long-term benefits?
- Does class size influence students' achievement?
- Are some instructional programs more effective than others?

All of these questions have been the subject of high-quality studies that have implemented cutting-edge designs and applied innovative methods of data analysis. We refer to these high-quality studies throughout our book as we explain a variety of innovative approaches for making causal inferences from empirical data. In fact, by the end of our book, you will find that the phrase "high-quality" itself eventually becomes code for referring to studies that effectively employ the approaches we describe.

Notice that all of the educational policy questions listed here concern the impact of a particular action on one or more outcomes. For example, does the provision of financial aid affect families' decisions to send a child to secondary school? This is a distinctive characteristic of causal questions, and learning to answer such questions is the topic of this book. In our work, we distinguish such causal questions from descriptive questions, such as whether the gap between the average reading achievement of black students and that of white students closed during the 1980s. Although there are often significant challenges to answering descriptive questions well, these challenges are typically less difficult than the challenges you will face when addressing causal questions.

We have written this book not only for those who would like to conduct causal research in education and the social sciences, but also for those who want to interpret the results of such causal research appropriately and understand how the results can inform policy decisions. In presenting these new designs and methods, we assume that you have a solid background in quantitative methods, that you are familiar with the notion of statistical inference, and that you are comfortable with statistical techniques up to, and including, ordinary least-squares (OLS) regression analysis. However, as an interested reader can see by skimming ahead in the text, ours is not a highly technical book. To the contrary, our emphasis is not on mathematics, but on providing *intuitive explanations* of key ideas and procedures. We believe that illustrating our technical explanations with data from exemplary research studies makes the book widely accessible.

We anticipate that you will obtain several immediate benefits from reading our book carefully. First, you will learn how alternative research designs for making causal inferences function, and you will come to understand the strengths and limitations of each innovative approach. In addition, you will learn how to interpret the results of studies that use these research designs and analytic methods, and will come to understand that careful interpretation of their findings, although often not obvious, is critical to making the research useful in the policy process.

What to Read Next

We conclude every chapter with a brief list of additional resources you may want to consult, to learn more about the topics that were discussed in the chapter. In this introductory chapter, the extra readings that we suggest deal primarily with the history of educational research. In subsequent chapters, many of our suggestions are to scholarly papers that

provide specialized treatments of the particular technical issues raised in the chapters they accompany.

To learn more about the reasons why the NIE failed to fulfill its much-publicized promise, we suggest reading the 1978 book by Lee Sproull and her colleagues entitled *Organizing an Anarchy*. Jonah Rockoff's (2009) paper "Field Experiments in Class Size from the Early Twentieth Century" provides an interesting and brief history of attempts to estimate the causal impact of class size on student achievement. Grover "Russ" Whitehurst's thoughtful reports (Whitehurst, 2008a, 2008b) on the research agenda that the Institute of Education Sciences developed and supported during the period 2003–2008 describe the challenges of supporting research that is both rigorous and relevant to improving the quality of the education that children receive.

2

The Importance of Theory

A question that governments around the world ask repeatedly is whether using scarce public resources to educate children is a good social investment. Beginning in the late 1950s, and sparked by the pioneering work of Nobel Prize winners Theodore Schultz and Gary Becker, economists developed a theoretical framework within which to address this question. The resulting framework, which became known as *human capital theory*, provided the foundation for a vast amount of quantitative research in the ensuing years. Among the many insights from human capital theory and the empirical work that it generated are the important role education plays in fostering a nation's economic growth, the reason education has its biggest labor market payoffs in economies that are experiencing rapid technological change, and why employers are often willing to pay to train workers to do specific tasks, such as use a new technology, but are typically unwilling to pay for training that improves workers' reasoning skills and writing ability.[1]

Over subsequent decades, social scientists refined the theory of human capital in a variety of ways. These refinements led to new hypotheses and to important new evidence about the payoffs to investments in education, some of which are described in later chapters. The salient point for the

1. For many references to the evidence of the role of education in fostering economic growth, see Hanushek and Woessman (2008). For evidence on the especially valuable role of education in increasing productivity in environments experiencing technological change, see Jamison & Lau (1982). The classic reference on the reasons why employers are typically willing to pay for specific training, but not general training, is Becker (1964).

moment is that human capital theory provides a powerful illustration of the role that theory plays in guiding research, especially research into cause and effect. We will return to human capital theory later in this chapter. First, however, we explain what we mean by the term *theory* and the roles that it plays in guiding research in the social sciences and education.

What Is Theory?

According to the *Oxford English Dictionary* (OED, 1989), a theory is "a scheme or system of ideas or statements held as an explanation or account of a group of facts or phenomena; a hypothesis that has been confirmed or established by observation or experiment, and is propounded or accepted as accounting for the known facts; a statement of what are held to be the general laws, principles, or causes of something known or observed."[2] All three parts of this definition contain the notion that, within a theory, a general principle of some kind—the OED calls it a "scheme," a "system," "general laws,"—is intended to "explain" or "account for" particular instances of what we observe on a day-to-day basis.

Theory plays important roles in guiding empirical research in the social sciences and education by providing guidance about the questions to ask, the key constructs to measure, and the hypothesized relationships among these constructs. For example, at the core of human capital theory is the idea that individuals compare benefits and costs in making decisions about whether to undertake additional education. This framework leads researchers to ask what factors should be included among the benefits and costs of acquiring additional education, and how to measure differences among individuals in these benefits and costs or changes in their values over time. Theory also suggests the direction of hypothesized relationships. For example, theory suggests that a decline in the earnings of college graduates relative to those of high school graduates would lead to a decline in the percentage of high school graduates who decide to enroll in college.

Of course, theory is never static. For example, in the first round of an investigation, research questions are often broad and undifferentiated, and any hypothesized intervention is treated simply as a "black box." However, in answering the first-round research question, investigators

2. Accessed at the following webpage: http://dictionary.oed.com.ezp-prod1.hul.harvard.edu/cgi/entry/50250688?query_type=word&queryword=theory&first=1&max_to_show=10&sort_type=alpha&result_place=1&search_id=wpON-NKMyN3-6203-&hilite=50250688

can refine their theory and propose more sophisticated, finer-grained questions that sometimes shed light on the causal machinery within the box.

The development of human capital theory illustrates this. One pattern common across many countries that the theory sought to explain was that, on average, the more formal education that workers had completed, the higher were their labor market earnings. In its initial formulation in the late 1950s, economists treated "formal schooling" as a black box—increases in formal schooling were theorized to improve subsequent wages because they led to increases in the productivity of workers. In a 1966 paper, Richard Nelson and Edmund Phelps unpacked this straightforward idea, and thereby refined human capital theory, by suggesting that additional education increased productivity because it increased workers' ability to understand and make use of new information. This led them to hypothesize that education would have a greater impact on worker productivity in settings in which technologies were changing than in settings in which technology was static.

Subsequent quantitative research tested this new hypothesis and found support for it. For example, Jamison and Lau (1982) found that education had a larger impact on productivity in agriculture in settings in which green revolution seeds and fertilizers were changing agricultural methods than it did in settings in which techniques were stable and had been passed down orally from one generation to the next. Later contributions to human capital theory developed the idea that if additional education did improve individuals' skills at processing and making use of new information, it would not only increase their productivity at work, it would also result in improved health and better parenting.[3] These subsequent theoretical refinements catalyzed a still growing body of quantitative research on the payoffs of education.

Good theory often leads researchers to new ideas that raise questions about the tenets of existing theory. For example, building on the work of Kenneth Arrow and others, Michael Spence (1974) developed a challenge to human capital theory. Spence proposed an alternative theory, which he called *market signaling*. In a simple market-signaling model, high-productivity individuals obtain additional schooling not because it enhances their skills, but because it is a way to signal to potential employers that they possess exceptional qualities, and consequently should be paid higher salaries than other applicants. Thus, a market-signaling model

3. In Chapter 10, we describe one important paper in this line of research, written by Janet Currie and Enrico Moretti (2003).

could explain the positive relationship between educational attainments and labor market wages even if education did not enhance students' skills. Market signaling continues to pose an alternative to human capital theory in explaining education–earnings relationships in some settings. Unfortunately, it has proven very difficult to design quantitative research to test the two theories unequivocally, head to head. In fact, many social scientists would argue that both human capital theory and market signaling theory play roles in explaining wage patterns in many societies. For example, the earnings premium that graduates of elite universities enjoy stems in part from the skills they acquired at the university and partly from the signal of high ability that admission to, and graduation from, an elite university conveys.[4]

The French sociologist Pierre Bourdieu posed another alternative to human capital theory to explain the role that education plays in Western societies. In Bourdieu's theoretical framework, education sorts students in ways that contribute to the reproduction of existing social hierarchies. Children in elite families graduate from the best universities and obtain access to prestigious, well-paying careers. Children from lower-class families obtain education that only provides access to low-prestige, lower-paying jobs. In Bourdieu's theory, many sorting mechanisms contribute to this pattern. One is the allocation of educational opportunities by scores on standardized tests that favor the types of knowledge that children from well-to-do families acquire at home. Another is an educational-finance system that favors students from families that already have financial resources. Most objective social scientists recognize that Bourdieu's theory of *social reproduction* sheds light on the role that education plays in many societies.[5] However, as with market signaling, it has proven difficult to compare Bourdieu's theory with human capital theory head to head. In fact, the two theories provide complementary insights into explaining the role that education plays in many settings.

Philosophers of science distinguish between two modes of inquiry, one based on *deductive logic* and the other based on *inductive logic.* Deductive reasoning involves the development of specific hypotheses from general theoretical principles. In the exercise of deductive reasoning, you engage in the reverse of deduction. You begin by observing an unexpected pattern, and you try to explain what you have observed by generalizing it. In other words, with inductive reasoning, you go from particular observations to general principles. The origins of most important theories involve

4. For an accessible discussion of human capital and market signaling models, see Weiss (1995).
5. For an introduction to Bourdieu's theory, see Lane (2000).

a mixture of the two types of reasoning. Induction is critical because the theorist is trying to make sense of a pattern he or she has observed or learned about. At the same time, having a rudimentary theory in mind directed the theorist's attention to the pattern. Once theories are formulated, deduction typically becomes preeminent in the formal design and execution of new theory-based research. However, induction often provides the post-hoc insight that is instrumental in refining existing theory.

Both deductive and inductive reasoning have played roles in the development of human capital theory. For example, economists used deductive reasoning to formulate a variety of specific hypotheses based on the general statement of human capital theory. One was that the lower the interest rate high school graduates had to pay on loans for college tuition, the more probable it was that they went to college. Economists also used insights from human capital theory to inform the design of research aimed at estimating the rate of return to a society of investing in education. A consistent result was that the social benefits from universal primary education far exceeded the social costs in most developing countries. Indeed, in most countries, the estimated social rate of return on investments in primary education far exceeded the social rate of return to other possible governmental use of resources, such as investing in physical infrastructure (Psacharopoulos, 2006).

Despite the compelling evidence that primary education was a good social investment in most countries, social scientists observed that many families in developing countries choose not to send their children to school. This observation led researchers to engage in inductive reasoning, in order to formulate possible explanations. One alternative was that families did not have access to primary schools—a supply problem. A second was that families were unaware of the payoffs to education—an information problem. A third was that families could not borrow the money at reasonable interest rates to pay the cost of schooling, costs that might include replacing the labor that the child provided at home—a problem of capital market failure. Yet another hypothesis was that paying for children's schooling was not a good personal investment for parents in cultures in which children did not feel a strong moral obligation to support their parents later in life—a cultural explanation. These hypotheses, all stemming from the observation that the educational investment decisions of many families seemed inconsistent with insights from human capital theory, led to increased attention in human capital theory to the supply of schools, the information available to parents, their ability to borrow at reasonable interest rates, and cultural norms about children's responsibilities to parents. In turn, these hypotheses led to studies that

examined the relative importance of these different possible explanations for the educational investment decisions of parents.

Theory in Education

Every educational system involves a large and diverse array of actors whose decisions interact in a great many ways. Governments make decisions about the types of organizations that may offer schooling services and that are eligible to receive partial or full payment for their services from tax revenues. Parents make decisions about the schools their children will attend and the amount of time and energy they will devote to shaping their children's skills and values at home. Children make decisions about how much attention they will pay to school-related work and to the types of interactions in which they will engage with peers. Educators decide where they will work, how they will teach, and how much attention they will pay to individual children. The decisions of these many actors interact in important ways. For example, parents' choices are influenced by their children's efforts in school and by the quality and resources of their local schools. The actions of policymakers regarding licensing requirements and compensation structures influence the career decisions of potential teachers.[6]

The number of different players who contribute to education, and the complexity of their interactions, make it difficult to formulate parsimonious, compelling theories about the consequences of particular educational policies. In contrast, physics is a field with very strong theory—well-developed general principles expressed in mathematical terms from which stem many clearly defined hypotheses that can be tested empirically. In thinking about the role of theory in the social sciences and education, it is important to remember that physics is the exception rather than the rule. In most other fields of scientific endeavor, theory is commonly expressed in words rather than in mathematics, and the general principles are less clearly defined than they are in physics. The reason we mention this is to encourage researchers to define theory broadly, so that it includes a clear description of the policy intervention to be evaluated, the outcomes it may influence, and conjectures about the mechanisms through which the intervention may influence outcomes. (Some writers use the term *theory of action* to refer to these steps.) Rarely will such a description be expressed

6. The ideas we describe in this paragraph are taken from Shavelson and Towne (eds., 2002).

in mathematical terms, and it does not need to be. What is important is clear thinking, which is typically informed by a deep knowledge of previous research in the relevant area and a solid grounding in a social science.

One vital part of the work of using theory to inform the design of empirical work investigating causal relationships in education and the social sciences concerns the measurement of key concepts. For example, a hypothesis of great interest in many countries is that reducing class sizes in elementary schools will result in improved student achievement. A little thinking brings the realization that the key conceptual variables relevant to this hypothesis—class size and student achievement—could be measured in many different ways, and the choices could affect the results of the research. For example, in schools in which the student population is mobile, counting the number of students who appear on a class roster would provide a very different measure of class size than would counting the number of students in attendance on any single day. Developing a measure of student achievement raises even more questions. Do scores on standardized reading tests measure literacy skills effectively? Would the research results differ if other measures of student achievement were chosen, such as success at the next level of schooling? We see the process of thinking hard about these measurement issues as part of the task of applying theory to the design of empirical work.

It is often useful to distinguish between two kinds of theories that can inform the design of causal research in education and the social sciences. *Partial equilibrium theories* can shed light on the likely consequences of policy interventions of modest scale undertaken in a particular setting. An example would be the application of human capital theory to predict the consequences of a policy initiative that would offer zero-interest loans for college expenses to high school graduates from low-income families in a particular community. Since only a modest number of students would be affected by the proposed policy, it would be reasonable to assume that the loan program would have no impact on the relative earnings of high school graduates and college graduates.

In contrast, in considering the consequences of a policy initiative that would offer zero-interest college loans to all low-income students in the United States, it would be important to take into consideration that an increase in the supply of college graduates would lower the earnings of this group relative to the earnings of high school graduates. Theories that take into account such indirect effects of policy initiatives are called *general equilibrium theories.*

An important question when choosing a particular theoretical framework to guide the design of causal research in education is whether a partial equilibrium approach will suffice, or whether a general equilibrium

approach is necessary. The advantage of a partial equilibrium framework is usually its relative simplicity. However, the simplicity is achieved by the assumption that the intervention is of sufficiently small scale that secondary effects can be ignored. The advantage of a general equilibrium framework is that it provides tools to examine those secondary effects, which are likely to be more important the larger the scale of the policy initiative. One accompanying cost, however, is greater complexity. An even greater cost of adopting a general equilibrium framework as the basis for a social science experiment is that if an intervention has broadly distributed secondary effects, it is very difficult for the investigator to define an appropriate comparison or control group that would not be influenced indirectly by the intervention. For that reason, we share the view of many methodologists that random-assignment experiments and the other analytic techniques that we promote in this book cannot capture the full general equilibrium effects of large-scale policy interventions (Duflo, Glennerster, & Kremer, 2008).

In the next section, we provide an example of how a prominent theory regarding the consequences of *educational vouchers* became more refined over time. We also show the ways that researchers used both partial equilibrium and general equilibrium versions of the theory to shed light on the consequences of particular educational voucher policies.

Voucher Theory

Perhaps the most vigorously contested educational policy issue in the world in recent years has concerned the consequences of using public tax revenues to pay for children's education at private schools. Writing in the early 1960s (Friedman, 1962), the American economist—and later Nobel Prize winner—Milton Friedman argued that the prevailing system of public schools in the United States restricted the freedom of parents to choose the schools that would best serve their children. He advocated the introduction of an "educational voucher" system, which would, in his view, both expand freedom of choice and improve the quality of American education. Friedman's initial statement of voucher theory was elegant. The key policy recommendation was that government should provide educational vouchers of equal value to all parents of school-age children. Parents could then use the vouchers to pay for the education of their children at a public school of their choice or use them to pay part or all of the tuitions at a private school of their choice.

Friedman envisioned several desirable outcomes from a universal voucher system, some easier to measure (increased student achievement

and lower schooling costs) than others (enhancement of freedom). The mechanism through which a voucher system would achieve these outcomes was the force of market competition. Part of Friedman's argument was that the introduction of a system of educational vouchers would have its greatest positive impact on the quality of education available to children from low-income families because they have the fewest schooling choices under the prevailing educational system.

Implicit in Friedman's voucher theory were two critical assumptions, both of which came from the application to education of the economic theory of *competitive markets*. The first assumption was that consumers would be free to choose any school for which they could pay the tuition. The second was that the schooling choices that parents made for their children would be independent of the schooling choices that other parents made for their children. These assumptions made sense in competitive markets for consumer goods such as bread. Typically, shoppers can buy any brand of bread that they feel is worth the market price, and their decisions are not directly influenced by the choices made by other consumers. These assumptions simplify enormously the development of theories that predict how competitive markets function.

In the decades following publication of Friedman's voucher theory, however, a growing number of studies documented that the two critical assumptions implicit in Friedman's voucher theory did not hold. One reason is that some children are more expensive to educate than others. For example, children with disabilities, such as dyslexia or hearing problems, require additional resources to help them to master critical skills (Duncombe & Yinger, 1999). If schools are constrained to charge the same tuition to all students, and if the value of the education voucher provided by government is the same for all children, then school administrators have incentives to avoid accepting children who would be expensive to educate.

A second challenge to the assumptions underlying Friedman's voucher theory is that parents recognize that the quality of the education their child receives in a particular school depends on the skills and behaviors of other children attending the same school (Graham, 2008; Hoxby, 2000). These influences, which sociologists call "peer-group effects" and economists call "externalities," complicate the way that educational voucher systems would operate in practice. In particular, schools that attempted to attract students from particular types of families (such as those with well-educated, affluent parents) would seek to refuse admission to children whom their desired clientele would not like to have as classmates.

Taking advantage of advances in computer-based simulation, a number of social scientists developed theoretical models that incorporated cost

differentials and peer-group effects. Many of the models also incorporated details of public school finance systems. These are complex general-equilibrium models that treat the school-choice decisions of families as interdependent. Researchers used these theoretical models not only to explore how the introduction of voucher plans with particular designs would affect families' schooling choices, but also how they would influence things like housing prices and families' decisions about where to live.[7] A hypothesis stemming from many of these theoretical models and policy simulations is that a universal educational-voucher system in which the value of the voucher was the same for all children would lead to significant sorting of students from specific backgrounds into particular schools. Subsequent studies of universal voucher systems in Chile (Hsieh & Urquiola, 2006) and in New Zealand (Fiske & Ladd, 2000), in which the vouchers did have the same value for all children, provided evidence supporting this hypothesis. For example, Hsieh and Urquiola (2006) showed that children from the poorest families in Chile tended to be concentrated in low-performing public schools, whereas children from relatively affluent families were concentrated in particular private schools.[8]

Evidence of the importance of cost differentials and peer-group effects has resulted in two subsequent refinements to voucher theory. The first has been the creation of theoretical models predicting the consequences of voucher systems in which the value of the voucher that individual children receive depends on their characteristics.[9] The logic is that a system with appropriately differentiated voucher values might prevent the sorting by socioeconomic status that took place under the single-valued voucher systems in Chile and New Zealand. The second development has been the formulation (and testing) of relatively simple partial-equilibrium models in which only children from low-income families are eligible to receive vouchers. The logic underlying these models is that the family-income limits for participation would reduce the threat of sorting by

7. See Hoxby (2003), and Nechyba (2003) for discussions of the importance of general equilibrium models for understanding the consequences of particular voucher plans. For examples of such equilibrium models, see Nechyba (2003, pp. 387–414); Epple and Romano (1998, pp. 33–62); Hoxby (2001); Fernandez and Rogerson (2003, pp. 195–226).
8. Concerned with the sorting by socioeconomic status that took place under its equal-value voucher system, the Chilean government modified its national voucher system in 2008. Under the new system, the vouchers distributed to children from the one-third poorest families in the country (called Priority students) are worth 50% more than those distributed to more affluent families. Private schools that receive higher-valued vouchers are prohibited from charging Priority students any tuition or fees in excess of the value of their voucher.
9. See, for example, Hoxby (2001); Fernandez and Rogerson (2003).

socioeconomic status. We discuss the evaluation of a trial of one such voucher program in Chapter 4.

The consequences of educational vouchers continue to be debated as heatedly today as in the years following Friedman's publication of *Capitalism and Freedom* (1962). However, the debate has become vastly more sophisticated, primarily because of advances in understanding that have stemmed from the interplay of theoretical developments and new evidence. For example, there is widespread recognition today that some children are more expensive to educate than others, that peer-group effects influence student achievement, and that many parents need help in collecting the information that is necessary to make good school choices. These patterns all have implications for the design of future voucher systems and for evaluations of their impacts. The wheel of science continues to turn, and our theories evolve!

What Kind of Theories?

In this chapter, we have chosen our examples primarily from the field of economics because it is the social science discipline we know best. However, theories drawn from other social-science disciplines can also inform the design of causal educational research. Examples include theories of social capital drawn from sociology and theories of child development from psychology. The choice of a theoretical framework within which to embed the design of quantitative research depends on the nature of the causal question being asked and the knowledge base of the investigators.

We do want to emphasize, however, the distinction between social-science theory and statistical theory. In recent decades, important advances have been made in statistical theory that have led to new research designs and analytic methods, many of which are presented in later chapters of this book. New resampling methods for conducting hypothesis tests and methods for estimating statistical power when individuals are clustered in classrooms and/or schools provide two examples. The point we want to emphasize here is that statistical theory, and the methods stemming from advances in statistical theory, are methodological complements to substantive social-science theory, not substitutes.

What to Read Next

For readers interested in learning more about the role of theory in informing causal research in general, and causal research in education in

particular, there is much to read. One place to start, which provides an entrée into the relevant literature, is the brief volume *Scientific Research in Education* (Shavelson & Towne, 2002), which is the thoughtful report of a National Research Council Committee in the United States. The 2003 paper by David Cohen, Stephen Raudenbush, and Deborah Loewenberg-Ball entitled "Resources, Instruction, and Research" provides an insightful theory about the conditions under which school resources influence student learning. For a provocative view of the role of theory written by a major figure in 20th-century educational research, see John Dewey's 1929 book, *The Sources of a Science of Education.*

3

Designing Research to Address Causal Questions

One of the first actions that Grover "Russ" Whitehurst, the first director of the Institute of Education Sciences, took after assuming office in 2002 was to commission a survey of educational practitioners and policymakers in order to learn what they wanted from educational research.[1] Not surprisingly, the survey results showed that the priorities of educators depended on their responsibilities. Superintendents and other local education officials were most interested in evidence about particular curricula and instructional techniques that were effective in increasing student achievement. State-level policymakers wanted to learn about the consequences of standards-based educational reforms and the impact of particular school intervention strategies. Congressional staff wanted to know about the effectiveness of different strategies for enhancing teacher quality. Educators at all levels wanted to know about the effect of differences in resource levels, such as class sizes, in determining students' achievement.

Whereas the priorities of educators depended on their responsibilities, the striking commonality in their responses was that practitioners and policymakers—at all levels—wanted to know the answers to *questions about cause and effect*. They wanted to know *if A caused B*, and wanted IES to commission research that would provide them with answers. In this chapter, we discuss the conditions that must be satisfied for such causal questions to be addressed effectively in education, and we introduce some of the major concepts and terms that we use throughout the rest of the book.

1. See Huang et al. (2003).

Conditions to Strive for in All Research

Before we begin our discussion of how best to address the causal questions that are so central to educators, we begin with a brief description of the classical elements of good research design in the social sciences and education. We do this because designing *causal* research requires us to pay attention to the central tenets of all good research. Then, within this larger domain, causal research must satisfy an additional set of constraints, and it is these that form the central topic for the rest of our book. We used the expression "strive for" in the title of this section because it is typically difficult to satisfy all of the conditions we describe. We use examples throughout the book to clarify the consequences of not satisfying particular elements of the classical description of effective research design. As you will learn, violation of some of the tenets of appropriate design makes it impossible to make a defensible causal inference about the consequences of an educational policy or intervention. Violation of other tenets does not threaten the ability to make a causal inference, but does limit the ability to determine to whom the results of the study apply. We will return to these issues. However, we begin by stating these elements of good research design.

First, in any high-quality research, whether it be purely descriptive or able to support causal inference, it is critically important that it begin with a clear statement of the research question that will drive the project and the theory that will frame the effort. These two key elements ultimately drive every aspect of the research design, as they provide the motivation and the rationale for every design decision that you ultimately make. They have also been the topics of our first two chapters and, as we have argued, they are completely intertwined. As theories are refined, it becomes possible to pose more complex questions, and these, in their turn, inform refinements of the theory. Light, Singer, and Willett (1990) referred to this as the "wheel of science."

An explicit statement of the research question makes it possible to define the *population of interest* clearly and unambiguously. This is critical in any research. If we do not do it, we cannot build a suitable sampling frame, nor can we know to whom we can generalize the findings of our research. In addition, it pays to be explicit, rather than vague, about the nature of the population of interest. For example, in studying the impact of class size on children's reading skills, it might make sense to define the population of interest to be "all children without special needs in first-grade classrooms in urban public schools in the United States," rather than just "children." Defining the population clearly enables readers who have a particular concern, such as the impact of class size on the learning of autistic children, to judge the relevance of our results to their concern.

Once we have defined the population of interest clearly, we must work hard to sample representatively from that population. Thus, in an investigation of the impact of class size on student achievement in the population defined earlier, we need to decide whether it is feasible to obtain a simple random sample of students from the population of first graders without special needs attending urban public schools in the United States. Alternatively, we might decide that we want to use a more complex sampling plan, such as a multistage cluster sample of school districts, schools, and grades. However we go about sampling, it is critical that the analytic sample that we use in our research be fully representative of the population. This ensures what methodologists call the *external validity* of the research. This refers to the ability to generalize our findings credibly to a known population of interest.

The next important step in any research project is to choose appropriate measures of the key variables that are central to the research, and to ensure their construct validity and reliability for the population under investigation. We should use our knowledge of the research question and its supporting theory to distinguish three important classes of variables: (a) the outcome variable; (b) the principal question predictor, defined as the variable that provides our research question; and (c) the covariates or control predictors. These distinctions will recur consistently throughout our account of causal research, as they do through the account of any high-quality descriptive research project. In our hypothetical investigation of class size and academic achievement, for instance, we might decide to focus on two specific academic outcomes, such as children's reading and mathematics achievement. Our principal question predictor would be a measure of class size. Covariates or control variables might include student demographic characteristics and measures of teacher experience. We would need to exercise care in determining just how we would measure each of these variables. For example, we would need to decide whether we want to measure class size by the number of students enrolled in a class on a particular day, or perhaps by the average of the number of students enrolled on several prespecified dates. We would also want to be sure to measure each student's reading and mathematics achievement using age-appropriate normed and suitably scaled tests. Our decisions should be guided by our research question, our theoretical framework, and the background literature in which they are embedded.

At this point, we want to point out explicitly the one and only distinction between descriptive and causal research. It concerns the principal question predictor that forms the centerpiece of the research design. The critical question for causal research is how the values of the question predictor are determined for each of the participants in the sample.

In our class-size example, if the actions of children, teachers, parents, or school administrators determine the size of the class into which each child is placed, all manner of unobserved forces and choices would undermine our ability to make inferences about the causal impact of class size on children's achievement. On the other hand, if we were to randomly assign children and teachers to classes of different sizes, thereby determining their values on the principal question predictor, we may be able to credibly estimate the causal impact of class size on the achievement of children in the population from which the analytic sample was drawn. The difference is simply in the way that the values of the question predictor, class size, have been determined for each child in the analytic sample and for their teachers. This single issue and its consequences for design, data analysis, and interpretation distinguish credible causal research from all other research. It is the central concern of the rest of our book.

One final step is to ensure that the research is replicated in other samples drawn from the same population. This is important because of the uncertainty that exists in measurement and is built into the probabilistic nature of statistical inference. We will devote considerable attention in this book to describing how different kinds of statistical errors can influence the findings from statistical analysis.

Making Causal Inferences

In their excellent book on the design of social science research, Shadish, Campbell, and Cook (2002, p. 6) cite 19th-century philosopher John Stuart Mill's description of three critical conditions that must be met in order to claim that one thing *causes* another. The first condition is that the hypothesized *cause* must *precede* its anticipated *effect* in time. For example, in investigating whether student achievement depends upon the number of students in the class, it is important to ensure that students had been taught in class settings of a particular size *before* their achievement was measured.

The second of Mill's conditions is that if the levels of the cause differ in some systematic way, then there must be corresponding variation in the effect. For example, if our theory suggests that children taught in classes with fewer students achieved at higher levels, we would anticipate that as the number of students in classes got smaller, the students' achievement would be higher, on average.

The third of Mill's conditions is by far the most important and the most difficult to satisfy in practice. It stipulates that the researcher must be able to discount all other plausible explanations—other than the anticipated causal one—for the link observed between the hypothetical cause and effect.

In the case of an investigation of the impact of class size on student achievement, we must be able to argue compellingly that any observed association between class sizes and subsequent student achievement is not a consequence, for example, of choices that parents may have made about where to send their children to school or decisions by school administrators to assign students with particular characteristics to classes of particular sizes.

The most persuasive way to conduct research that satisfies Mills' three conditions—and thereby successfully address causal questions—is for the researcher to conduct an *experiment*. Following Shadish, Campbell, and Cook (2002, p. 511), we define an experiment as an empirical investigation in which the levels of a potential cause are manipulated by an outside agent functioning independently of the participants in the research, and after which the consequences for an important outcome are measured.

Furthermore, as illustrated in Figure 3.1, we distinguish between two kinds of experiments: randomized experiments and quasi-experiments. The most compelling evidence for making causal attributions typically comes from randomized experiments, defined as experiments in which units are assigned to experimental conditions by a random process, such as the toss of a fair coin (Shadish, Campbell, & Cook, 2002, p. 12). Notice that well-executed *randomized experiments* satisfy Mills's three conditions for making causal inferences: (a) cause precedes effect, (b) different levels of cause can lead to different levels of effect, and (c) random assignment obviates all other plausible explanations for any differences in effect detected. In fact, the random assignment of students and teachers to classes of different sizes by an independent investigator ensures that the children and teachers who were in the different class-size "treatments" are equal on all characteristics—on average—before the experiment begins. Because of randomization, any small and idiosyncratic differences that exist among the groups prior to treatment will fall within the noise that is

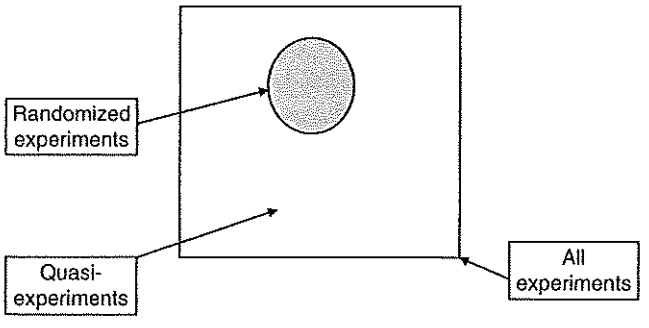

Figure 3.1 Two kinds of experiments.

accounted for naturally by statistical methods used to analyze the resulting outcome data. As we describe more fully in Chapter 4, when individuals are assigned by randomization to different experimental conditions, we say that the groups so-formed are *equal in expectation.*

Quasi-experiments are experiments in which units are not assigned to conditions randomly (Shadish, Campbell, & Cook, 2002, p. 12). It is sometimes possible to make legitimate causal inferences using data from quasi-experiments. Indeed, we devote several chapters of this book to methods for doing so. However, as we illustrate with many examples, researchers need to be prepared to deal with a variety of *threats to the internal validity* of research based on data from quasi-experiments. As we discuss in more detail in Chapter 4, this term refers to threats to the validity of a statement that the relationship between two variables is causal (Shadish, Campbell, & Cook, 2002, pp. 53–61).

Although the interpretation of the evidence from any experiment will depend on the details of the particular case, we want to emphasize one general point. Randomized experiments and quasi-experiments typically provide estimates of the total effect of a policy intervention on one or more outcomes, not the effects of the intervention holding constant the levels of other inputs (Todd & Wolpin, 2003). This matters, because families often respond to a policy intervention in a variety of ways, and the experiment provides evidence about the net impact of all of the responses on measured outcomes. For example, we will describe several experiments in which parents were offered scholarships to help pay for the education of a particular child at a private school. A common outcome in such experiments is a measure of the cognitive skills of children at a later point in time. One response to the policy is to increase the probability that parents send to a private school the child offered a scholarship. However, another response may be that the parents reduce the amount of money that they spend on providing tutoring and enrichment activities for that child in order to free up resources to devote to other children. The experiment provides evidence about the net impact of these two responses (as well as any others). It does not provide an estimate of the impact of the scholarship offer on children's subsequent achievement, holding constant the level of parental resources devoted to tutoring and enrichment.

Past Approaches to Answering Causal Questions in Education

Unfortunately, until fairly recently, most educational researchers did not address their causal questions by conducting randomized experiments or

by adopting creative approaches to analyzing data from quasi-experiments. Instead, they typically conducted *observational studies*, defined as analyzing data from settings in which the values of all variables—including those describing participation in different potential "treatments"—are observed rather than assigned by an external agent (Shadish, Campbell & Cook, 2002, p. 510). For instance, hundreds of observational studies have been conducted on the association between class size and academic achievement using achievement data collected from students during the normal operation of a school district. In these settings, the number of students in various classes differs as a result of demographic patterns, the decisions of parents about where to live, and the decisions of school administrators about placement of students into classes.

In observational studies, the skills and motivations of students in small classes may differ from those in larger classes, irrespective of any impact that class size itself may have had ultimately on their achievement. This could be the result of a variety of mechanisms. For example, families with the resources to invest in their children's education may purchase or rent homes in the attendance zones of schools with reputations for having small classes. As a result, the average achievement of students in the schools with relatively small classes may be higher than that in schools with larger classes, even if class size did not have a causal effect on student achievement. The reason could be that those parents who chose to live near schools with small classes used their resources to provide their children with educationally enriched environment at home. This is an example of what methodologists would call an *endogenous* assignment of participants to treatments. By this we mean that assignment to levels of the treatment is a result of actions by participants within the system being investigated—in this case, the decisions of parents with resources to take advantage of the relatively small classes offered in particular schools.

Of course, well-trained quantitative researchers recognized that, as a result of the decisions of parents and school administrators, students placed endogenously in classes of different sizes may differ from each other in respects that are difficult to observe and measure. For many years, researchers responded to this dilemma in one of two ways. One common response was to include increasingly larger and richer sets of covariates describing the students and their families in the statistical models that were used to estimate the effect of treatment on outcome. The hope was that the presence of these control predictors would account for differences in the outcome that were due to all of the unobserved—and endogenously generated—differences among students in classes of different size. Sociologists Stephen Morgan and Christopher Winship (2007, p. 10) refer to the period in which researchers relied on this strategy as

"the age of regression." Seminal studies published in the 1980s threw cold water on this "control for everything" strategy by demonstrating that regression analyses that contained a very rich set of covariates did *not* reproduce consistently the results of experiments in which individuals were assigned randomly to different experimental conditions.[2]

A second response, especially common among developmental psychologists, was to accept that analysis of observational data could not support causal inference and to simply avoid using causal language in both the framing of research questions and in the interpretation of research results. For example, researchers would investigate whether children placed in center-based child care had better subsequent performance on cognitive tests than did observationally similar children in family-based child care, and would simply caution that causal attribution was not justified on the basis of their findings. In our view, there are at least two problems with this approach. First, the cautions presented in the "Methods" and "Results" sections of research papers were often forgotten in the "Discussion" section, where researchers would suggest policy implications that depended on an unsupported causal interpretation of their findings. Second, their use of noncausal language meant that these researchers were not accustomed to considering explicitly alternative explanations for the statistical relationships they observed.

Fortunately, in more recent years, social scientists have developed a variety of new research designs and analytic strategies that offer greater promise for addressing causal questions about the impact of educational policies. Many of these new approaches also make use of standard techniques of multiple regression analysis, but apply them in new ways. Explaining these strategies, and illustrating their use, is a central goal of this book.

The Key Challenge of Causal Research

In conducting causal research in education and the social sciences, our central objective is to determine how the outcomes for individuals who receive a treatment differ from what the outcomes would have been in the absence of the treatment. The condition to which the research subjects would have been exposed in the absence of the experimental treatment is called the *counterfactual*. From a theoretical standpoint, the way to obtain an ideal counterfactual would be to use the same participants under both

2. See Angrist & Pischke (2009, pp. 86–91) for a discussion of this evidence.

a treatment (e.g., "small" class size) and a "control" (e.g., "normal" class size) condition, resetting all internal and external conditions to their identical initial values before participants experienced either condition. So, you might draw a representative sample of participants from the population, administer the treatment to them, and measure their outcome values afterward. Then, to learn what the outcomes would be under the counterfactual condition, you would need to transport these same participants back to a time before your research was conducted, erase all their experiences of the treatment and the outcome measurement from their memories, and measure their values of the outcome again, after their lives had transpired under the control condition. If this were possible, you could argue convincingly that any difference in each participant's outcome values between the two conditions must be due *only* to their experiences of the treatment.

Then, because you possessed values of the outcome for each individual obtained under both "factual" and 'counterfactual" conditions, you would be able to estimate the effect of the treatment for each participant. We call this the *individual treatment effect* (ITE). You would do this simply by subtracting the value of the outcome obtained under the counterfactual condition from the value obtained under the treated condition. In this imaginary world, you could then average these estimated ITEs across all members of the sample to obtain the estimated *average treatment effect* (ATE) for the entire group. Finally, with a statistical technique like a simple paired *t*-test, you could seek to reject the null hypothesis that the population mean difference in participants' outcomes between the treated and counterfactual conditions was zero. On its rejection, you could use your estimate of the ATE as an *unbiased* estimate of the *causal* effect of the treatment in the population from which you had sampled the participants.

Since time travel and selective memory erasure lie in the realm of imagination rather than research, in practice you always have a "missing data" problem. As we illustrate in Figure 3.2, you never actually know the value of the outcome for any individual under both the treatment and control conditions. Instead, for members of the treatment group, you are missing the value of the outcome under the control condition, and for members of the control group, you are missing the value of the outcome under the treatment condition. Consequently, you can no longer estimate the individual treatment effects and average them up to obtain an estimate of the average treatment effect.

So, you must devise an alternative, practical strategy for estimating the average treatment effect. The reason that this is so difficult to do in practice is that actors in the educational system typically care a lot about which experimental units (whether they be *students* or *teachers* or *schools*) are

	... the value of the outcome in the ***Treatment Group*** is ...	... the value of the outcome in the ***Control Group*** is ...
For members of the ***Treatment Group*** ...	Known	Missing
For members of the ***Control Group*** ...	Missing	Known

Figure 3.2 The challenge of the counterfactual.

assigned to particular educational treatments, and they take actions to try to influence these assignments. In other words, the assignment of participants to treatments is typically *endogenous* in educational research. A consequence of this is that, in an investigation of the impact of class size on academic achievement, students assigned endogenously to differently sized classes are likely to differ from each other, and not only on dimensions that can be *observed* (such as gender, age, and socioeconomic status), but also on dimensions that remain unobserved (such as intrinsic motivation and parental commitment, both of which are likely to be associated with achievement outcomes).

One positive way to restate this point—and to satisfy Mills's third condition for making compelling causal inferences—is to insist that the assignment of participants to treatments be *exogenous* rather than *endogenous*. According to the *Oxford English Dictionary*, *exogenous* means "relating to *external* causes," and is the natural opposite of *endogenous*, which means "relating to an *internal* cause or origin." In the context of our book, these words have similar, though more refined and specific meanings. When we say that there is "exogenous variation" in the educational treatments that students receive, we mean that the assignment of students to treatments has *not* been determined by participants *within* the educational system—that is, by the students, parents, teachers, or administrators—themselves. Instead, their placement in a particular treatment condition has been determined "externally" by the investigator or some other independent agency.

Of course, you might argue that it is not good enough for assignment to treatment condition to be simply *exogenous*. It is possible, for instance, that even external agents may be biased or corrupt in their assignment of participants to treatment conditions. Typically, though, when we say that assignment to experimental conditions is exogenous, we are assuming

that the external agent has exercised his or her opportunity to assign participants in a way that supports causal inference directly. One very simple and useful way that such exogenous variation in experimental conditions can be created is for the investigator to assign participants randomly to treatments. Such an approach was taken in the Tennessee Student/Teacher Achievement Ratio (STAR) experiment (Krueger, 1999).

In the mid-1980s, the Tennessee state legislature appropriated funding for a randomized experiment to evaluate the causal impact of class-size reduction on the reading and mathematics achievement of children in the primary grades. More than 11,000 students and 1,300 teachers in 79 public schools throughout the state participated in the experiment, which became known as Project STAR. In each participating school, children entering kindergarten in the fall of 1985 were assigned randomly by investigators to one of three types of classes: (a) a small class with 13 to 17 children, (b) a class of regular size with 22 to 25 students, or (c) a class of regular size staffed by both a teacher and a full-time teacher's aide. Teachers in each school were also assigned randomly to classrooms. Finally, the research design called for students to remain in their originally designated class type through third grade.

A major theme of our book is that some element of exogeneity in the assignment of units to a treatment is necessary in order to make causal inferences about the effects of that treatment. Expressed in the formal terms used by statisticians and quantitative social scientists, a source of exogenous assignment of units to treatments is necessary to *identify* the causal impact of the treatment. So, when a social scientist asks what *identification strategy* was used in a particular study, the question is about the source of the exogeneity in the assignment of units to treatments. In subsequent chapters, we show that randomization is not the only way of obtaining useful exogenous variation in treatment status and consequently of identifying the causal impact of a treatment. Sometimes, it is possible to do so with data from a quasi-experiment. Sometimes, it is even possible to do so with data from an observational study, using a statistical method known as *instrumental-variables estimation* that we introduce in Chapter 10.

The Tennessee STAR experiment, which the eminent Harvard statistician Frederick Mosteller called "one of the most important educational investigations ever carried out" (Mosteller 1995, p. 113), illustrates the difficulties in satisfying all of the conditions for good research that we described earlier in this chapter. After the Tennessee legislature authorized the experiment in 1985, the State Commissioner of Education invited all public school systems and elementary schools in the state to

apply to participate. Approximately 180 schools did so, 100 of which were sufficiently large to satisfy the design criterion of having three classes at each grade level from kindergarten through grade 3. The research team then chose 79 schools to participate.

The process of selecting schools to participate in the STAR experiment illustrates some of the compromises with best research practice that are sometimes necessary in even extremely well-planned experiments. First, the research sample of schools was chosen from the set of schools that volunteered to participate. It is possible that the schools that volunteered differed from those that did not in dimensions such as the quality of leadership. Second, only quite large schools met the design requirements and consequently the STAR experiment provided no evidence about the impact of class size on student achievement in small schools. Third, although the research team was careful to include in the research sample urban, suburban, and rural schools, as the enabling legislation mandated, it did not randomly select 79 schools from the population of 100 schools that volunteered and met the size criteria (Folger, 1989). A consequence of the sample selection process is that the definition of the population of schools to which the results of the experiment could be generalized is not completely clear. The most that can be said is that the results pertain to large elementary schools in Tennessee that volunteered to participate in the class-size experiment. It is important to understand that the lack of clarity about the population from which the sample is taken is a matter of *external validity*. The sampling strategy did not threaten the *internal validity* of the experiment because students and teachers within participating schools were randomized to treatment conditions.

The STAR experiment also encountered challenges to *internal validity*. Even though children in participating schools had originally been randomly and exogenously assigned to classes of different sizes, some parents were successful in switching their children from a regular-size class to a small class at the start of the second school year. This endogenous manipulation had the potential to violate the principal assumption that underpinned the randomized experiment, namely, that the average achievement of the students in regular-size classes provided a compelling estimate of what the average achievement of the students placed in the small classes would have been in the absence of the intervention. The actions of these parents therefore posed a threat to the internal validity of the causal inferences made from data collected in the STAR experiment about the impact of a second year of placement in a small class.

This term, *threat to internal validity*, is important in the annals of causal research and was one of four types of validity threats that Donald Campbell

(1957), a pioneer in developing methods for making causal inferences, described more than a half century ago. As mentioned earlier, it refers to rival explanations for the statistical relationships observed between educational treatments and outcomes. If we can remove all threats to internal validity, we have eliminated all alternative explanations for the link between cause and effect, and satisfied Mills's third condition. Devising strategies to respond to threats to internal validity is a critical part of good social science. Of course, in quasi-experimental and observational research, ruling out *all* potential rival explanations for the hypothesized link between "cause" and "effect" is extraordinarily difficult to do. How do you know when you have enumerated and dismissed all potential rival explanations? The short answer is that you *never* do know with certainty (although, of course, with each rival explanation that you do succeed in ruling out explicitly, the stronger is your case for claiming a causal link between treatment and outcome, even in quasi-experimental and observational research). As we explain in the next chapter, one of the great advantages of the classic randomized experimental design, in which a sample of participants is assigned randomly to different treatments, is that this process eliminates all alternative explanations for any relationship between class size and student achievement. But, even in randomized experiments, things can go wrong, and you may have to provide evidence for the internal validity of your work. In Chapter 5, we describe some of the problems that can crop up in randomized experiments and how skilled researchers have dealt with them.

Perhaps the most important lesson to take away from this chapter is that the active behaviors of the participants in the educational system—teachers, administrators, parents, and students—have enormous impacts on the quality of the education provided in particular schools and classrooms. These active behaviors often make it very difficult to conduct internally valid evaluations of the impacts of educational initiatives, whether they involve the placement of students in smaller classes, the use of new curricula and instructional methods, the installation of new ways to prepare teachers, or the creation of new governance structures. In the chapters that follow, we show how new sources of data, new approaches to research design, and new data-analytic methods have improved our ability to conduct internally valid studies of the causal impact of educational initiatives on student outcomes. We will make use of the terms introduced in this chapter, including *randomized experiment*, *quasi-experiment*, *observational study*, *exogenous*, *endogenous*, and *threats to internal and external validity*. By the time you have finished reading our book, these terms will be old friends.

What to Read Next

For readers who wish to follow up on the ideas we have raised in this chapter, we recommend Shadish, Campbell, and Cook's comprehensive book (2002) on the design of research, *Experimental and Quasi-Experimental Designs*, and Morgan and Winship's insightful book (2007), *Counterfactuals and Causal Inference*.