cal glaciation periods. He based the hypothesis on astronomical observations showing that the regions above and below the ecliptic are laden with cosmic dust, which would cool the planet. Muller’s “inclination theory” received major support when Kenneth Farley (1995) published a paper on cosmic dust in sea sediments.

Farley had begun his research project in an effort to refute the Muller inclination model, but discovered—to his surprise—that cosmic dust levels did indeed wax and wane in sync with the ice ages. As an immediate cause of the temperature change, Muller proposed that dust from space would influence the cloud cover on Earth and the amount of greenhouse gases—mainly carbon dioxide—in the atmosphere. Indeed, measurements of oxygen isotopes in trapped air bubbles and other properties from a 400,000-year-long Antarctic ice core by paleo-oceanographer Nicholas Shackleton (2001) provided more confirming evidence.

To gain greater understanding of these processes, geochronologists are seeking new “clocks” to determine more accurately the timing of events in the Earth’s history (e.g., Feng and Vasconcelos, 2001), while geochemists look for new ways of inferring temperature from composition of gasses trapped deep in ice or rock (see Pope and Giles, 2001). Still, no one knows how orbital variations would send the carbon dioxide into and out of the atmosphere. And there are likely to be other significant geologic factors besides carbon dioxide that control climate. There is much work still to be done to sort out the complex variables that are probably responsible for the ice ages.

culture, socioeconomic status, and poverty are classical models in anthropology, sociology, and political science. In program evaluation, program developers have ideas about the mechanism by which program inputs affect targeted outcomes; evaluations translate and test these ideas through a “program theory” that guides the work (Weiss, 1998a).

Theory enters the research process in two important ways. First, scientific research may be guided by a conceptual framework, model, or theory
that suggests possible questions to ask or answers to the question posed.\textsuperscript{4} The process of posing significant questions typically occurs before a study is conducted. Researchers seek to test whether a theory holds up under certain circumstances. Here the link between question and theory is straightforward. For example, Putnam based his work on a theoretical conception of institutional performance that related civic engagement and modernization.

A research question can also devolve from a practical problem (Stokes, 1997; see discussion above). In this case, addressing a complex problem like the relationship between class size and student achievement may require several theories. Different theories may give conflicting predictions about the problem's solution, or various theories might have to be reconciled to address the problem. Indeed, the findings from the Tennessee class size reduction study (see Box 3-3) have led to several efforts to devise theoretical understandings of how class size reduction may lead to better student achievement. Scientists are developing models to understand differences in classroom behavior between large and small classes that may ultimately explain and predict changes in achievement (Grissmer and Flanagan, 2000).

A second more subtle way that theoretical understanding factors into the research process derives from the fact that all scientific observations are “theory laden” (Kuhn, 1962). That is, the choice of what to observe and how to observe it is driven by an organizing conception—explicit or tacit—of the problem or topic. Thus, theory drives the research question, the use of methods, and the interpretation of results.

**SCIENTIFIC PRINCIPLE 3**

**Use Methods That Permit Direct Investigation of the Question**

Research methods—the design for collecting data and the measurement and analysis of variables in the design—should be selected in light of a research question, and should address it directly. Methods linked directly to problems permit the development of a logical chain of reasoning based

\textsuperscript{4}The process of posing significant questions or hypotheses may occur, as well, at the end of a study (e.g., Agar, 1996), or over the course of an investigation as understanding of the facets of the problem evolves (e.g., Brown, 1992).
on the interplay among investigative techniques, data, and hypotheses to reach justifiable conclusions. For clarity of discussion, we separate out the link between question and method (see Principle 3) and the rigorous reasoning from evidence to theory (see Principle 4). In the actual practice of research, such a separation cannot be achieved.

Debates about method—in many disciplines and fields—have raged for centuries as researchers have battled over the relative merit of the various techniques of their trade. The simple truth is that the method used to conduct scientific research must fit the question posed, and the investigator must competently implement the method. Particular methods are better suited to address some questions rather than others. The rare choice in the mid 1980s in Tennessee to conduct a randomized field trial, for example, enabled stronger inferences about the effects of class size reduction on student achievement (see Box 3-3) than would have been possible with other methods.

This link between question and method must be clearly explicated and justified; a researcher should indicate how a particular method will enable competent investigation of the question of interest. Moreover, a detailed description of method—measurements, data collection procedures, and data analyses—must be available to permit others to critique or replicate the study (see Principle 5). Finally, investigators should identify potential methodological limitations (such as insensitivity to potentially important variables, missing data, and potential researcher bias).

The choice of method is not always straightforward because, across all disciplines and fields, a wide range of legitimate methods—both quantitative and qualitative—are available to the researcher. For example when considering questions about the natural universe—from atoms to cells to black holes—profoundly different methods and approaches characterize each sub-field. While investigations in the natural sciences are often dependent on the use of highly sophisticated instrumentation (e.g., particle accelerators, gene sequencers, scanning tunneling microscopes), more rudimentary methods often enable significant scientific breakthroughs. For example, in 1995 two Danish zoologists identified an entirely new phylum of animals from a species of tiny rotifer-like creatures found living on the mouthparts of lobsters, using only a hand lens and light microscope (Wilson, 1998, p. 63).
BOX 3-3

Does Reducing Class Size Improve Students' Achievement?

Although research on the effects of class size reduction on students' achievement dates back 100 years, Glass and Smith (1978) reported the first comprehensive statistical synthesis (meta-analysis) of the literature and concluded that, indeed, there were small improvements in achievement when class size was reduced (see also Glass, Cahen, Smith, and Filby, 1982; Bohrnstedt and Stecher, 1999). However, the Glass and Smith study was criticized (e.g., Robinson and Wittebol, 1986; Slavin, 1989) on a number of grounds, including the selection of some of the studies for the meta-analysis (e.g., tutoring, college classes, atypically small classes). Some subsequent reviews reached conclusions similar to Glass and Smith (e.g., Bohrnstedt and Stecher, 1999; Hedges, Laine, and Greenwald, 1994; Robinson and Wittebol, 1986) while others did not find consistent evidence of a positive effect (e.g., Hanushek, 1986, 1999a; Odden, 1990; Slavin, 1989).

Does reducing class size improve students' achievement? In the midst of controversy, the Tennessee state legislature asked just this question and funded a randomized experiment to find out, an experiment that Harvard statistician Frederick Mosteller (1995, p. 113) called “... one of the most important educational investigations ever carried out.” A total of 11,600 elementary school students and their teachers in 79 schools across the state were randomly assigned to one of three class-size conditions: small class (13-17 students), regular class

If a research conjecture or hypothesis can withstand scrutiny by multiple methods its credibility is enhanced greatly. As Webb, Campbell, Schwartz, and Sechrest (1966, pp. 173-174) phrased it: “When a hypothesis can survive the confrontation of a series of complementary methods of testing, it contains a degree of validity unattainable by one tested within the more constricted framework of a single method.” Putnam’s study (see Box 3-1) provides an example in which both quantitative and qualitative methods were applied in a longitudinal design (e.g., interview, survey, statistical estimate of institutional performance, analysis of legislative docu-
(22-26 students), or regular class (22-26 students) with a full-time teacher's aide (for descriptions of the experiment, see Achilles, 1999; Finn and Achilles, 1990; Folger and Breda, 1989; Krueger, 1999; Word et al., 1990). The experiment began with a cohort of students who entered kindergarten in 1985, and lasted 4 years. After third grade, all students returned to regular size classes. Although students were supposed to stay in their original treatment conditions for four years, not all did. Some were randomly reassigned between regular and regular/aide conditions in the first grade while about 10 percent switched between conditions for other reasons (Krueger and Whitmore, 2000).

Three findings from this experiment stand out. First, students in small classes outperformed students in regular size classes (with or without aides). Second, the benefits of class-size reduction were much greater for minorities (primarily African American) and inner-city children than others (see, e.g., Finn and Achilles, 1990, 1999; but see also Hanushek, 1999b). And third, even though students returned to regular classes in fourth grade, the reduced class-size effect persisted in affecting whether they took college entrance examinations and on their examination performance (Krueger and Whitmore, 2001).*

*Interestingly, in balancing the size of the effects of class size reduction with the costs, the Tennessee legislature decided not to reduce class size in the state (Ritter and Boruch, 1999).

ments) to generate converging evidence about the effects of modernization on civic community. New theories about the periodicity of the ice ages, similarly, were informed by multiple methods (e.g., astronomical observations of cosmic dust, measurements of oxygen isotopes). The integration and interaction of multiple disciplinary perspectives—with their varying methods—often accounts for scientific progress (Wilson, 1998); this is evident, for example, in the advances in understanding early reading skills described in Chapter 2. This line of work features methods that range from neuroimaging to qualitative classroom observation.
We close our discussion of this principle by noting that in many sciences, measurement is a key aspect of research method. This is true for many research endeavors in the social sciences and education research, although not for all of them. If the concepts or variables are poorly specified or inadequately measured, even the best methods will not be able to support strong scientific inferences. The history of the natural sciences is one of remarkable development of concepts and variables, as well as the tools (instrumentation) to measure them. Measurement reliability and validity is particularly challenging in the social sciences and education (Messick, 1989). Sometimes theory is not strong enough to permit clear specification and justification of the concept or variable. Sometimes the tool (e.g., multiple-choice test) used to take the measurement seriously under-represents the construct (e.g., science achievement) to be measured. Sometimes the use of the measurement has an unintended social consequence (e.g., the effect of teaching to the test on the scope of the curriculum in schools).

And sometimes error is an inevitable part of the measurement process. In the physical sciences, many phenomena can be directly observed or have highly predictable properties; measurement error is often minimal. (However, see National Research Council [1991] for a discussion of when and how measurement in the physical sciences can be imprecise.) In sciences that involve the study of humans, it is essential to identify those aspects of measurement error that attenuate the estimation of the relationships of interest (e.g., Shavelson, Baxter, and Gao, 1993). By investigating those aspects of a social measurement that give rise to measurement error, the measurement process itself will often be improved. Regardless of field of study, scientific measurements should be accompanied by estimates of uncertainty whenever possible (see Principle 4 below).

**SCIENTIFIC PRINCIPLE 4**

**Provide Coherent, Explicit Chain of Reasoning**

The extent to which the inferences that are made in the course of scientific work are warranted depends on rigorous reasoning that systematically and logically links empirical observations with the underlying theory and the degree to which both the theory and the observations are linked to the question or problem that lies at the root of the investigation. There
is no recipe for determining how these ingredients should be combined; instead, what is required is the development of a logical “chain of reasoning” (Lesh, Lovitts, and Kelly, 2000) that moves from evidence to theory and back again. This chain of reasoning must be coherent, explicit (one that another researcher could replicate), and persuasive to a skeptical reader (so that, for example, counterhypotheses are addressed).

All rigorous research—quantitative and qualitative—embodies the same underlying logic of inference (King, Keohane, and Verba, 1994). This inferential reasoning is supported by clear statements about how the research conclusions were reached: What assumptions were made? How was evidence judged to be relevant? How were alternative explanations considered or discarded? How were the links between data and the conceptual or theoretical framework made?

The nature of this chain of reasoning will vary depending on the design of the study, which in turn will vary depending on the question that is being investigated. Will the research develop, extend, modify, or test a hypothesis? Does it aim to determine: What works? How does it work? Under what circumstances does it work? If the goal of the research is to test a hypothesis, stated in the form of an “if-then” rule, successful inference may depend on measuring the extent to which the rule predicts results under a variety of conditions. If the goal is to produce a description of a complex system, such as a subcellular organelle or a hierarchical social organization, successful inference may rather depend on issues of fidelity and internal consistency of the observational techniques applied to diverse components and the credibility of the evidence gathered. The research design and the inferential reasoning it enables must demonstrate a thorough understanding of the subtleties of the questions to be asked and the procedures used to answer them.

Muller (1994), for example, collected data on the inclination of the Earth's orbit over a 100,000 year cycle, correlated it with the occurrence of ice ages, ruled out the plausibility of orbital eccentricity as a cause for the occurrence of ice ages, and inferred that the bounce in the Earth's orbit likely caused the ice ages (see Box 3-2). Putnam used multiple methods to subject to rigorous testing his hypotheses about what affects the success or failure of democratic institutions as they develop in diverse social environments to rigorous testing, and found the weight of the evidence favored
the assertion that civic tradition matters more than economic affluence (see Box 3-1). And Baumeister, Bratslavsky, Muraven, and Tice (1998) compared three competing theories and used randomized experiments to conclude that a “psychic energy” hypothesis best explained the important psychological characteristic of “will power” (see “Application of the Principles”).

This principle has several features worthy of elaboration. Assumptions underlying the inferences made should be clearly stated and justified. Moreover, choice of design should both acknowledge potential biases and plan for implementation challenges.

Estimates of error must also be made. Claims to knowledge vary substantially according to the strength of the research design, theory, and control of extraneous variables and by systematically ruling out possible alternative explanations. Although scientists always reason in the presence of uncertainty, it is critical to gauge the magnitude of this uncertainty. In the physical and life sciences, quantitative estimates of the error associated with conclusions are often computed and reported. In the social sciences and education, such quantitative measures are sometimes difficult to generate; in any case, a statement about the nature and estimated magnitude of error must be made in order to signal the level of certainty with which conclusions have been drawn.

Perhaps most importantly, the reasoning about evidence should identify, consider, and incorporate, when appropriate, the alternative, competing explanations or rival “answers” to the research question. To make valid inferences, plausible counterexplanations must be dealt with in a rational, systematic, and compelling way. The validity—or credibility—of a hypothesis is substantially strengthened if alternative counterhypotheses can be ruled out and the favored one thereby supported. Well-known research designs (e.g., Campbell and Stanley [1963] in educational psychology; Heckman [1979, 1980a, 1980b, 2001] and Goldberger [1972, 1983] in

---

5In reporting, too, it is important to clarify that rival hypotheses are possible and that conclusions are not presented as if they were gospel. Murphy and colleagues call this “fair-dealing”—wariness of presenting the perspective of one group as if it defined a single truth about the phenomenon, while paying scant attention to other perspectives” (Murphy, Dingwall, Greatbatch, Parker, and Watson, 1998, p. 192).
economics; and Rosenbaum and Rubin [1983, 1984] in statistics) have been crafted to guard researchers against specific counterhypothenses (or “threats to validity”). One example, often called “selectivity bias,” is the counterhypothesis that differential selection (not the treatment) caused the outcome—that participants in the experimental treatment systematically differed from participants in the traditional (control) condition in ways that mattered importantly to the outcome. A cell biologist, for example, might unintentionally place (select) heart cells with a slight glimmer into an experimental group and others into a control group, thus potentially biasing the comparison between the groups of cells. The potential for a biased—or unfair—comparison arises because the shiny cells could differ systematically from the others in ways that affect what is being studied.

Selection bias is a pervasive problem in the social sciences and education research. To illustrate, in studying the effects of class-size reduction, credentialed teachers are more likely to be found in wealthy school districts that have the resources to reduce class size than in poor districts. This fact raises the possibility that higher achievement will be observed in the smaller classes due to factors other than class size (e.g., teacher effects). Random assignment to “treatment” is the strongest known antidote to the problem of selection bias (see Chapter 5).

A second counterhypothesis contends that something in the research participants’ history that co-occurred with the treatment caused the outcome, not the treatment itself. For example, U.S. fourth-grade students outperformed students in other countries on the ecology subtest of the Third International Mathematics and Science Study. One (popular) explanation of this finding was that the effect was due to their schooling and the emphasis on ecology in U.S. elementary science curricula. A counterhypothesis, one of history, posits that their high achievement was due to the prevalence of ecology in children’s television programming. A control group that has the same experiences as the experimental group except for the “treatment” under study is the best antidote for this problem.

A third prevalent class of alternative interpretations contends that an outcome was biased by the measurement used. For example, education effects are often judged by narrowly defined achievement tests that focus on factual knowledge and therefore favor direct-instruction teaching tech-
niques. Multiple achievement measures with high reliability (consistency) and validity (accuracy) help to counter potential measurement bias.

The Tennessee class-size study was designed primarily to eliminate all possible known explanations, except for reduced class size, in comparing the achievement of children in regular classrooms against achievement in reduced size classrooms. It did this. Complications remained, however. About ten percent of students moved out of their originally assigned condition (class size), weakening the design because the comparative groups did not remain intact to enable strict comparisons. However, most scholars who subsequently analyzed the data (e.g., Krueger and Whitmore, 2001), while limited by the original study design, suggested that these infidelities did not affect the main conclusions of the study that smaller class size caused slight improvements in achievement. Students in classes of 13–17 students outperformed their peers in larger classes, on average, by a small margin.

**SCIENTIFIC PRINCIPLE 5**

**Replicate and Generalize Across Studies**

Replication and generalization strengthen and clarify the limits of scientific conjectures and theories. By replication we mean, at an elementary level, that if one investigator makes a set of observations, another investigator can make a similar set of observations under the same conditions. Replication in this sense comes close to what psychometricians call reliability—consistency of measurements from one observer to another, from one task to another parallel task, from one occasion to another occasion. Estimates of these different types of reliability can vary when measuring a given construct: for example, in measuring performance of military personnel (National Research Council, 1991), multiple observers largely agreed on what they observed within tasks; however, enlistees’ performance across parallel tasks was quite inconsistent.

At a somewhat more complex level, replication means the ability to repeat an investigation in more than one setting (from one laboratory to another or from one field site to a similar field site) and reach similar conclusions. To be sure, replication in the physical sciences, especially with inanimate objects, is more easily achieved than in social science or education; put another way, the margin of error in social science replication is usually
much greater than in physical science replication. The role of contextual factors and the lack of control that characterizes work in the social realm require a more nuanced notion of replication. Nevertheless, the typically large margins of error in social science replications do not preclude their identification.

Having evidence of replication, an important goal of science is to understand the extent to which findings generalize from one object or person to another, from one setting to another, and so on. To this end, a substantial amount of statistical machinery has been built both to help ensure that what is observed in a particular study is representative of what is of larger interest (i.e., will generalize) and to provide a quantitative measure of the possible error in generalizing. Nonstatistical means of generalization (e.g., triangulation, analytic induction, comparative analysis) have also been developed and applied in genres of research, such as ethnography, to understand the extent to which findings generalize across time, space, and populations. Subsequent applications, implementations, or trials are often necessary to assure generalizability or to clarify its limits. For example, since the Tennessee experiment, additional studies of the effects of class size reduction on student learning have been launched in settings other than Tennessee to assess the extent to which the findings generalize (e.g., Hruz, 2000).

In the social sciences and education, many generalizations are limited to particular times and particular places (Cronbach, 1975). This is because the social world undergoes rapid and often significant change; social generalizations, as Cronbach put it, have a shorter “half-life” than those in the physical world. Campbell and Stanley (1963) dubbed the extent to which the treatment conditions and participant population of a study mirror the world to which generalization is desired the “external validity” of the study. Consider, again, the Tennessee class-size research; it was undertaken in a set of schools that had the desire to participate, the physical facilities to accommodate an increased number of classrooms, and adequate teaching staff. Governor Wilson of California “overgeneralized” the Tennessee study, ignoring the specific experimental conditions of will and capacity and implemented class-size reduction in more than 95 percent of grades K-3 in the state. Not surprisingly, most researchers studying California have
concluded that the Tennessee findings did not entirely generalize to a different time, place, and context (see, e.g., Stecher and Bohrnstedt, 2000).  

**SCIENTIFIC PRINCIPLE 6**  
**Disclose Research to Encourage Professional Scrutiny and Critique**

We argue in Chapter 2 that a characteristic of scientific knowledge accumulation is its contested nature. Here we suggest that science is not only characterized by professional scrutiny and criticism, but also that such criticism is essential to scientific progress. Scientific studies usually are elements of a larger corpus of work; furthermore, the scientists carrying out a particular study always are part of a larger community of scholars. Reporting and reviewing research results are essential to enable wide and meaningful peer review. Results are traditionally published in a specialty journal, in books published by academic presses, or in other peer-reviewed publications. In recent years, an electronic version may accompany or even substitute for a print publication. Results may be debated at professional conferences. Regardless of the medium, the goals of research reporting are to communicate the findings from the investigation; to open the study to examination, criticism, review, and replication (see Principle 5) by peer investigators; and ultimately to incorporate the new knowledge into the prevailing canon of the field.

---

6A question arises as to whether this is a failure to generalize or a problem of poor implementation. The conditions under which Tennessee implemented the experiment were not reproduced in California with the now known consequence of failure to replicate and generalize.

7The committee is concerned that the quality of peer review in electronic modes of dissemination varies greatly and sometimes cannot be easily assessed from its source. While the Internet is providing new and exciting ways to connect scientists and promote scientific debate, the extent to which the principles of science are met in some electronically posted work is often unclear.

8Social scientists and education researchers also commonly publish information about new knowledge for practitioners and the public. In those cases, the research must be reported in accessible ways so that readers can understand the researcher's procedures and evaluate the evidence, interpretations, and arguments.
The goal of communicating new knowledge is self-evident: research results must be brought into the professional and public domain if they are to be understood, debated, and eventually become known to those who could fruitfully use them. The extent to which new work can be reviewed and challenged by professional peers depends critically on accurate, comprehensive, and accessible records of data, method, and inferential reasoning. This careful accounting not only makes transparent the reasoning that led to conclusions—promoting its credibility—but it also allows the community of scientists and analysts to comprehend, to replicate, and otherwise to inform theory, research, and practice in that area.

Many nonscientists who seek guidance from the research community bemoan what can easily be perceived as bickering or as an indication of “bad” science. Quite the contrary: intellectual debate at professional meetings, through research collaborations, and in other settings provide the means by which scientific knowledge is refined and accepted; scientists strive for an “open society” where criticism and unfettered debate point the way to advancement. Through scholarly critique (see, e.g., Skocpol, 1996) and debate, for example, Putnam’s work has stimulated a series of articles, commentary, and controversy in research and policy circles about the role of “social capital” in political and other social phenomena (Winter, 2000). And the Tennessee class size study has been the subject of much scholarly debate, leading to a number of follow-on analyses and launching new work that attempts to understand the process by which classroom behavior may shift in small classes to facilitate learning. However, as Lagemann (2000) has observed, for many reasons the education research community has not been nearly as critical of itself as is the case in other fields of scientific study.

APPLICATION OF THE PRINCIPLES

The committee considered a wide range of literature and scholarship to test its ideas about the guiding principles. We realized, for example, that empiricism, while a hallmark of science, does not uniquely define it. A poet can write from first-hand experience of the world, and in this sense is an empiricist. And making observations of the world, and reasoning about their experience, helps both literary critics and historians create the
interpretive frameworks that they bring to bear in their scholarship. But empirical method in scientific inquiry has different features, like codified procedures for making observations and recognizing sources of bias associated with particular methods, and the data derived from these observations are used specifically as tools to support or refute knowledge claims. Finally, empiricism in science involves collective judgments based on logic, experience, and consensus.

Another hallmark of science is replication and generalization. Humanists do not seek replication, although they often attempt to create work that generalizes (say) to the “human condition.” However, they have no formal logic of generalization, unlike scientists working in some domains (e.g., statistical sampling theory). In sum, it is clear that there is no bright line that distinguishes science from nonscience or high-quality science from low-quality science. Rather, our principles can be used as general guidelines for understanding what can be considered scientific and what can be considered high-quality science (see, however, Chapters 4 and 5 for an elaboration).

To show how our principles help differentiate science from other forms of scholarship, we briefly consider two genres of education inquiry published in refereed journals and books. We do not make a judgment about the worth of either form of inquiry; although we believe strongly in the merits of scientific inquiry in education research and more generally, that “science” does not mean “good.” Rather, we use them as examples to illustrate the distinguishing character of our principles of science. The first—connoisseurship—grew out of the arts and humanities (e.g., Eisner, 1991) and does not claim to be scientific. The second—portraiture—claims to straddle the fence between humanistic and scientific inquiry (e.g., Lawrence-Lightfoot and Davis, 1997).

Eisner (1991, p. 7) built a method for education inquiry firmly rooted in the arts and humanities, arguing that “there are multiple ways in which the world can be known: Artists, writers, and dancers, as well as scientists, have important things to tell about the world.” His method of inquiry combines connoisseurship (the art of appreciation), which “aims to

---

We do not claim that any one investigator or observational method is “objective.” Rather, the guiding principles are established to guard against bias through rigorous methods and a critical community.
appreciate the qualities . . . that constitute an act, work, or object and, typically . . . to relate these to the contextual and antecedent conditions” (p. 85) with educational criticism (the art of disclosure), which provides “connoisseurship with a public face” (p. 85). The goal of this genre of research is to enable readers to enter an event and to participate in it. To this end, the educational critic—through educational connoisseurship—must capture the key qualities of the material, situation, and experience and express them in text (“criticism”) to make what the critic sees clear to others. “To know what schools are like, their strengths and their weaknesses, we need to be able to see what occurs in them, and we need to be able to tell others what we have seen in ways that are vivid and insightful” (Eisner, 1991, p. 23, italics in original).

The grounds for his knowledge claims are not those in our guiding principles. Rather, credibility is established by: (1) structural corroboration—“multiple types of data are related to each other” (p. 110) and “disconfirming evidence and contradictory interpretations” (p. 111; italics in original) are considered; (2) consensual validation—“agreement among competent others that the description, interpretation, evaluation, and thematics of an educational situation are right” (p. 112); and (3) referential adequacy—“the extent to which a reader is able to locate in its subject matter the qualities the critic addresses and the meanings he or she ascribes to these” (p. 114). While sharing some features of our guiding principles (e.g., ruling out counterinterpretations to the favored interpretation), this humanistic approach to knowledge claims builds on a very different epistemology; the key scientific concepts of reliability, replication, and generalization, for example, are quite different. We agree with Eisner that such approaches fall outside the purview of science and conclude that our guiding principles readily distinguish them.

Portraiture (Lawrence-Lightfoot, 1994; Lawrence-Lightfoot and Davis, 1997) is a qualitative research method that aims to “record and interpret the perspectives and experience of the people they [the researchers] are studying, documenting their [the research participants’] voices and their visions—their authority, knowledge, and wisdom” (Lawrence-Lightfoot and Davis, 1997, p. xv). In contrast to connoisseurship’s humanist orientation, portraiture “seeks to join science and art” (Lawrence-Lightfoot and Davis, 1997, p. xv) by “embracing the intersection of aesthetics and empiricism” (p. 6). The standard for judging the quality of portraiture is authenticity,
“... capturing the essence and resonance of the actors’ experience and perspective through the details of action and thought revealed in context” (p. 12). When empirical and literary themes come together (called “resonance”) for the researcher, the actors, and the audience, “we speak of the portrait as achieving authenticity” (p. 260).

In *I've Known Rivers*, Lawrence-Lightfoot (1994) explored the life stories of six men and women:

... using the intensive, probing method of “human archeology”—a name I [Lawrence-Lightfoot] coined for this genre of portraiture as a way of trying to convey the depth and penetration of the inquiry, the richness of the layers of human experience, the search for ancestral and generational artifacts, and the painstaking, careful labor that the metaphorical dig requires. As I listen to the life stories of these individuals and participate in the ‘co-construction’ of narrative, I employ the themes, goals, and techniques of portraiture. It is an eclectic, interdisciplinary approach, shaped by the lenses of history, anthropology, psychology and sociology. I blend the curiosity and detective work of a biographer, the literary aesthetic of a novelist, and the systematic scrutiny of a researcher (p. 15).

Some scholars, then, deem portraiture as “scientific” because it relies on the use of social science theory and a form of empiricism (e.g., interview). While both empiricism and theory are important elements of our guiding principles, as we discuss above, they are not, in themselves, defining. The devil is in the details. For example, independent replication is an important principle in our framework but is absent in portraiture in which researcher and subject jointly construct a narrative. Moreover, even when our principles are manifest, the specific form and mode of application can make a big difference. For example, generalization in our principles is different from generalization in portraiture. As Lawrence-Lightfoot and Davis (1997) point out, generalization as used in the social sciences does not fit portraiture. Generalization in portraiture “... is not the classical conception ... where the investigator uses codified methods for generalizing from specific findings to a universe, and where there is little interest in findings that reflect only the characteristics of the sample. ...” By contrast, the portraitist seeks to “document and illuminate the complexity
and detail of a unique experience or place, hoping the audience will see itself reflected in it, trusting that the readers will feel identified. The portraitist is very interested in the single case because she believes that embedded in it the reader will discover resonant universal themes” (p. 15). We conclude that our guiding principles would distinguish portraiture from what we mean by scientific inquiry, although it, like connoisseurship, has some traits in common.

To this point, we have shown how our principles help to distinguish science and nonscience. A large amount of education research attempts to base knowledge claims on science; clearly, however, there is great variation with respect to scientific rigor and competence. Here we use two studies to illustrate how our principles demonstrate this gradation in scientific quality.

The first study (Carr, Levin, McConnachie, Carlson, Kemp, Smith, and McLaughlin, 1999) reported on an educational intervention carried out on three nonrandomly selected individuals who were suffering severe behavioral disorders and who were either residing in group-home settings or with their parents. Since earlier work had established remedial procedures involving “simulations and analogs of the natural environment” (p. 6), the focus of the study was on the generalizability (or external validity) to the “real world” of the intervention (places, caregivers).

Using a multiple baseline design, baseline frequencies of their problem behaviors were established, and these behaviors were remeasured while an intervention lasting for some years was carried out. The researchers also took several measurements during the maintenance phase of the study. While care was taken in describing behavioral observations, variable construction and reliability, the paper reporting on the study did not provide clear, detailed depictions of the interventions or who carried them out (research staff, staff of the group homes, or family members). Furthermore, no details were given of the changes in staffing or in the regimens of the residential settings—changes that were inevitable over a period of many years (the timeline itself was not clearly described). Finally, in the course of daily life over a number of years, many things would have happened to each of the subjects, some of which might be expected to be of significance to the study, but none of them were documented. Over the years, too, one might expect some developmental changes to occur in the aggressive behavior displayed by the research subjects, especially in the two teen-
agers. In short, the study focused on generalizability at too great an expense relative to internal validity. In the end, there were many threats to internal validity in this study, and so it is impossible to conclude (as the authors did) from the published report that the "treatment" had actually caused the improvement in behavior that was noted.

Turning to a line of work that we regard as scientifically more successful, in a series of four randomized experiments, Baumeister, Bratslavsky, Muraven, and Tice (1998) tested three competing theories of "will power" (or, more technically, "self-regulation")—the psychological characteristic that is posited to be related to persistence with difficult tasks such as studying or working on homework assignments. One hypothesis was that will power is a developed skill that would remain roughly constant across repeated trials. The second theory posited a self-control schema "that makes use of information about how to alter one's own response" (p. 1254) so that once activated on one trial, it would be expected to increase will power on a second trial. The third theory, anticipated by Freud's notion of the ego exerting energy to control the id and superego, posits that will power is a depletable resource—it requires the use of "psychic energy" so that performance from trial 1 to trial 2 would decrease if a great deal of will power was called for on trial 1. In one experiment, 67 introductory psychology students were randomly assigned to a condition in which either no food was present or both radishes and freshly baked chocolate chip cookies were present, and the participants were instructed either to eat two or three radishes (resisting the cookies) or two or three cookies (resisting the radishes). Immediately following this situation, all participants were asked to work on two puzzles that unbeknownst to them, were unsolvable, and their persistence (time) in working on the puzzles was measured. The experimental manipulation was checked for every individual participating by researchers observing their behavior through a one-way window. The researchers found that puzzle persistence was the same in the control and cookie conditions and about 2.5 times as long, on average, as in the radish condition, lending support to the psychic energy theory—arguably, resisting the temptation to eat the cookies evidently had depleted the reserve of self-control, leading to poor performance on the second task. Later experiments extended the findings supporting the energy theory to situations involving choice, maladaptive performance, and decision making.
However, as we have said, no single study or series of studies satisfy all of our guiding principles, and these will power experiments are no exception. They all employed small samples of participants, all drawn from a college population. The experiments were contrived—the conditions of the study would be unlikely outside a psychology laboratory. And the question of whether these findings would generalize to more realistic (e.g., school) settings was not addressed.

Nevertheless, the contrast in quality between the two studies, when observed through the lens of our guiding principles, is stark. Unlike the first study, the second study was grounded in theory and identified three competing answers to the question of self-regulation, each leading to a different empirically refutable claim. In doing so, the chain of reasoning was made transparent. The second study, unlike the first, used randomized experiments to address counterclaims to the inference of psychic energy, such as selectivity bias or different history during experimental sessions. Finally, in the second study, the series of experiments replicated and extended the effects hypothesized by the energy theory.

CONCLUDING COMMENT

Nearly a century ago, John Dewey (1916) captured the essence of the account of science we have developed in this chapter and expressed a hopefulness for the promise of science we similarly embrace:

Our predilection for premature acceptance and assertion, our aversion to suspended judgment, are signs that we tend naturally to cut short the process of testing. We are satisfied with superficial and immediate short-visioned applications. If these work out with moderate satisfactoriness, we are content to suppose that our assumptions have been confirmed. Even in the case of failure, we are inclined to put the blame not on the inadequacy and incorrectness of our data and thoughts, but upon our hard luck and the hostility of circumstances. . . . Science represents the safeguard of the [human] race against these natural propensities and the evils which flow from them. It consists of the special appliances and methods... slowly worked out in order to conduct reflection under conditions whereby its procedures and results are tested.
In Chapter 3 the committee argues that the guiding principles for scientific research in education are the same as those in the social, physical, and life sciences. Yet the ways that those principles are instantiated—in astrophysics, biochemistry, labor economics, cultural anthropology, or mathematics teaching—depend on the specific features of what is being studied. That is, each field has features that influence what questions are asked, how research is designed, how it is carried out, and how it is interpreted and generalized. Scholars working in a particular area establish the traditions and standards for how to most appropriately apply the guiding principles to their area of study (Diamond, 1999).

In this chapter, we describe how our principles of science translate in the study of education—a rich tapestry of teaching, learning, and schooling. In particular, we briefly discuss five features of education that shape scientific inquiry, and describe how these features affect research. We argue that a key implication of these features of education is the need to account for influential contextual factors within the process of inquiry and in understanding the extent to which findings can be generalized. These features sharpen the conception of scientific research quality we develop in Chapter 3. We also discuss three features of education research that are essential to understanding the nature and conduct of the professional work.

To set the stage for our discussion of the particulars of scientific education research, we reiterate our position that there are substantial similarities between inquiry in the physical and social worlds. We have argued in
previous chapters that our principles of science are common across disciplines and fields and that the accumulation of knowledge progresses in roughly the same way. Furthermore, profoundly different methods and approaches characterize each discipline and field in the physical sciences, depending on such things as the time frame, the scale of magnitude, and the complexity of the instrumentation required. The same is true in the social sciences and education, where questions ranging from individual learning of varied subject matter to fundamental social patterns to cultural norms determine the length of time, the number of people, and the kind of research instruments that are needed in conducting the studies.

Differences in the phenomena typically under investigation do distinguish the research conducted by physical and social scientists. For example, the social and cultural work of sociologists and cultural anthropologists often do not lend themselves to the controlled conditions, randomized treatments, or repeated measures that typify investigations in physics or chemistry. Phenomena such as language socialization, deviancy, the development of an idea, or the interaction of cultural tradition with educational instruction are notoriously impervious to the controls used in the systematic investigations of atoms or molecules. Unlike atoms or molecules, people grow up and change over time. The social, cultural, and economic conditions they experience evolve with history. The abstract concepts and ideas that are meaningful to them vary across time, space, and cultural tradition. These circumstances have led some social science and education researchers to investigative approaches that look distinctly different from those of physical researchers, while still aligning with the guiding principles outlined in Chapter 3.

Another area that can notably distinguish research between the social and physical sciences concerns researcher objectivity in relation to bias. In some physical and life sciences, investigators are often deliberately kept ignorant of the identity of research participants, and controls are instituted through such devices as double-blind or randomization procedures. This strategy is often used in medical trials to ensure that researchers’ perspectives are not influenced by their knowledge of which participants received which treatment, and similarly, that this knowledge does not alter the behavior of the research participants. In many areas of the social sciences, in contrast, the investigator is recognized as an “engaged participant
observer," involved with the experience and action of those observed (Blumer, 1966; Denzin, 1978; Kelly and Lesh, 2000). In such "naturalistic research paradigms" (Moschkovich and Brenner, 2000), investigators do not seek to distance themselves from research participants, but rather to immerse themselves in the participants' lives, with conscious attention to how the investigator affects and contributes to the research process. Such strategies were developed to allow the researcher to observe, analyze, and integrate into the research process unexpected, constantly changing, and other potentially influential aspects of what is being studied. These approaches are often particularly important in studying how changes in school subject matter or the development of new technologies can be incorporated into educational practice. In collecting and coding such qualitative data, convergence can be demonstrated with repeated instances, more than one observer, and multiple raters. Also essential to the process is the examination of competing interpretations, contrasting cases, and disconfirming evidence. Regularity in the patterns across groups and across time—rather than replication per se—is a source of generalization. The goal of such scientific methods, of course, remains the same: to identify generalized patterns.

Uses of theory also tend to distinguish work in the social and physical sciences. Theory in the physical sciences leads to predictions about things that will happen in the future. Strong theories include causal mechanisms that predict what will happen and give insights into why. Theory in the social sciences is predictive, but more often it serves to understand things that happened in the past, serving a more diagnostic or explanatory purpose. Understanding the past often enables social science researchers to explain why things happened. Though understanding the past can sometimes predict the future, it does so only in broad outline and with a lesser degree of certainty. For instance, researchers have documented the regularity of certain misconceptions and patterns of error as students learn scientific or mathematical ideas. Although one cannot predict exactly when they will occur, awareness of them permits teachers to interpret student comments more effectively and to create assessment items to test for evidence of them.

A related and final point is that the level of certainty with which research conclusions can be made is typically higher in the physical sciences than in
the social sciences. As we discuss in Chapter 3, many scientific claims have some degree of uncertainty associated with them—that is, they are probabilistic rather than deterministic. We include within our principles the idea that careful estimation and reporting of uncertainty is crucial to science. However, because theories that model social phenomena—human behavior, ideas, cultures—are not as well developed as those for some physical phenomena and because they are often out of the direct control of the researcher, results are always probabilistic and tend to be more tentative than in the physical sciences. In technical terms, this means that the “error limits” associated with scientific inferences (not unlike confidence intervals typically cited in public opinion polls) tend to be larger in social and behavioral research, often due to the “noise” caused by difficulties precisely measuring key constructs and major contextual factors. The influential role of context in many social and behavioral research inquiries is a fundamental aspect of studying humans. However, it does make replication—the key to boosting certainty in results and refining theory—more difficult and nuanced. In sum, the degree of precision associated with current social science findings tends to be lower than that in the physical and life sciences.

Although education research has its roots in the social and behavioral sciences, it is also an applied field akin in important ways to medicine and agriculture. Some scholars have likened education research to the engineering sciences, arguing that it is an enterprise fundamentally aimed at bringing theoretical understanding to practical problem solving. Like other applied fields, education research serves two related purposes: to add to fundamental understanding of education-related phenomena and events, and to inform practical decision making. Both are worthy, both require researchers to have a keen understanding of educational practice and policy, and both can ultimately lead to improvements in practice. Education research with the sole aim of explaining, describing, or predicting closely resembles “traditional” scientific inquiry of the kind we describe in the previous chapter. Research whose direct aim is to aid educational practice, decision making, and policy in the near term also must meet scientific principles for rigor, but it has an action orientation. The dual purposes of education research suggest that there must be a balance of considerations of the factors of the validity of the knowledge claims, the credibility of the
research team, and the utility and relevance of the work to situations of educational practice.

Scientific education research, whether it is aimed primarily at uncovering new knowledge or meeting the dual goals of generating knowledge and informing practice, is influenced by the unique configuration of characteristic features of the educational enterprise.

FEATURES OF EDUCATION

Education is a complex human endeavor ultimately aimed at enhancing students' cognitive, civic, and social learning and development. Like medicine, law, or farming, education is a craft—a practical profession requiring specialized skill. Researchers studying teachers have documented that teaching is a complex, interactive exchange as the teachers seek to engage students in learning new material; to relate it to their prior knowledge; to respond to the heterogeneous needs of children with varied backgrounds, interests, and ideas; and to assess the depth and endurance of student learning. Education can occur in school classrooms, private homes, museums, community centers and through information accessible on the Web. Even formal schooling varies in profound ways from community to community, and from preschoolers to adults. Its institutions are many and multilayered—elementary schools, middle schools, high schools, 2-year and vocational colleges, 4-year colleges and universities, and adult learning centers. As an institution, its clientele frequently move, for example, from one school or college to another. The variability and complexity of education is mirrored by the practice of education. In the exercise of their craft, educators draw on, and are influenced by, practical wisdom, professional relationships, and values, as well as scientifically grounded theory and fact. Indeed, it is this real world of research in education that led columnist Miller to lament, “If only education reforms came in a pill” (2001, p. A14).

The character of education not only affects the research enterprise, but also necessitates careful consideration of how the understanding or use of results can be impeded or facilitated by conditions at different levels of the system. Organizational, structural, and leadership qualities all influence how the complex education system works in practice. Results may have
“shelf lives” that vary with cultural shifts and resource changes (Cronbach, 1975).

In the section that follows we discuss some of the salient features of education and their effects on scientific research: values and politics; human volition; variability in education programs; the organization of schools; and the diversity of the many individuals involved in education.

**Values and Politics**

Aristotle once opined that it is impossible to talk about education apart from some conception of “the good life” (Cremin, 1990, p. 85). Indeed, education is a field in which values appropriately play a central role, because what people hope to attain in education—especially the education of children—is intimately connected with people’s views about individual human potential, their hopes and expectations of what society can become, and their ideas about how social problems can be alleviated. In this way, social ideals inevitably and properly influence the education system and in turn, the research that is carried out. More subtly, but crucially, these values also affect the choice of outcomes to study and measure, as they are proxies for the myriad goals of education: basic knowledge and skills, community service, job training, social development, and problem solving. We comment further on the implications of these disagreements about goals in discussing the role of a federal education research agency in Chapter 6.

A more global implication of the role of values in education research concerns the extent to which research in education is truly akin to an engineering science. The question of why education research has not produced the equivalent of a Salk vaccine is telling. After all, medical research is something of an engineering science in that it brings theoretical understanding in the life sciences to bear on solving the practical problems of prolonging life and reducing disease. Education research is similar, with the key difference that there is less consensus on the goal. Medical research often has clearer goals—for example, finding a cure for cancer. Because values are so deeply embedded in education in so many different ways, education researchers do not have a singular practical goal that drives their inquiry (Bruner, 1996).
Local, state, and federal politicians, teacher unions, special interest groups, higher education faculty, and other interested citizens who have a stake in education are often moving in different directions and driven by different sets of incentives. These stakeholders make decisions that influence education policy and practice, and thus have an impact on the research that attempts to model and understand it. At any given time, schools and school systems may be responding to a configuration of possibly conflicting demands from these stakeholders, while trying to serve their primary clients—children, parents, and community members. This dynamic creates a turbulent environment for research. Furthermore, political motivations can affect the uses of research; some stakeholders may have strong incentives to resist the findings or interpretations of researchers or to over-interpret the results if they indicate even modest degrees of evidentiary support.

Another potential consequence of the role of stakeholders is that education research can be interrupted by a change in policy or political support for a particular type of reform. In California, the mathematics and science standards crafted in the late 1980s—which served as important examples for the current national mathematics and science standards—were abruptly changed because of political shifts. Just as the state was gearing up its curriculum, teaching, and accountability system to implement the new standards in a systematic way, the political environment changed, and so did the standards and accountability system (Kirst and Mazzeto, 1996). Research on the reform, too, ended abruptly. Such changes occur as a result of the democratic system of educational governance in the United States, and can have practical implications for research planning (e.g., limiting opportunities to conduct long-term studies).

Human Volition

Education is centrally concerned with people: learners, teachers, parents, citizens, and policy makers. The volition, or will, of these individuals decreases the level of control that researchers can have over the process. For example, in some cases, people cannot be randomly assigned to treatment groups; they will not agree to let themselves or their children be “controlled” for the purposes of experimental trials. This lack of control can also cause problems of noncompliance with research protocols and
instances of missing data because, for example, parents have the interests of their individual child in mind and may have priorities and needs that conflict with those of the research process.

Human movement and change have, for example, affected efforts to study the effects of education vouchers on student achievement. Many voucher studies (Witte, 2000; Peterson, 1998; Rouse, 1997; Peterson, Howell, and Greene, 1999; Myers, Peterson, Mayer, Chou, and Howell, 2000; Peterson, Mycrs, and Howell, 1999)—some designed as randomized trials and some not—face challenges because significant percentages of families did not return the year after baseline data were collected, did not fill out all the questionnaire items, or did not complete the standardized tests. A study of a New York City choice program (Barnard, Frangakis, Hill, and Rubin, 2002) featured a design that anticipated these noncompliance issues, and incorporated the use of sophisticated statistical (Bayesian) modeling to estimate the “treatment” effects of the program under these conditions.

A related point is that the U.S. population is a highly mobile one, with people often moving from one geographical area to another, from one home to another, and from one job to another. And their children follow suit, moving among classrooms, schools, districts, and states. According to data collected by the U.S. Census Bureau, 16 percent of the population changed households between March 1999 and March 2000 (Schaefer, 2001). This mobility characterizes not only precollege students, but college students as well: nearly one-third of students attend at least two institutions of higher education before completing their undergraduate studies (National Center for Education Statistics, 1996). Students are quite likely to experience different curricula, different teaching methods, and different standards for performance depending on the particular classroom, school or university, district, and state. Thus, researchers engaged in longitudinal research in schools are often faced with substantial shifts in the student population—and thus their study sample—which complicates the tracking of students’ learning trajectories over time.

**Variability of Educational Programs**

Researchers typically must accommodate a rapidly changing reform environment that tends to promote frequent changes in the core education
programs a learner encounters. The current education reform movement can be traced back 18 years ago to the report of a Presidential commission, *A Nation at Risk* (National Commission on Excellence in Education, 1983). Since then, the nation has been in a constant process of reforming the schools, and there is no sign that this "tinkering towards utopia" (Tyack and Cuban, 1995) will end soon. Historically, education reform seems to be the norm, not the novelty, in U.S. education, dating back at least to the nineteenth century (Tyack and Cuban, 1995). As one reform idea replaces another, instability in curriculum, standards, and accountability mechanisms is the norm.

Even within reform movements, the state and local control of education significantly shapes the ways that instructional programs and other changes to schooling are implemented, making generalizations difficult. For example, charter schools—public schools that operate under contract with either a state agency or a local school board—take very different forms according to their states’ authorizing statutes and the particular contracts (charters) under which the schools operate (RPP International, 2000). While all charter schools are characterized by some degree of flexibility from state education statutes, their educational programming and student populations vary considerably across and within states. The statute that authorizes charter schools in the state of Minnesota, for example, specifically encourages serving children with special needs. By contrast, many (though not all) charter schools in Colorado were founded by well-to-do parents who wanted rigorous academic programs for their children. Consequently, trying to answer a seemingly straightforward question like “Are charter schools more effective in improving student achievement than traditional public schools?” is not particularly useful if one wishes to understand the impact of instructional innovation because the educational environments and programs that fall under the rubric of “charter schools” are so varied that there is no common instructional intervention to evaluate.

Evaluations of changes in curriculum are also influenced by variability in programs. The implementation of curricula is a cyclic process that is governed by a complex mix of state review, teacher input, district leadership, and public comment. Further, new initiatives often require a significant commitment of funds for professional development, which may or may not be available. High stakes accountability systems and national college
entrance exams also may complicate the evaluation of the effectiveness of curricular change. Like others we discuss in this chapter, these typical circumstances require that researchers be careful to specify caveats and conditions under which findings are produced.

**Organization of Education**

Formal schooling takes place in an interdependent, multilayered system. In the preK-12 system, for example, students are assigned to classes, classes are organized by grade level within a school, schools are organized into school districts, school districts may be organized within counties, and counties are subdivisions of states. In addition, within classrooms, students are often placed into different instructional groups based on instructional needs or related issues. And all are influenced by federal education policy. The implication for research is that to understand what is happening at one level, it is often necessary to understand other levels. Thus, a study of how students come to understand key themes in U.S. history, for example, may be influenced by a teacher's approach to history instruction, the value a principal places on history within the curriculum (which influences how much time the teacher has to teach history and the child to learn it), the curriculum adopted by the district (and related decisions to implement the curriculum), and different familial and community factors (e.g., parent and community support of approach to history instruction). In subject areas such as science and mathematics, where accomplishment in later courses is heavily dependent on the quality of early learning, preK-12 school structures can be designed to either facilitate successful remediation or to systematically exclude increasing numbers of students from these courses over time. These differences demand that researchers consider the nature of the vertical organization of the system in their work.

Education researchers have long investigated the interrelationship of these various levels of the system. Statistical methods, for example, can help estimate educational effects on students' history achievement while at the same time accounting for the effects of the multiple layers of the K-12 system (Bryk and Raudenbush, 1988). A study that examined the mechanism by which Catholic schools achieve equitable outcomes for students used such a technique (see Box 5-3).
Diversity

The U.S. population is becoming increasingly diverse in a number of ways, and demographic projections indicate that the trend will continue (Day, 1996). Mirroring the diversity of the broader population, education takes place in specific neighborhoods with their particular geographical, historical, social, ethnic, linguistic, economic, and cultural mixes. For example, students representing dozens of native languages may attend a single school; in some school districts students speak more than 125 languages (Crawford, 1992). This linguistic diversity that characterizes many U.S. schools illustrates the influence of diversity on research. Students from immigrant families are often defined by a characteristic they commonly share—a lack of English fluency. Scratching just below the surface, however, reveals stark differences. Schools serve students who are new immigrants—often unfamiliar with American life beyond what they might have seen in movies—as well as many Hispanics, African Americans, Asian Americans, and American Indians whose families have lived here for generations and who have varying degrees of English proficiency.

Along with linguistic diversity comes diversity in culture, religion, and academic preparation. Some students visit their home country frequently, while others have no contact with their or their parents' birthplaces. Some immigrant students have had excellent schooling in their home countries before coming to the United States; others have had their schooling interrupted by war; and still others have never attended school. Some are illiterate in their own language, and some have languages that were only oral until recently; others come from cultures with long literary traditions. The differences between these students—their age and entry into U.S. schools, the quality of their prior schooling, their native language and the number of native languages represented in their class, their parents' education and English language skills, and their family history and current circumstances—will affect their academic success much more than their common lack of English (Garcia and Wiese, in press). Incorporating such linguistic and sociocultural contexts into the research process is critical to understanding the ways in which these differences influence learning in diverse classrooms.

In sum, the features that shape the application of our principles of science to education research—values and politics, human volition, variability in education programs, the organization of education institutions,
and diversity—underscore the important role of context. A specific implication of the role of contextual factors in education research is that the boundaries of generalization from scientific research need to be carefully delineated. Our discussion of diversity above is illustrative: to what extent, for example, is it possible to generalize results of research on suburban middle-class children of Western European descent to inner-city, low-income, limited-English students from Central America or Southeast Asia? Naive uses and expectations of research that do not recognize such contextual differences can lead to simplistic, uninformed, and narrow interpretations of research and indiscriminate applications. To build theory, formulate research questions, design and conduct studies, and draw conclusions, scientific education research must attend to such contextual conditions.

This attention to context also suggests that advancing understanding in complex and diverse education settings may require close coordination between researchers and practitioners, interdisciplinary work, and the interplay between varying forms of education research. It also means a far greater emphasis on taking stock of the inherent diversity of the education experience and its results for different populations of students. In short, it requires specific attention to the contexts of research more frequently and more systematically than has been the case for much of the work in education to date (National Research Council, 1999c).

**FEATURES OF EDUCATION RESEARCH**

In addition to the features of education that influence research, there are also aspects of education research as a field that help clarify the nature of scientific inquiry in education. A perspective of education research as an enterprise points to some of the infrastructure supports that sustain it, a topic we take up in our consideration of the federal role in supporting education research (Chapter 6). Three of these education research characteristics are noteworthy in this regard: its multidisciplinary nature, ethical considerations, and its reliance on relationships with education practitioners.

**Multiple Disciplinary Perspectives**

The variability and complexity of education are the grist for the academic’s disciplinary mill. Multiple scientific disciplines study education
and contribute knowledge about it. Economists study the incentive structures of schooling to understand the relationship between interventions designed to change behavior and educational outcomes. Developmental psychologists and subject-matter specialists study fundamental processes of cognition, language, and socialization. Physicists, chemists, and biologists study science curriculum, teaching, and assessment. Organizational sociologists study systems that are organized to meet education goals. Cultural anthropologists study the character and form of social interactions that characterize students’ formal and informal educational experiences. Political scientists study the implementation of large-scale institutional change, like charter schools.

The presence of many disciplinary perspectives in education research has at least three implications. First, since several disciplinary perspectives focus on different parts of the system, there are many legitimate research frameworks and methods (Howe and Eisenhart, 1990). But because many disciplines are focusing on different parts of the system, contradictory conclusions may be offered, adding fuel to the debates about both the specific topic and the value of education research. The challenge for the diverse field of education is to integrate theories and empirical findings across domains and methods. Researchers from a range of disciplines working together, therefore, can be particularly valuable. Ongoing work at the Park City Mathematics Institute (see http://www.admin.ias.edu/ma/) provides an example of the potential for interdisciplinary inquiry in education to enhance understanding and promote effective instruction. A diverse group of researchers (from mathematics education, statistics, and psychology) and practitioners (teachers and teacher educators) have joined to conduct research collaboratively on how students understand statistical concepts (e.g., distributions) in order to provide advice to curriculum developers (Jackson, 1996; Day and Kalman, 2001).

A second implication is that advances in education research depend in no small part on advances in related disciplines and fields. Work in the traditional scientific disciplines, as well as in such applied fields as public health may be necessary as infrastructure support for scientific studies in education.

Finally, this proliferation of frameworks, coupled with the sheer scope of the myriad fields that contribute to understanding in education, make