1

Asking the Right Questions

Does education research have any impact on the instructional practices, curricula, and policies in your classroom, school, or district? Probably not, if you are like many educators we know. You may even secretly believe that your own common sense and experience are far more trustworthy than the experiments and observations of researchers. We all know individuals who wouldn't dream of buying a new car or choosing a treatment for a medical condition without researching the options. Yet on the job, they will commit hundreds of thousands of dollars of their schools' or districts' budgets to an innovative or supposedly exemplary program without carefully evaluating the available research findings.

One elementary school principal explained the problem this way: "We tend to move from one fad to another in order to demonstrate that we are 'state of the art' even though most of the activities have little impact. There is big money in selling education programs and consultants use 'research says' to sell programs that purportedly can fix just about anything. Most . . . teachers and administrators can't differentiate viable research from poor research" (Walker, 1996, p. 41).

One can certainly understand why some practitioners dismiss education research as irrelevant to their daily lives and continue to "do their own thing." Even insiders concede that there are problems with it: poor research designs and sloppy statistics (Cook, 1999), divisive bickering (Gage, 1989; Snow, 2001), and petty politics (Shaker & Heilman, 2002). Others are more optimistic about the potential of education research to inform practice: "Research is the most powerful instrument to improve student achievement—if only we would try it in a serious and sustained manner" (National Educational Research Policy and Priorities Board, 2000a, p. 1).

This statement serves as a challenge to both researchers *and* educators. Researchers have an obligation to produce useable knowledge for practitioners, but educators are no less accountable for applying what is already known to the practice of schooling. We bear an additional responsibility as well—that of holding publishers, curriculum developers, and consultants accountable for evaluating their products and models using rigorous research techniques and then making that research available to practitioners.

WHAT IS *OUR* APPROACH TO MAKING SENSE OF RESEARCH?

Before proceeding, it is important to clarify some fairly broad assumptions that we will make about research: (a) that one *can* frame a meaningful question related to educational practice, (b) that one *can* develop a hypothesis related to that question, (c) that one *can* design a study—whether quantitative, qualitative or ideally a skillful combination of both—and collect data to assess the hypothesis, (d) that one *can* assess whether the data support the hypothesis with some degree of certainty (or uncertainty), and (e) that one *can* apply this knowledge, within reason, to inform decision making, whether at the classroom, school, district, state, or national level.[1]

Even if research appears to follow the preceding steps, one must be circumspect about accepting the "facts" that it purports to establish. In the manner of Sherlock Holmes, we have to sift through research evidence to determine what it is really saying. The issue of causality is a good example, one that receives particular emphasis in this book. Life in schools—and, indeed, life in general—is rife with causal statements: "Our test scores went up *because* of the new reading program." "Teachers are leaving the school in droves *because* of low salaries."

If this book accomplishes nothing else, it aims to convince the reader that causal statements cannot be made in a cavalier fashion. Schools are complex and multifaceted. Causal links are difficult to establish with certainty, if only because there are usually alternative—and sometimes equally plausible—causes for that which we seek to explain. For example, it may be that test scores went up not because of a new reading program, but *because* of an influx of well-to-do students or a sudden exclusion of low-achieving students from the testing. Or it may be that salaries, low as they are, have nothing whatever to do with the mass exodus of teachers. They are leaving *because* the principal makes their lives

Causal links are difficult to establish with certainty, if only because there are usually alternative—and sometimes equally plausible—causes for that which we seek to explain.

miserable. Although establishing clear causal links is a daunting assignment, the task is easier if one pays close attention to rules-of-thumb (to be introduced throughout the book) that have been developed and refined by generations of social scientists. This book will repeatedly emphasize that identifying "good" research hinges on understanding—and critiquing—the causal underpinnings of that research.

Five Questions About Research

> Our district has tried numerous strategies: we lifted a school day; we increased time on-task; we increased the graduation requirement; we mandated exit testing; and we put in a no-driver's-license-if-you-drop-out provision. Many other school boards have tried instituting similar enhancement policies. Locally we try to deal with attendance and discipline rules, but these measures alter the nature of the system without addressing the root causes of the problem. We have audited our rules for compliance purposes. What needs to be examined now is the unhappy consequence of these efforts: there have been no significant improvements in student achievement patterns. These innovations have failed to eliminate poor instruction and ineffective and redundant curricula. This raises the question of exactly what our professional roles are going to be to help more students become prepared for a new century. (Dorn, 1995, p. 7)

"What exactly am I supposed to be doing as a principal?"

You can read between the lines of this lament by a Florida high school principal. His superintendent and school board no doubt issued a mandate: "Raise student achievement." This is a tough assignment at the high school level, or at any other level for that matter. Principal Dorn and his staff seemingly tried every strategy, idea, and innovation they could think of and nothing has worked. His frustration is palpable as he raised a very critical question: "What exactly am I supposed to be doing as a principal?"

This book does not—it cannot—offer a single answer to Dorn or to others that share his goal of improving schools. In fact, we are deeply skeptical of consultants, salespeople, and project leaders who purport to provide such answers. Rather, it suggests five questions that Dorn and educators like him should ask of research. By presenting these questions, we do not aim to waffle (thus evoking Harry Truman's plea for a one-armed economist, so that he might never hear "on the one hand" and "on the other hand"). The questions simply acknowledge two hard realities about education research.

First, *all* authors think that their research is "good" and worthy of a receptive audience. How can the beleaguered practitioner separate the wheat from the chaff? Some quality-control mechanisms already exist, of course. There are academic journals with more rigorous quality standards than others (enforced by impartial and anonymous reviewers). A great deal of the worst research is never published at all. Yet even good journals publish studies with overstatements, misstatements, and downright fabrications. Further, the proliferation of sub-standard journals, self-published Web sites, and advocacy research organizations means that it is increasingly difficult *not* to find an outlet for publication. Education is full of well-intentioned, but occasionally ineffective, attempts to synthesize and communicate research findings (e.g., Berliner & Casanova, 1993; Zemelman, Daniels, & Hyde, 1998). At the end of the day, a practitioner's common sense may have to be the final arbiter of what constitutes "good" research. We firmly believe that every practitioner can become a more informed and critical consumer of research, if armed with the right questions.

Even good journals publish studies with overstatements, misstatements, and downright fabrications.

The second hard reality is that most researchers have the relative luxury of not having to worry about the implementation of their findings in classrooms and schools. Unfortunately, research findings do not always translate easily to the ambiguities of educational practice. Thousands of victims of botched staff development or reforms gone haywire can attest to that. What works in one context may fail miserably in another. Thus, research cannot be analyzed solely on the basis of the hermetically sealed bubble in which it was conducted; it must be evaluated in light of the context in which the findings will be applied, whether a classroom, school, district, or state. Once again, practitioners are often in the best position to decide what might work for them, based on their informed reading of the research.

Practitioners are often in the best position to decide what might work for them, based on their informed reading of the research.

To aid practitioners in their quest for understanding, we offer five questions:

1. The causal question: Does it work?
2. The process question: How does it work?
3. The cost question: Is it worthwhile?
4. The usability question: Will it work for me?
5. The evaluation question: Is it working for me?

Apply the first four questions to research *before* you adopt or implement a program, method, or policy. For example, if you (and your team) are considering the implementation of "multiple

intelligences" as a way to organize curricula and design instruction, or contemplating the use of block scheduling as a means of raising achievement in your high school, seek out every available research study—both quantitative and qualitative. Read them carefully to determine if *it* (i.e., multiple intelligences or block scheduling) actually works, how *it* works, if *it's* worthwhile, and if *it* will work for you. Ask these questions *before* you commit to adoption.

Once you have made the decision to adopt and actually begin implementation, ask the fifth and final question—is *it* working for me. Ask this question at several points *during* implementation (i.e., a formative evaluation to fine-tune the implementation process) and then ask it again *after* implementation (i.e., a summative evaluation to determine the overall effectiveness of the method or policy). Figure 1.1, Asking the Right Questions, shows how and when the five questions can be used to make sense of research.

How many times have you adopted a program and then hurriedly moved on to another priority, failing to investigate whether the program you *believed* would work based either on research or a glowing testimonial, is actually working? Question 5 can only be answered by going "on location" in your classroom, school, or district and conducting what we call *user-driven research*.

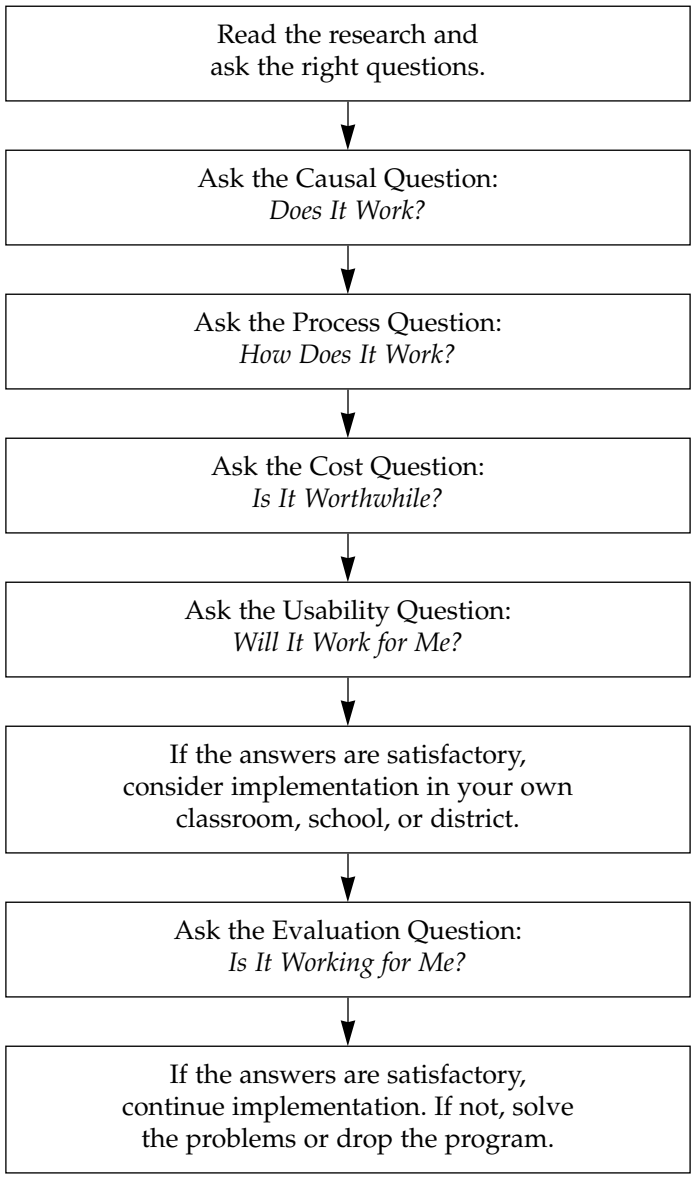
The bottom line is whether the intervention being considered "causes" a particular outcome to change.

The Causal Question: Does It Work?

Many research studies in education, primarily those with a quantitative bent, test whether a causal relationship exists between interventions and outcomes.[2] The interventions are infinite, ranging from ability grouping to whole-school reform. The tested outcomes are equally diverse, although the biases of experimental researchers, as well as most parents (Public Agenda, 2002), tend to favor some type of standardized measure of achievement.

The bottom line of these studies is whether the intervention being considered "causes" a particular outcome to change. For example, did added teacher training in a school district *cause* students' test scores to rise? Drawing this inference is rarely as simple as we would like to believe. Even if we observe that test scores rose shortly after teachers participated in training, we must immediately rule out *every other potential explanation* for the change in test scores. Only then can we credibly infer that a causal relationship exists between training and test scores. The range of alternate explanations (and whether they are remotely plausible) will depend on the particulars of the research context. Perhaps the school district enrolled a new group of wealthy (and high-achieving) students. In this instance, the rise in test scores would have occurred even in the absence of staff development.

Figure 1.1 Asking the Right Questions



High-quality research in education can spare readers from this guesswork. It does so by adhering to tried and true elements of good research design. One of the best designs is the randomized experiment. There is a treatment group (of students, schools, or other

units) that participates in an intervention. Likewise, there is a control group that does not receive the intervention. The sine qua non of an experiment, however, is randomized assignment to treatment and control groups. That is, the assignment of a particular student or school to either group depends on nothing more than chance (or, quite literally, the flip of a coin). If implemented properly, a good experiment allows the researcher to convincingly establish whether an intervention is the *only* plausible explanation for gains (or losses) in valued outcomes. Why? The answer lies in the fact that with randomized assignment we can be virtually certain that the only difference between the two groups is their exposure to the intervention. They were essentially identical at the outset of the experiment. Chapter 3 will define the elements of a randomized experiment in much greater detail, and describe why it is such a powerful means of obtaining strong answers to the causal question: Does it work?

The sine qua non of an experiment is randomized assignment to treatment and control groups.

Unfortunately for those of us looking for more certainty, randomized experiments are often the exception in education research (though not in other fields, such as health). The fallback position of most education researchers is the quasi-experiment. This is a much-used (and abused) term that refers to research designs meant to replicate experimental conditions. Like experiments, quasi-experiments usually rely on comparing the outcomes of a treatment group that received an intervention to a comparison group that did not.[3] Unlike experiments, there is no randomized assignment to these groups—hence the prefix "quasi." You might already be imagining the pitfalls to be encountered in a poorly done quasi-experiment. Let's say we are trying to assess whether a new reading curriculum "works" by comparing the test scores of the students in a classroom that received the treatment (i.e., the new reading program) to the scores of students in another classroom who did not experience the new program. Suppose that the "treated" class is taught by a highly effective veteran teacher, whereas the other class is taught by an embittered, burnt-out one. Did the curriculum work? Any differences in outcomes between the two classes *might* be due to the reading program, but the quality of the instruction could also be the explanation for the higher scores.

Of course, a very good quasi-experiment would go to extraordinary lengths to rule out explanations besides the curriculum, thus deriving a strong "causal" conclusion. It would do so through careful matching procedures or sophisticated statistical techniques or any number of other methods. (Shadish et al., 2002). We will provide a sampling of these in Chapter 4, as well as some of the hazards in their application and interpretation.

The Process Question: How Does It Work?

Even if we succeed in uncovering a causal relationship, that knowledge does not tell us exactly *why* a relationship exists between a treatment and an effect. In other words, we do not know the possible classroom and school processes or interactions that may have mediated the causal relationship, or lack thereof. If reduced class size is found to improve achievement, for example, it presumably does not happen by magic. Something occurs in the instructional setting of smaller-size classes that produces higher student achievement: Teachers may devote more time to students with learning challenges, or smaller classes may promote a better disciplinary climate, or students may be more inclined to participate in a less crowded classroom. The "how" of this causal relationship may be a combination of all of these explanations—or none of them. The point is that causal description—evidence that something happened via an experiment or quasi-experiment—does not always imply or provide a causal explanation of how something happened (Shadish et al., 2002).

Yet understanding the processes that undergird a positive (or negative) causal finding can greatly enrich our knowledge. How was the intervention implemented and were there any significant setbacks or roadblocks? Was this implementation responsible, to some degree, for the observed outcomes? Were some elements of the intervention more effective than others, and might these be profitably emphasized in subsequent implementations? And how, exactly, did the intervention succeed in altering the outcomes?

Researchers may find such information helpful in testing and even refining the theory that led them to conduct the research in the first place. Even more germane from this book's perspective is that practitioners may find it useful as they evaluate the relevance of a research study for their own purposes. Knowledge of such processes can be vital in deciding whether and how the findings of a research study might be replicated—or improved upon—in an alternate setting.

The "thick descriptions" necessary to understand processes can often be obtained from research with a more qualitative bent.

The "thick descriptions" necessary to understand processes can often be obtained from research with a more qualitative bent (although the best quantitative studies will also gather information on processes). A variety of techniques—ranging from case studies to ethnographies—fall into this category. Chapter 5 will describe some of these and show how they can be a useful means of answering perplexing questions about *how* education interventions work.

The Cost Question: Is It Worthwhile?

Teachers, principals, and even superintendents often perceive costs as an issue for someone else—the district business manager or accountants. Even most books on education research and evaluation take up the issue of costs as an afterthought, if it is given any consideration at all. In contrast, we think the issue of costs is central for all practitioners, particularly when they are contemplating new programs or policies. How many schools, for example, have introduced a whole-school reform package, only to discover that it stretched their personnel to the breaking point and caused other important priorities to be shelved? How many districts have enthusiastically pursued class size reduction, only to find that limited space has consigned students and their newly hired teachers to converted janitorial closets?

Teachers, principals, and even superintendents often perceive costs as an issue for someone else—the district business manager or accountants.

These are precisely the conundrums faced by schools that act upon research findings—however glowing the "causal" findings—without considering the cost implications of decisions. You will discover that our definition of costs is broader than the average accountant's. To be sure, it includes typical resource costs like expenditures on salaries, instructional materials, and facilities; salaries are especially important in the labor-intensive work of education. However, it also includes costs that never pass through an accountant's balance sheet. Imagine a policy that is apparently "costless"—say, the implementation of a mandatory period of sustained silent reading for all students in an elementary school. No new teachers are hired and no money exchanges hands. And yet, teacher and student time is now occupied in silent reading that would have been devoted to other activities like tutoring, studying, and grading. Unless these tasks were *entirely* unproductive, something was sacrificed by reallocating the time of teachers and students. The value of this lost opportunity is a cost—perhaps difficult to quantify—but a cost nonetheless. It should clearly factor in to the decisions of practitioners.

Chapter 6 will introduce two general approaches to considering costs: cost-feasibility and cost-effectiveness. They are familiar enough concepts in everyday life, but somehow they have become estranged from decision making in the educational arena. Cost-feasibility corresponds to the essential question: Can we afford it? In other words, what are the monetary (and nonmonetary) costs of a particular intervention, and is it feasible for our district (or school or classroom) to bear those costs?

Cost-effectiveness corresponds to a broader and *comparative* question: Is there a less costly means of accomplishing the same thing? Suppose that a particular intervention has been shown to be effective across a wide range of schools, including yours. And yet, it comes

with a hefty price tag. We might legitimately ask, can the same effects be obtained via another intervention at a lower cost? Or reversing the question, does another intervention, costing the same amount, provide even larger effects? In short, what intervention will give us the most "bang for our buck"?

Can the same effects be obtained via another intervention at a lower cost?

The Usability Question: Will It Work for Me?

Let's suppose that you do find a research study that credibly demonstrates that an intervention "works," succeeds in describing why, and is worthwhile in terms of both cost-feasibility and cost-effectiveness. Consider, as an example, an after-school tutoring program that has been shown to increase mathematics achievement. Do those results automatically imply that the program will work everywhere? More to the point, do the results imply that it will work in *your* school or district? The answer is often "no," which won't surprise practitioners who are occasionally the willing or unwilling victims of "proven" reforms.

Let's consider several hypothetical instances in which the positive results from the after-school program research study may not apply to your school. Imagine that the research study was conducted in schools that enroll mostly upper-middle-class students. Your student body, on the other hand, contains students from poor and working-class backgrounds, many of whom receive free-and-reduced lunches. These students have after-school obligations, ranging from child care for their younger siblings to part-time jobs. Targeting and retaining them in such a program in your school would be difficult.

In addition, the tutoring in the research study was done by a group of trained parent volunteers, many of whom were given release time from jobs at the local university. In your community, parent volunteers are in short supply due to the economic constraints just cited. Instead, your tutors will be drawn from the ranks of local high-school students who will receive only an abbreviated orientation. You suspect that the altered version of the treatment might work, though with a diminished impact on mathematics achievement.

Last, the research study only examined the impact of mathematics achievement, using a special test instrument developed by the researcher, aligned with that school's curriculum. You plan to evaluate mathematics achievement using a very different instrument—the one mandated by your school district. And although mathematics achievement is certainly a focus of your school's efforts to improve, there are also great concerns about raising reading achievement. You wonder if the tutoring program might have ancillary benefits for students'

reading, particularly given the text-intensive approach to teaching mathematics that is favored in the tutoring program.

The preceding example suggests that caution is warranted when generalizing the causal conclusions of a particular research study, whether quantitative or qualitative, to your context. Perhaps the student populations are different. Or the treatment will undergo modifications. Or different outcomes are considered. This bad news is compounded by the undeniable reality that you will *never* find an exhaustive description of precisely where and how the research would be most applicable.

All is not lost, however. Many researchers are quite careful in evaluating the conditions under which their findings might be generalized. While a single study might provide only a few clues—perhaps limited by the scope of its data—you can often locate wide-ranging summaries of the available research on a given topic that use a group of statistical techniques known as meta-analysis. In doing so, one can learn a great deal about whether causal conclusions hold across a wide range of students, treatment variations, and outcomes. And even in the absence of firm evidence about generalizability, there are common sense rules for answering the question, Will it work for me? Chapter 7 will describe them.

The Evaluation Question: Is It Working for Me?

Education research is often—and unnecessarily—confined to the rarefied atmosphere of academia. There, highly trained intellectuals often focus on very narrow or particular aspects of schooling. They often zero in, whether quantitatively or qualitatively, without regard for the problems or possibilities inherent in replicating or implementing what they have researched. Once the findings are disseminated, practitioners are left "on their own" to figure out how what is purported to be a fabulous instructional innovation or a breakthrough in learning can make a faithful leap from the drawing board to the classroom.

We believe that the implementation of such "research-based" programs must be accompanied by practitioner-initiated and site-specific research that never stops asking the question, Is it working for me—in my classroom, school, or district? Posing *this* question means applying the questions we have previously asked with regard to actual research studies in the context of your own setting. First, you will seek to know *if* a program is actually doing what it is supposed to be doing; then, determine *how* it is doing it; next, investigate its costs; and last, consider whether it makes sense to modify, sustain, or even expand these efforts.

We believe that the implementation of "research-based" programs must be accompanied by practitioner-initiated and site-specific research that never stops asking the question, Is it working in my school or district?

Answering these questions need not dictate that districts will design their own randomized experiment (although they certainly could, and even should if a program's effects are uncertain). Nor does it necessarily mean that teachers will prepare case studies of their own implementation experiences (although they might and could even receive some type of credit or stipend for doing so). Chapter 8 will provide examples of user-driven research that you can adapt for your own use.

Use the five questions to make sense of these debates and disagreements for yourself.

Four Case Studies of Education Research

The four case studies introduced in this section will appear again in Chapters 3 to 7, not only to illustrate key concepts about the kind of research under discussion but also to demonstrate how to ask and find answers to the first four questions using actual studies. The bodies of research that we have chosen are somewhat controversial in nature. There are nearly equal numbers of scholars and researchers on either side of these issues—both claiming to know how the research should be interpreted and what it means for practitioners. We submit that you needn't rely exclusively on the opinions of these researchers. Instead, you can use the five questions to make sense of these debates and disagreements for yourself.

CASE STUDY

Class Size Reduction: Does Size Matter?

Ask any parent whether they prefer to have their child in a small class or a large one, and you might be greeted with a look of disbelief. Among many, it is almost an article of faith that reducing class sizes will improve student outcomes, by virtue of allowing teachers to spend more time with individual students. Apparently, many politicians agree with this logic. Legislatures in more than 20 states have decided to spend large amounts of money on lowering class sizes, usually with the expressly stated goal of improving student performance.[4] One of the most well-known examples of class size reduction occurred in California. In 1996, an education budget was approved by the California State Legislature, part of Establish CRS Program Bill of 1996, that allowed local districts to receive state funds for class size reductions in the early grades (Johnston, 1996). As of 2000-2001, districts received $850 for each K-3 student enrolled in a class of 20 or fewer students (Stecher & Bornstedt, 2002).

You might assume that that an avalanche of high-quality research on class size reduction has swiftly followed these investments—but you would be wrong. It is true that some states have devoted more resources to research than others; California, Tennessee, and Wisconsin are good examples (Jacobson, 2001). But in the vast majority of cases, it is *assumed* that class size reduction is producing benefits for schools and students. In Nevada, for example, "officials at the state education department still have little proof that the long- running initiative is really improving student achievement," despite four previous evaluations (Jacobson, 2001).

Yet most proponents of class size reduction still claim to have research on their side. That research is drawn from just a few states, with findings from Tennessee occupying center-stage. You have likely heard of the Tennessee STAR project. In 1985, the Tennessee legislature funded an experiment in 79 schools in which students were randomly assigned to "regular" classes (22-26 students) and "small" classes (13-17 students). Ignore for the moment the fact that some teachers would trade their planning period for what constitutes a regular-sized class in Tennessee. The experiment showed that standardized test scores rise because of attendance in smaller classes and that these differences seem to persist (Grissmer, 1999).[5]

Predictably, there are researchers who dispute the rosy conclusions of the Tennessee STAR experiment. Stanford University economist Eric Hanushek is one of the most prominent voices of dissent. He has published influential reviews of the education "production function"

(Continued)

(Continued)

literature, in which authors use statistical methods to search for the correlation between school resources and student achievement. He concludes from these non-experimental studies that the "evidence does not suggest that any substantial achievement gains would accrue to general class size reduction policies of the type recently discussed and implemented in various jurisdictions around the United States" (Hanushek, 1999, p. 144). Hanushek's opinions are oft repeated by opponents of class size plans, and Hanushek himself has testified in a wide array of court cases where the effectiveness of school resources is at issue.

So who is right? As you might have surmised by now, we think this is the wrong question—or at least an incomplete one. From the vantage point of the decision maker or practitioner, the evidence on class size needs to be subjected to the questions described earlier. Does class size have a *causal* relationship to outcomes? If so, why does this relationship exist? How much will it cost, and is this expense justified? Even if research shows a causal relationship, will it necessarily carry over to *my* school or district?

CASE STUDY

Phonics: Can It Teach Them All to Read?

There are few more contentious curricular forums than that of reading.[6] In no other instructional arena, save perhaps math or bilingual education, are the issues as hotly debated. Given that the ability to gain meaning from the printed page is one of the most important building blocks of academic success, making sense of the reading debate is essential for educators at every level.

Phonics (a method for teaching students how to decode or sound out words) and *whole language* (an instructional philosophy variously interpreted by teachers and academics but usually centered on learning to read using a whole-word approach in the context of reading literature)[7] are the terms that have characterized the debate in the past. Most recently, the debate has revolved around studies sponsored by the National Institute of Child Health and Human Development (NICHHD; e.g., Foorman, Fletcher, Francis, Schatschneider, & Mehta, 1998)[8] and a report on reading instruction from The National Reading Panel (2000). The Houston Study (Foorman et al., 1998), named after the school district in which the research took place, is the most controversial of the NICHHD research studies; it compared classroom reading instructional methods containing three types of phonics instruction—explicit, embedded, and implicit. It concluded that an advantage exists for reading instructional approaches that emphasized explicit instruction in the alphabetic principle (phonics) for at-risk children.

Barbara Taylor, Richard Anderson, Kathryn Au, and Taffy Raphael (2000), respected and high profile reading academics, vigorously critiqued the study and condemned the way it was "overly promoted by the media and misused by some policy makers and educational leaders to support a simple solution to the complex problem of raising the literacy of young children in high-poverty neighborhoods" (p. 16). Taylor and her colleagues took issue with what they believed were four erroneous assumptions on the part of Foorman and her colleagues (1998): (a) reading is simply a matter of reading words as opposed to gaining meaning, (b) reading difficulties occur because students are deprived of instruction and also suffer from lack of home preparation in understanding the alphabetic principle, (c) instructional methods equal teaching, and (d) "training" teachers in a specific methodology equates with professional development. The critics seemed less concerned with the research design or its findings about the effectiveness of phonics instruction than with the publicity the study had received and the

(Continued)

(Continued)

associated political fallout for researchers whose agendas were not experimental or quasi-experimental in nature.

Foorman, Fletcher, Francis, and Schatschneider (2000) issued a lengthy rebuttal, and reinforcements for the phonics forces arrived in the form of a meta-analysis regarding the effects of systematic phonics instruction compared to nonphonics instruction on learning to read (Ehri, Nunes, Stahl, & Willows, 2001). Researchers concluded, based on 66 treatment-control comparisons derived from 38 experiments and quasi-experiments that "phonics instruction proved effective and should be implemented as part of literacy programs to teach beginning reading as well as to prevent and remediate reading difficulties" (p. 393). How can teachers and principals develop a balanced literacy program, if reading researchers can't agree? *Is* phonics necessary for every child? What are the implications of the phonics research for instruction in your school?

CASE STUDY

Vouchers: Are Private Schools in the Public Good?

For most of the 1990s, the merits of private-school vouchers were debated by politicians and on editorial pages. A voucher is simply a coupon that is redeemable for tuition at private schools. Some proposals would make vouchers available to all children; in others, they would be limited to disadvantaged children. In some, private schools that accept vouchers would be subject to state regulations about curriculum and personnel; in others, private schools would be subject to minimal regulation.[9]

In its various permutations, the concept of vouchers has existed for quite some time. Thomas Paine (as cited in West, 1967) suggested a kind of voucher plan in *The Rights of Man*, and Nobel-Prize-winning economist Milton Friedman (1955) made a forceful case for vouchers in the 1950s. The recent furor over vouchers and similar reforms was spurred by the polemical writings of John Chubb and Terry Moe (1990), who argued that America's public schools are effectively "broken," their effectiveness hamstrung by a dysfunctional bureaucracy.

For all the debate surrounding the "ideal" voucher plan, the implementation of vouchers has been limited to a few high-profile cases. For many years parents in Maine and Vermont have been eligible to receive state-funded vouchers, but only in areas without sufficient public school coverage. In the 1990s, the cities of Milwaukee and Cleveland began distributing vouchers to low-income children. A much-heralded statewide program in Florida offers vouchers to students in public schools that are judged to be persistently failing. An even larger number of programs in New York, Washington, DC, and elsewhere offer privately funded vouchers—essentially scholarships—to low-income children in public schools.[10] In some cases, like California and Michigan in the fall elections of 2000, referenda on vouchers have been rejected by state voters.

No matter what a voucher program looks like, it typically rests on an important assumption: private schools are more effective than public schools at improving student outcomes. Or to phrase it in the manner of this book, attending a private school (instead of a public school) *causes* student outcomes to improve. Since the early 1980s, researchers have produced reams of evidence that purports to test that assertion.

Perhaps not surprisingly, they have a hard time agreeing, even when they are assessing the same program. For example, John Witte, the state-appointed evaluator of Milwaukee's voucher program found "no consistent differences" in test scores of voucher recipients and a comparison group (Witte, 1998, p. 241).

(Continued)

(Continued)

Harvard professor Paul Peterson and his colleagues (as quoted in Olson, 1996) countered that Witte's work is "so methodologically flawed as to be worthless." Their own work, conducted after Witte's, finds uniformly positive results (Greene, Peterson, & Du, 1998). However, a third study by Princeton professor Cecilia Rouse found yet another set of results that appeared to split the difference. She found that voucher recipients in Milwaukee experienced gains in math achievement but not in reading (Rouse, 1998). Similar controversies have dogged the voucher debate in other contexts, including evaluations in Cleveland, Florida, New York City, and elsewhere.

A cynic might chalk it all up to politics; we have a different view. At the core of researchers' disagreements, there usually lies a clear disagreement—over research methods and data. It is hard to know which side to believe unless one possesses an appreciation of these issues. It is even harder to decide if research might have something useful to say about one's own school, district, or state. In subsequent chapters, we will explore the multiple layers of the research debate over private-school effectiveness, and assess how its findings can be used to answer the questions posed in the previous sections.

CASE STUDY

Whole-School Reform: Greater Than the Sum of Its Parts?

Traditionally, efforts to improve schools have proceeded in a piecemeal fashion. Once problems were identified, they were addressed with a bewildering and not always coherent array of "solutions" that have waxed and waned in their popularity. These have ranged from block scheduling to cooperative learning to computer-assisted instruction. In the 1980s, a new approach began to take hold. Partisans of "whole-school" reform suggested the only means of obtaining large and sustained improvements in schools was to drastically change the way schools operated, rather than simply grafting policies onto existing institutions (Levin, 2002).

The Northwest Regional Educational Laboratory (2001) has produced a catalog that lists more than 28 models for whole-school reform. Some of the most recognized include James Comer's Comer School Development Program (2002), the Success for All approach developed by Robert Slavin (Slavin & Madden, 2001), and the Accelerated Schools Project of Henry Levin (National Center for Accelerated Schools, 2002a). Although they all qualify as "whole-school" reforms, they sometimes recommend vastly different strategies and goals (Levin, 2002). For example, the School Development Project emphasizes the improvement of the interpersonal relationships and social climate in a school. Success for All relies upon a prescribed curriculum to improve early-elementary school reading proficiency. The Accelerated Schools Project relies upon a constructivist learning theory in which schools have great latitude to determine their strategies for improvement.

Whole-school reform initiatives received some important boosts in the 1990s. A nonprofit organization called New American Schools provided funding to some whole-school reform models. Some school districts, such as Memphis, required schools to choose among the varied reform packages (although Memphis abandoned its initiative in 2001). And since 1997 the federal government has provided millions of dollars to support "comprehensive school reform."

In light of so much agreement on the need for whole-school reform, one might presume that a strong research base exists that can aid in discerning which models are most effective. In fact, the amount of high-quality research is not large, especially compared with the surfeit of program models and the long line of willing sponsors of whole-school reform. There are only a few randomized experiments, all conducted on Comer's School Development Project (Cook, Habib, Phillips, Settersten, Shagle, & Degirmencioglu, 1999;

(Continued)

(Continued)

Cook, Murphy, & Hunt, 2000). There is a voluminous amount of quasi-experimental research conducted on the Success for All program, and a growing research base on other projects such as Accelerated Schools and the models included in New American Schools (Barnett, 1996; Bloom et al., 2001; Levin, 2002). In subsequent chapters, we will place some of these studies under the microscope, and assess whether they can provide answers to the vital questions delineated in this chapter.

FURTHER READING

If you are interested in pursuing any of the case studies independently, the Resource section provides a complete listing of the pertinent references as well as Web sites in Chapters 3 through 7, grouped by topic.

NOTES

1. Our assumptions may well evoke strong reactions from two types of readers. Many quantitative researchers will take exception to our statement that qualitative research *can* test hypotheses and play a role in uncovering or confirming causal relationships. They believe that qualitative research by its nature is incapable of hypothesis testing and that uncovering causal relationships between interventions and outcomes is the sole prerogative of quantitative research. On the other hand, many qualitative researchers will find all of our assumptions to be suspect. We refer readers to King et al., (1994, pp. 3-9) and Shadish et al. (2002, pp. 478-484) for discussions of these issues from both qualitative and quantitative perspectives.

2. We caution the reader that although qualitative research *can* test causal relationships, it is only when the methods and logic of qualitative research are made explicit and when qualitative data are collected very systematically, that causal inferences (albeit imperfect ones) can be made based on observational data.

3. A common convention is to refer to "control groups" when assignment is randomized and "comparison groups" when it is not.

4. For an overview of current state initiatives, see the following class size advocacy Web site: http://www.reduceclasssizenow.org/state_of_the_states.htm.

5. Later chapters will review these findings in greater detail. Much of the recent debate over the Tennessee STAR experiment is contained in the Summer 1999 issue of the *Journal of Educational Evaluation and Policy Analysis.*

6. Jeanne Chall and Marie Carbo debated phonics in the pages of *Phi Delta Kappan* in the late 1980s (Carbo, 1988; Chall, 1989). Carbo, however, later backed off her antiphonics position (1996). Gerald Coles and Reid Lyon took up the debate in the late 1990s in *Education Week* (Coles, 1997; Lyon, 1997).

7. For a complete discussion of this debate, see McEwan (2002).

8. See the discussion at http://cars.uth.tmc.edu/debate.htm.

9. For a review of different approaches, see Levin (1991).

10. For an early review of these programs, see Moe (1995).