CHAPTER 5

Scientific Thinking and Science Literacy

RICHARD LEHRER and LEONA SCHAUBLE

Although there is a long tradition of research on the development of scientific reasoning, the impact of this research on science education has been limited and not always constructive. As Metz (1995) pointed out, to the extent that research has provided any guiding picture of development to inform science education, the most enduring influence has come from outmoded misinterpretations of Piagetian research. As a consequence, even now, ideas about children and science are dominated by untested conclusions about what children *cannot* do—or worse, claims about deficits that have already been refuted by evidence, but that somehow continue to hang around like unwelcome relatives, exerting their influence on education via texts, science standards, and the beliefs of educators. These assumptions about what children cannot learn show up with particular frequency in evaluations of the "developmental appropriateness" of approaches to science education or specific topics of study. Metz, for example, charts the influence of these assumptions on the national discussions about science standards and argues convincingly that the standards seriously underestimate young children's capability to learn and do science.

In the previous volume of this *Handbook*, Strauss (1998) suggested several reasons why the best of developmental psychology does not always contribute to the best of science education. He proposes, among other reasons, that developmental psychology and science education share little overlap in content, focus, underlying assumptions, and methods of inquiry. However, since his chapter was published, there has been an acceleration of activity in the intersection between these two fields. Science educators have become increasingly interested in and knowledgeable about learning and development. And some developmental scholars have begun to pursue education in a more serious and committed way. For example, there are now a number of research programs, described later in the chapter, in which investigators are deeply involved not just in studying scientific thinking, but also in changing its course in contexts of education. New programs of research emphasize the coordinated

Order of authorship is alphabetical; the contribution of both authors was equal. We gratefully acknowledge the contributions of Steven McGee and Kathleen Metz, who reviewed the manuscript and offered comments that were extremely helpful.

design and study of science learning in school classrooms, consistent with a wider appreciation of the fact that studying interesting forms of scientific thinking cannot progress very far unless these forms of thinking are brought into being. As a result, research on the development of scientific reasoning is increasingly becoming entwined with the search for effective ways to catalyze and support it.

Typically, this approach to research entails designing and implementing instruction and then studying the resulting student learning over a relatively extended period of time (ideally, several years). These long time periods are required because the forms of thinking that are of interest do not emerge within a few months or even a year. The emphasis in this research is not on describing "naturally occurring" forms of thinking, whatever those may be, but on systematically testing effective ways to support the development of students' reasoning and knowledge over the long term. In addition, many of these projects pursue a secondary interest in the professional development of the teachers who conduct the instruction or in the institutional structures of schooling that both facilitate and constrain educational potential. Because these programs take a longitudinal perspective, they offer the opportunity for a more serious test of accounts of development than do studies that last only a few days or weeks (an opportunity, however, on which it is difficult to capitalize, as we will discuss). Moreover, they are tests of development under conditions in which development is brought into being and sustained by cultural and semiotic tools. As we will explain, the field is currently struggling to decide the extent to which mechanisms of development such as language, tasks, forms of argument, and tools, need to be incorporated into theoretical and empirical accounts.

This general approach to studying development and learning, in which intervention and investigation are conducted as part of a coordinated enterprise, has been called "design experiments" or "design studies." The merits and limitations of this approach are currently being explored and debated (Brown, 1992; Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Shavelson & Towne, 2002; Sloane & Gorard, 2003). However, these conversations are occurring almost exclusively within the field of education research rather than the field of development. Our interest in design studies is in their potential to shed light on both origins and pathways of development, an issue that we take up in the second section of the chapter. In that section, we survey the landscape of contemporary design studies that are informed by and, in turn, inform our knowledge about the development of scientific thinking.

In spite of the emergence of this new research in the overlap between developmental psychology and science education, we can by no means congratulate ourselves that the fields of psychology and education have achieved a comfortable and general consensus about common goals for and conclusions about children's learning. There appear to be two main reasons for this gap. First, not only between, but also within these fields, there are long-standing disagreements about what it means to learn or understand science. These disagreements are partly due to the lack of shared vision in our society about the purposes for education in general. More particular to *science* learning, there are also competing views of the nature of science, so that we lack consensus on the character of the phenomenon under investigation. Second, within the field of developmental psychology there are long-standing differences of opinion about the nature and mechanisms of development and how developmental research can best inform and be informed by the educational enterprise. These disagreements are also at play, and views of how best to study development color perspectives about how learning should be supported.

For example, some scholars emphasize mechanisms that are conceived primarily as internal qualities of the developing individual and especially emphasize those forms of development that appear to be universal to the human species and therefore relatively robust across varying contexts and cultures. Others have argued that psychology attends too much to explanations of development that are based on presumed inner mental processes, traits, or constraints operating at the level of the individual organism. These scholars argue that an adequate account of development needs to include the local and distal contexts that support and shape it. From this perspective, the focus of study should be on the structures, goals, and values associated with the activities that people are habitually immersed in; the kinds of tasks and problems they encounter in contexts of learning; the content and structure of their prior knowledge; their histories of learning; the cultural expectations, tools, and behavioral patterns that are part of an individual's world; and the social and historical contexts that shape contemporary activity. Of course, this tension between

explanations based on mental qualities of individuals versus the physical and social environments is an old and ongoing story in developmental psychology, one that seems to continually reshape itself as the field evolves.

In sum, different views of science literacy and learning are at least partly the result of differences in answers to two questions: What is developing when children learn science? and What is development? Where progress is being made, it has been by reformulating and testing the implications of different answers to these enduring questions. Therefore, we begin the chapter by considering different images of the nature of science, because these images have either explicitly or tacitly guided the conduct of developmental research. The second section of the chapter revisits some familiar territory—studies of the growth of scientific reasoning—but reconsiders them in light of the images of science that they assume and also in light of longer-term studies where development is (deliberately) shaped by education. This section examines the assumptions about scientific thinking and development that inspired longer-term investigations of development and summarizes how both traditional and design approaches contribute to what we understand about learning and development.

The design studies emphasize somewhat different views of the nature of science and, taken as a group, entail a contrasting set of educational designs based on different "bets" about how to catalyze development over the long haul. This new research is important for both developmental psychologists and science educators to understand. For science educators, it is providing a beginning empirical base to inform the debates about the nature of science and resulting implications for education. For developmental psychologists, it may reframe our expectations about trajectories of cognitive development and the influences that can shape or change those trajectories.

As we will explain, classroom design studies encounter a host of challenges that laboratory research typically does not. For example, taking a long-term view of learning and development often requires a fundamental rethinking of the subject matter under consideration. Historically, decisions about what is worth teaching and learning have been informed not by knowledge about learning and development, but by politics and custom. These decisions are often strongly influenced by the organizational structure and constraints of schooling. The curricular shape of a school discipline is laid down by historical tradition and can be very difficult to reenvision. The way a subject has been previously taught comes to take on canonical status as it is encapsulated in textbooks, standards, tests, and preservice teacher education, and (equally important) in the expectations of parents and the public at large. These historically entrenched views about what science learning or history learning or mathematics learning should be like can be very difficult to change (Dow, 1991), as the current "math wars" amply illustrate. Yet, as we will show, a developmental perspective, coupled with longitudinal research on learning, tends to raise fundamental questions about the status quo vision of school disciplines. Taken seriously, thinking developmentally may change the landscape considerably, both for what should be learned and for how it is learned.

The third and final major section of the chapter illustrates in greater detail how these issues play out, using as an illustrative case a design investigation conducted over 10 years by the authors. Although in principle, any of the examples in this chapter might serve as the case for this analysis, the issues we discuss in this section require exposing the way design research works under the hood, information that is usually known well only to those close to the project in question. Matters usually dismissed as "implementation" or "logistical" issues seldom appear in journals or other public presentations, but in design research they should be accounted for as part of the theory of action, rather than dismissed as side issues. The purpose of this final section is to show how this form of investigation requires researchers to find new ways of addressing research concerns such as representativeness, generalizability, and replication, which cannot always be handled in the same ways as in laboratory investigations (although closer inspection of experimental laboratory studies suggests some clear parallels, especially in new domains of research; see, e.g., Gooding, 1990).

A word on what the chapter will *not* address. There are many fields of research that bear on the issues that are discussed here. They include science education, social studies of science, semiotics, the history and philosophy of science, and cognitive models of learning and development. To avoid taking the chapter too far afield, we keep our central focus trained on classroom studies that take a developmental approach to science learning and scientific reasoning. Research in related fields is

introduced only as it bears directly on the chapter's primary focus.

IMAGES OF SCIENCE

Images of the nature of science set the stage for the study of development. They inform what researchers choose to study and suggest appropriate means of study.

We have identified three images that appear to have attracted broad research support: science-as-logic, science-as-theory, and science-as-practice. Here we briefly describe each of these views of science and then further exemplify these positions by contrasting their stance toward the idea of experiment, which is an epistemic form characteristic of and central to the practice of science.

Science-as-Logical Reasoning

Science-as-logic emphasizes the role of domain-general forms of scientific reasoning, including formal logic, heuristics, and strategies, whose scope ranges across fields as diverse as geology and particle physics. This image figures prominently in three early programs of research that have been especially influential in the way researchers conceptualize scientific thinking. These include Inhelder and Piaget's (1958) pioneering work on formal operations; the Bruner, Goodnow, and Austin (1956) studies on concept development; and Wason's (1960, 1968) four-card task studies demonstrating that people tend to avoid evidence that disconfirms their prior theories. The image of scientist-as-reasoner continues to be influential in contemporary research (Case & Griffin, 1990). Learning to think scientifically is conceived as a matter of acquiring strategies for coordinating theory and evidence (D. Kuhn, 1989), mastering counterfactual reasoning (Leslie, 1987), distinguishing patterns of evidence that do and do not support a definitive conclusion (Fay & Klahr, 1996), or understanding the logic of experimental design (Chen & Klahr, 1999; Tschirgi, 1980). These heuristics and skills are considered important targets for research and for education because they are assumed to be widely applicable and to reflect at least some degree of domain generality and transferability (D. Kuhn, Garcia-Mila, Zohar, & Andersen, 1995).

A general feature of studies conducted in this vein is that researchers often attempt to rule out the use of knowledge by relying either on unfamiliar tasks based on knowledge that children are considered unlikely to have, or on tasks that are intrinsically content lean. For example, in a study of problem-solving strategies, D. Kuhn and Phelps (1982) asked children to investigate mixtures of clear, unlabeled chemical solutions in an attempt to find out "for sure" which mixtures, when added to a mixing liquid, would reliably turn pink. The content of this problem was considered unlikely to evoke participants' prior content knowledge in ways that would either help or hinder them in solving the problem, as preadolescent children typically know little about chemical solutions. Moreover, only alphabetical labels on the test tubes identified the chemicals, and all of the chemicals were indistinguishable clear liquids. The labels were changed after every trial, making it impossible for participants to develop cumulative knowledge about the materials over time. Indeed, the authors were not interested in how children think about chemical solutions; they chose this content because they wished to understand the kinds of evidence-generation and evidence-interpretation strategies children would employ in solving problems that involve multivariable causality and, in particular, how those strategies might evolve over repeated trials as children received feedback from observable changes in the physical materials.

A point on which there is no consensus is whether these forms of reasoning should be conceived of as specialized knowledge that is difficult to acquire and that emerges only gradually over development, and in many people never appears at all (D. Kuhn et al., 1995), or alternatively, whether they are appropriately viewed as the application of problem-solving strategies that are common to all kinds of thinking (Klahr, 2000). In either case, the task for developmental researchers is to identify origins, patterns of change, and underlying mechanisms of change in skills and strategies that are presumed to be useful across a wide variety of situations and problems particular to science (and perhaps everyday thinking as well).

Science-as-Theory Change

Science-as-theory change draws from philosophical studies of science and compares individual conceptual change to broader historical trends in science, especially

the periodization (i.e., normal and revolutionary science) of science identified by T. S. Kuhn (1962). Among others, Carey (1985b) and Koslowski (1996) have suggested that disciplinary knowledge evolves in ways that typically involve the gradual accretion of new facts (e.g., Kuhn's normal science) and knowledge or, occasionally, the replacement of one idea by another. At critical junctures there may even be wholesale restructuring of the theoretical landscape (e.g., Kuhn's scientific revolutions). In this case, the entire network of concepts and their relationships is reconfigured (Chi, 1992). Not only do new concepts enter the domain; in addition, existing concepts may change their meaning in fundamental ways because the theoretical structure within which they are situated radically changes. Consider, for example, the meaning of the concept *force* or *combustion*. *Force* in Aristotelian theory is not the same concept as *force* in Newtonian theory. Note, however, that we would be unlikely to conclude that scientists who believed in the phlogiston theory or who held Aristotelian notions of force and motion were illogical, in the sense of lacking or violating important canons of reasoning. Instead, we accept that scientists of earlier times reasoned in ways that depended on their knowledge and theories. Under different assumptions about the way the world worked, different kinds of conclusions and inferences would seem quite logical, perhaps even obvious.

If the development of scientific reasoning in individuals is like the development of scientific knowledge over the course of history, the argument goes, it is best conceived not as the mastery of domain-general logic, heuristics, or strategies, but as a process of conceptual or theory change. In fact, some of the research in this tradition is aimed toward demonstrating that children's reasoning per se does not differ in important ways from adults' (e.g., Carey, 1985a; Samarapungavan, 1992).

Carey (1985a), for example, claimed that there is nothing about the power or structure of children's logic that develops, at least beyond the preschool years. In her landmark studies challenging Piaget's (1962) earlier assertions about the "magical" or "animistic" thinking of preadolescent children, Carey (1985b) demonstrated that this apparent animism did not entail failures of children's reasoning, but instead reflected their theories about properties that distinguish living organisms from nonliving objects. Her results suggested that children lack some of the fundamental biological knowledge that adults have. Even more important, the knowledge that children do have is organized into conceptual systems (i.e., theories) that do not reflect either the overall structure or the categories typically possessed by adults. For example, when asked to provide examples of things that were "not alive," children's responses suggested that they were conflating a number of distinctions that an adult would honor into a general, undifferentiated alive/not alive opposition. As examples of things that are "not alive," children proposed organisms that had been alive but were now dead (a cat run over by a car) or extinct (dinosaurs), were representations (a drawing of an animal rather than a "real" animal) or imaginary. On the basis of responses like these (and a number of other clever experimental tasks), Carey showed that it may be a mistake to assume that when a child judges an example as "alive" or "not alive," he or she is relying on a conceptual system like the one that most adults have in mind. Carey concluded that there is no evidence that children think magically or illogically. Rather, their judgments make perfect sense given their conceptual understanding of the world. Developmental change, under this account, is conceived not as the mastery of thinking processes or a new form of logical or abstract thinking, but as changes over time in one's stock of knowledge about the meaning of terms like "alive," as children collect both first- and secondhand experience with organisms and their properties. These changes in the knowledge system accumulate, and when they reach a critical level the conceptual system restructures to accommodate the inconsistencies.

Indeed, all the relevant logical equipment can be presumed to be intact at least by the time children begin school. (Whether parts of this knowledge are already in place at birth, learned at very early ages, or governed by inborn constraints is a question being actively investigated.) Even participants in content-lean studies import knowledge in an attempt to make sense of the problems and tasks they encounter. Researchers in the science-as-logic tradition have generally acknowledged that it not really possible to rule out the influence of prior knowledge and have instead focused more directly on *how* knowledge and other factors might systematically influence participants' reasoning strategies and heuristics (D. E. Penner & Klahr, 1996; Schauble, 1990, 1996). From the theory change perspective, reasoning strategies and heuristics are tools for theory development. Epistemic commitments of theories are especially important targets for development, including, for example,

whether or not a new theory is free of contradiction, accords well with previous theoretical commitments, and accounts for evidence, both actual and potential (Posner, Strike, Hewson, & Gertzog, 1982).

Science-as-Practice

Science-as-practice is an image formulated from studies of science that emphasize observational studies of scientific activity, both in the short term (e.g., studies of activity in a particular laboratory or of a program of study) and historically (e.g., studies of laboratory notebooks, published texts, eyewitness accounts). Science-as-practice suggests that theory development and reasoning are components of a larger ensemble of activity that includes networks of participants and institutions (Latour, 1999); specialized ways of talking and writing (Bazerman, 1988); development of representations that render phenomena accessible, visualizable, and transportable (Gooding, 1989; Latour, 1990); and efforts to manage material contingency, because no theory ever specifies instrumentation and measurement in sufficient detail to prescribe practice. The alignment of instruments, measures, and theories is never entirely principled (e.g., Pickering, 1995). What the other two images of science take as foundational (reasoning and theory) together comprise only one leg of a triangle that also includes material procedures (e.g., making instruments and other contexts of observation, almost always involving machines) and models of how the material procedures function to render nature visible (Pickering, 1989).

The descriptions of science that are produced in this tradition of research suggest that science includes many different forms of practice, ranging from experiment to comparative study. For example, experimental physics tends to favor experiment as a critical form of argument, a tradition initiated several centuries ago (Sibum, 2004). As examples of this, see Shapin and Schaffer's (1985) description of the epistemic controversies aroused by Boyle's then novel experimental approach in the seventeenth century and Bazerman's (1988) description of Newton's role in the genesis of the experimental report. In contrast, even contemporary studies of evolution rely on comparative methods. For example, Van Valkenburgh, Wang, and Damuth (2004) recently tested tenets of natural selection by examining the fossil record of North American carnivores during the past 50 million years. Their argument was comparative in the sense that predictions were made about the effects of individual selection on extinction rates of large carnivores, and these were then compared to the extant fossil record.

Each of the components of practical activity cited in social studies of science appears critical for the overall success of the enterprise. Consider, for example, inscriptions (representations that are written). Latour (1990) suggests that systems of scientific inscription share properties that make them especially well suited for mobilizing cognitive and social resources in the service of scientific argument. His candidates include (a) the literal mobility and immutability of inscriptions, which tend to obliterate barriers of space and time and thus "fix" change so that it can be an object of reflection; (b) the scalability and reproducibility of inscriptions, which guarantees their economy but preserves the configuration of relations among elements of the represented phenomenon; (c) the potential for recombining and superimposing inscriptions, operations that generate structures and patterns that might not otherwise be visible or even conceivable; and (d) the control of reference, because inscriptions "circulate" throughout a program of study, taking the place of phenomena, yet maintaining an index to the original events that inspired their creation (Latour, 1999, p. 72). Lynch (1990) adds that inscriptions not only preserve change, they edit it as well: Inscriptions both reduce *and* enhance information.

Inscriptions serve epistemic commitments. Gooding (1989) examined how patterns made by iron filings in magnetic fields were transformed into displays featuring geometric curves and lines of force. These new technologies of display helped establish a language of description for the new phenomenon of electromagnetism "while also reinforcing the scientific values it embodied" (p. 186). Similarly, Kaiser (2000, pp. 76–77) suggested that the enduring and recurrent use of Feynman diagrams in particle physics was due to the diagrams sharing visual elements with the inscriptions of paths in bubble chambers, a correspondence that appealed to realism: "Feynman diagrams could evoke, in an unspoken way, the scatterings and propagation of real particles, with 'realist' associations for those physicists already awash in a steady stream of bubble chamber photographs."

Science-as-practice emphasizes the complicated and variable nature of science. What develops, then, must include logic and theory (Dunbar, 1993, 1998) but also ways of talking about phenomena and otherwise partic-

ipating in a community of practice (Gee & Green 1998; Lemke, 1990; Warren & Rosebery, 1996); inventing and appropriating display technologies, sometimes called representational competence (diSessa, 2002, 2004; Goodwin, 1994; Greeno & Hall, 1997; Roth & McGinn, 1998); becoming initiated into the lore of managing contingency within domains, including how to construct variables when Nature does not tell (e.g., Ford, 2004; Lehrer, Carpenter, Schauble, & Putz, 2000); and appreciating the different forms of method employed in different sciences. Because science-as-practice must, by definition, include opportunities to participate in these practices, studies of development that are guided by this image typically track long-term change in environments designed to support participation in scientific practices. As Warren and Rosebery (1996) summarize:

> From this perspective, learning in science cannot be reduced simply to the assimilation of scientific facts, the mastery of scientific process skills, the refinement of a mental model, or the correction of misconceptions. Rather, learning in science is conceptualized as the appropriation of a particular way of making sense of the world, of conceptualizing, evaluating, and representing the world. (p. 104)

Rethinking Images of Science: What Is Experiment?

A comparative analysis of *experiment* may serve to heighten the contrast among these images of science. Science-as-logic regards experiment as a form of reasoning dominated by a singular rationale: control of variables. To experiment is to control, and what develops is an appreciation of this logic. Experiments are valid with respect to the space of possible manipulations of variables. Science-as-theory takes a different tack, treating experiment as a "critical test" of a theory. Critical experiments under gird theory change because they have the potential to produce anomaly and thus initiate conceptual change. Science-as-practice regards experiment as a resolution of an apparent paradox (Latour, 1999). Experimental facts are made—with instruments, material, and ingenuity—and so never can be regarded simply as nature observed (Galison & Assmus, 1989). Theories thus always have a practical side. They rest on foundations of mediated activity (e.g., representations, apparatus, instrument readings, interactions with other participants, design of the experiment). Yet, this practical activity becomes less visible to those who routinely practice it. As initiates are taught to see in particular ways, the products of experiment are treated as ascendant, and the activity whereby they are made becomes transparent, so that experimental facts become unmoored from their original settings (Gooding, 1989, 1990; Shapin & Schaffer, 1985; Sibum, 2004). Thus, from the science-as-practice perspective, experiment is complex and textured.

Implications of Images of Science for Education and Development

As noted, the images of science-as-logic and science-as-theory have dominated the debate about appropriate explanations for developmental change. These two views seek their support in different forms of evidence. Moreover, they tend to be associated with different views of the most appropriate goals for science education. It is interesting that science education has also engaged in its own long-standing debate about the relative importance of scientific knowledge and theories, on the one hand, versus scientific thinking, on the other. In general, school science has tended to emphasize learning what Duschl (1990) calls "final form science," that is, its end products: concepts, facts, and theories. However, school texts that communicate this "rhetoric of conclusions" (Schwab, 1962) often fail to reveal how that knowledge was produced. Teaching facts, concepts, and theories as final form science may leave students in the dark about the way knowledge is generated and may also distort the nature of scientific knowledge, inappropriately conveying that it is unchangeable and uncontested. Partly as a corrective to traditional textbook approaches, educators in the 1960s began to argue that the focus of education should instead be on "science process skills," such as observing, predicting, measuring, and inferring. Indeed, one of the most influential post-*Sputnik* National Science Foundation curricula was titled *Science: A Process Approach* (American Association for the Advancement of Science, 1964). However, it quickly became evident that the learning of domain-general processes could easily become as ritualized and meaningless as the learning of textbook facts. Moreover, the application of these skills seems to be tightly tuned to particular situations, tasks, and content. They are not easily acquired in one realm and then transferred to others, even when their use would be advantageous. Perhaps for these reasons, "process skills" approaches have largely fallen out of

favor in science education research (although they still seem appealing to curriculum designers and school faculty; they appear regularly in published commercial curricula and school standards documents).

Science educators agree on the importance of helping students appreciate the epistemology of science, although there is little consensus on how to do so. National science standards, for example, emphasize the importance of providing an opportunity for students to get a taste of doing science at their own level of knowledge and expertise. Indeed, inquiry is a major theme in the National Science Education Standards (Minstrell & van Zee, 2000; National Research Council, 1996). The reference to inquiry (rather than reasoning or process skills) is intended to communicate that scientific knowledge *and* scientific thinking should be inseparable goals of education, always pursued hand in hand (Bransford, Vye, Kinzer, & Risko, 1990). In the context of developing and pursuing scientific investigations that are focused on scientific knowledge, students learn inquiry skills *and* science content. As yet, however, little agreement has been achieved on what these skills might be, the extent to which they are transferable across domains, or how (indeed, whether) their mastery can be assessed (see D. Kuhn, Black, Keselman, & Kaplan, 2000, for a discussion of these matters).

As in the education field's attempt to substitute the process/content dichotomy for an integrated emphasis on inquiry, the field of research has also increasingly acknowledged that science involves *both* characteristic ways of thinking *and* conceptual structures. In research, as in education, there has been growing interest in seeking to understand these as complementary aspects of scientific reasoning. Researchers are investigating how they coevolve and are building and testing models of thinking that coordinate these two aspects of science.

For example, Klahr and Dunbar's (1988) Scientific Discovery in Dual Spaces model describes scientific reasoning as a process of integrated search through two problem spaces: a space of hypotheses and a space of evidence. In this model, moves in each of these problem spaces affect the potential movements in the other, either by constraining potential moves or opening new possibilities. As described in much of the general research on problem solving, a scientific reasoner generates a mental representation of the problem (the "problem space"), and his or her solution of the problem is modeled as a heuristic search through that set of possibilities. In the dual search space model, goals include generating observations that may lead to the formulation of hypotheses, finding evidence that confirms or disconfirms hypotheses that are currently being entertained, or deciding among competing hypotheses. Therefore, the model incorporates hypotheses (which presumably have their origins in beliefs, concepts, or theories), strategies for generating and evaluating evidence, and descriptions of the interactions of search in these spaces in the course of scientific reasoning. In addition to this modeling approach, researchers (Klahr, 2000; D. Kuhn, Amsel, & O'Loughlin, 1988; D. E. Penner & Klahr, 1996; Schauble, 1990, 1996) have pursued empirical studies that systematically examine the effects of prior beliefs on students' strategies and heuristics for generating and evaluating evidence (and conversely, the effects of different strategies on changes in participants' theories).

Note, however, that whether a researcher believes that "what develops" is scientific concepts, scientific reasoning, or both, an assumption common to these perspectives is that the goal is to identify the most important aspect or essence of science, so that researchers can investigate its development and educators will know what to teach. Maybe, however, there is no such kernel. Perhaps what is most important about science is not its essence or core, but its variability. The science-as-practice image suggests that sciences span multiple epistemologies and practices. Moreover, perhaps what is important with respect to development is not characterizing changes that are internal to individuals, but understanding how individuals are initiated into and participate in these variable ways of knowing and doing science. From an educational perspective, the goal in that case would be to consider which forms of practice provide the greatest educational leverage, and then to understand how to assist students in beginning to participate. Primary attention would go not to investigating the developing knowledge or logic of individuals, but to characterizing the role of the systems in which cognition occurs, with special attention given to the array of semiotic and other tools that support and mediate thought.

WHAT IS DEVELOPMENT?

These views about the appropriate focus for research and education are closely associated with perspectives on the nature and mechanisms of development. This, of course, is the "What is development?" issue introduced

earlier in the chapter. From its origins as a field and throughout its history, developmental psychology has always preferred explanations based on the internal mental properties of individuals. There seems to be a bias toward seeking some form of biological essence as the ultimate explanation for development. This has been true from the origins of the field in Gesell's maturationist accounts to today's emphasis on identifying innate knowledge and genetically predetermined constraints on learning. It has been difficult in practice to conceptualize a developmental psychology that is not deeply rooted in assumptions about maturation and teleology. Indeed, for some investigators, what defines a phenomenon as *developmental* is that it has a universal character and appears to be governed at least in part by biological predispositions. With some important exceptions, the field of developmental psychology has largely regarded context, culture, history, and education primarily as noise, or at best, as *factors* that affect the course of development. Agreeing on how to legitimately bring these concerns into the purview of developmental study remains a struggle in the field.

As an alternative, one could conceive of development as inseparable from the means that support it, so that an account of "Under what conditions?" is considered an obligatory question that an adequate explanation of development must address. This kind of perspective is useful for scholars and practitioners who are concerned not just with describing or explaining development, but also with catalyzing and supporting it, or in some cases, changing its course in particular ways. Yet in general, mainstream developmental psychology has made little progress with the thorny problem of conceptualizing development and context. Indeed, the increasing attention in the field to younger and younger children could arguably be interpreted, at least in part, as an attempt to sidestep these difficult issues of culture and context.

Research on scientific reasoning that is conducted from a psychological perspective has relied mainly on cross-sectional investigations of individuals at different ages (less frequently, amount of education is used as an independent variable). A second, less frequently pursued methodology has been to track a group of individuals over the short term, conducting dense measurements to document the onset and pattern of change (D. Kuhn, 1989; D. Kuhn et al., 1988; D. Kuhn & Phelps, 1982). However, with one exception (Bullock & Ziegler, 1999), we know of no longitudinal research on scientific reasoning from a psychological perspective that extends beyond several weeks in duration. Indeed, cross-sectional studies (Klahr, Fay, & Dunbar, 1993; D. Kuhn et al., 1995) seem to suggest that there is more overlap than separation across age groups in the skills or heuristics typically investigated, and that education seems to be at least as important as whatever else is implicated by looking at individuals of different ages.

Although informed by the psychological research, much of the work featured in the second section of this chapter emphasizes the role of education and other semiotic means that constitute thinking. From this perspective, science entails the deployment of a set of very broad and eclectic psychological functions, marshaled in relationship to a web of complex and varying goals, pursued by a community over a changing history, and supported and shaped by culturally developed tools and semiotics. Under this view, there is no one psychological "essence" of science. Instead, science is regarded as a complex form of human practice. The term *practice* as used here refers not to the external organization of behavior, but to patterns of activity that are initiated and embedded within goals and thoroughly saturated with human meaning and intentions. "What develops" is a capability to participate in these practices of science. Researchers who pursue this perspective do not necessarily deny that scientific thinking entails logic, epistemology, and theory change. However, they argue that what is essential to account for is how these psychological functions are constructed by, contingent on, and expressed within social contexts and mediational means. Moreover, scientific reasoning is not conceived as knowing how to design experiments *plus* understanding patterns in evidence *plus* building a consistent and coherent knowledge base about a domain. Rather, each of these functions is viewed as fundamentally contingent on the others, so studying them as a collection of independent capabilities or *skills* may generate a distorted understanding of the intact enterprise.

This perspective on research tends to turn attention to sources and forms of variability, rather than to a search for universal or general forms of cognition. Variability is conceived as being understandable (and produced) by attending to the mediational features that support and provide meaning for scientific thinking or, from an educational perspective, that can be deployed as design features to instigate and support developmental change. These features may include histories of learning, teaching, and other forms of

assistance; cultural expectations of all levels and kinds; tasks and tools; genres of writing and argument; inscriptional and notational systems; and recurrent activity structures. Note that these items are conceptualized neither as internal psychological resources nor as external environmental stimuli; rather, they are understood to be externally instantiated (i.e., they have material expression) but imbued with meaning that is conferred by people.

The perspective of this chapter is not that either the psychological or practice view is "more right" than the other. However, one advantage that the practice view holds for education is that the elements that it takes as primary are potential instruments of change. One cannot directly engineer changes in people's psychological capabilities. Educating involves understanding and deploying tools, tasks, norms of argument, and classroom practices to bring about desired ends (Lehrer & Schauble, 2000c). Understanding how these and other designable features serve to generate and sustain cognition is, therefore, a useful goal for scholars and practitioners concerned with education.

Regardless of one's view on development, there remain unresolved questions concerning the characterization of science that is most appropriate for school science. The next section is devoted to describing current classroom investigations in which researchers work in partnership with teachers and others in school organizational structures to craft conditions that can best support the long-term development of students' participation in the practice of science. Each program emanates from prior developmental research, so we include these antecedents to situate the design studies. Taken collectively, the design investigations emphasize somewhat different views of scientific practice and, therefore, result in educational designs based on different bets about ways of conceiving scientific practice that serve to catalyze development. The way to understand the implication of these bets is to instantiate the designs and conduct longitudinal study on the development of student thinking that results. Debates about the best way to conceptualize scientific reasoning (for educational purposes, at least) are difficult to resolve unless the bets can be cashed in and the outcomes compared. Each approach is very likely to have both strengths and characteristic weaknesses; as in any design enterprise, these need to be evaluated as trade-offs.

CLASSROOM DESIGN STUDIES AND DEVELOPMENT

In this section, we describe current classroom studies in which scholars are working to coordinate two interrelated agendas. First, they seek to change educational practice in ways that foster the development of scientific thinking. As will become evident, each of the projects featured here exemplifies a somewhat different sense of "what develops." Thus, there is variability in what is taken as important early origins or precursors to scientific thinking, as well as in what is supported and studied along the way. Second, as these educational change experiments come into play and evolve over time, researchers study the cognitive and other forms of development that result among participating students. An important related goal is to understand the variety of means by which development is supported (Cobb et al., 2003), reflecting a general commitment to conceiving development as a culturally supported enterprise rather than a naturally occurring phenomenon.

Of course, there have been hundreds of classroom investigations that feature attempts to support students' scientific reasoning and knowledge. This chapter does not attempt to review all of them, or even all those that may be relevant to the development of children's scientific thinking and knowledge. Instead, we focus on a few cases that, collectively, exemplify the landscape of design studies in science, investigations in which scholars are pursuing the study of development by trying to change it. Examples that are featured here were selected for their fit to the following criteria:

- First and most important, these are projects that are developmental in their focus. In some cases, this means that the educational intervention was constructed on a foundation of knowledge from the literature in cognitive development. In others, the project may not be directly motivated by developmental studies, but it is conceptually consistent with current findings about development and makes new contributions to our understanding of development, typically by challenging what is "known" about development. These challenges often take the form of generating forms of thinking and learning that have not been previously documented. As a group, these investigations

are concerned both with identifying early origins or precursors of valued forms of thinking, and also with documenting change over time in the target forms of reasoning. In addition to describing the classroom studies on their own terms, for each, we also briefly summarize related research from developmental psychology that shows how the project links to the mainstream concerns of that field.

- In addition to focusing on the development of children's thinking, these projects take a developmental stance toward the domain of *school science*. Each embodies a perspective about how what is taught can contribute to a broader agenda of science literacy. The view of change is long term and looks well beyond the learning of a particular skill or concept. The typical grain size of interest is what can be accomplished over years of instruction, not within a lesson or a unit. All the work described here has given careful thought to what should count as a "big idea" in science education. As we will see, at this point the research agenda for most of this work still lags far behind the conceptualization.
- In each research project presented here, education is taken seriously. That is, the educational agenda is regarded as having intrinsic value. Accordingly, schools are not regarded merely as places to find participants for research, and education is conceived as more than tasks designed to tap some psychological function. The projects are situated in schools that are not unusually privileged with respect to student populations and resources. All of them have had to grapple with the actual conditions of schools, and all have had to address the thorny problem of sustainability.

Our intent is not to catalogue all work that fits these criteria, but to provide examples that illustrate the variety and breadth in the ways that investigators are conceiving of the intersection between science and development.

Not all the scholars whose work is reflected in this section identify themselves as conducting design research, but their research shows many of the commitments that design studies exemplify. Design studies are coordinated efforts to design learning environments and then to study the transitions in teaching and learning that follow. Those studies typically take many methodological forms, from traditional experiments or quasi-experiments to descriptive or ethnographic work. The distinguishing characteristic of this approach is not its use of any particular method, but a tight and cyclical interaction between two complementary aspects of work: instructional design and research. Working from a base of previous research, analysis of the domain, and theory, researchers plan and craft the design of a learning environment, which may vary with respect to scope. Concurrently, they conduct a careful and systematic program of research on the learning that results as the design coalesces. As the research proceeds, it produces findings that call for revisions to the design. Sometimes these changes are minor, sometimes radical. The changes, in turn, generate new questions for investigation.

An assumption of the design studies approach is that many forms of learning that are important targets of inquiry cannot, in fact, be studied unless the conditions for their generation are present. Thus, they are particularly applicable to the study of forms of development that require sustained education for their emergence. As we mentioned, each design investigation places different emphases on which practices are important to sustain over longer periods of time. Often, these "best bets" have roots in developmental approaches informed by one or more of the three images of science, although in practice, all prolonged studies are hybrids.

Supporting the Development of Scientific Reasoning

Inhelder and Piaget (1958) asserted that only at the onset of formal operations, around the beginning of adolescence, do children become capable of understanding the logic of scientific experimentation. This claim, like many others concerning presumed deficits in children's cognitive capabilities, eventually fell to evidence generated by subsequent research. Microgenetic studies conducted by D. Kuhn and her associates (1988, 1995; Schauble, 1990, 1996) confirmed that only small percentages of preadolescents initially produced valid scientific reasoning strategies or heuristics when attempting to solve multivariable problems without much guidance from adults. However, when given extended opportunities to conduct repeated trials in microgenetic designs (D. Kuhn & Phelps, 1982), most of the children in these studies began to show increasing use of more effective strategies for designing and interpreting experiments (D. Kuhn et al.,

1995; D. Kuhn, Schauble, & Garcia-Mila, 1992; Schauble, 1996). These strategies included investigating all relevant combinations of variables and their levels, controlling extraneous variation, and making inferences that are appropriately based on the available quality and quantity of evidence. Indeed, many of the participants went beyond simply beginning to use the new strategies to mastering and consolidating them. That is, they almost always used the new strategies when it was appropriate to do so; the earlier, flawed strategies were eventually abandoned altogether; and participants even transferred the new strategies to unfamiliar problems that did not share surface features with the original learning context (D. Kuhn et al., 1992).

Indeed, the origins of these heuristics are evident even as early as the preschool years. In a carefully constructed sequence of studies, Sodian, Zaitchik, and Carey (1991) demonstrated that preschoolers could consider two alternative tests of a hypothesis and reliably identify which would actually settle the question. However, their succeeding appeared to depend on a number of simplifying circumstances: that the alternatives did not confirm or challenge strongly held prior beliefs, that the number of choices and variables was kept very restricted, and that children were asked simply to evaluate alternatives rather than to propose an experimental design on their own. Nevertheless, these studies do show that at least in a rudimentary way, children can differentiate their beliefs from evidence that bears on those beliefs.

Promoting Understanding of Experimental Design via Instruction

Building from earlier findings that children can understand the logic of experimental design (Tschirgi, 1980), Chen and Klahr (1999) suggested that the kernel of a science education for children is mastery of the logic of the control of variables. They recommended that children should be taught to ignore or look behind particular content to focus on structural relationships. This is precisely what children were trained to do in Chen and Klahr's educational studies. In one investigation, students learned to evaluate the design of experiments by making judgments about the informativeness of pairs of trials presented by a researcher as a "test" of causes and effects in a multivariable context. Each trial included several potentially causal independent variables (that could be set at different levels) and an outcome variable (also with several levels). Students understood that the point of the comparisons was to make a decision about whether one of the independent variables was causally related to the outcome.

For example, in one context, students were told that their task was to evaluate an experimental trial's utility for helping to decide which factors determine how far a ball will roll down a ramp. The experimental comparison included two small ramps that could be set at either steep or shallow angles, with starting gates set at different positions on the ramps. The ramps were fitted with a reversible insert that would produce either a rough or a smooth ramp surface. Two different test balls were provided, a golf ball and a rubber squash ball. Children observed pairs of configurations of these materials and were asked whether or not each comparison supported a definitive conclusion.

Among the trials shown to children were various forms of invalid tests. For example, the two-ramp setups might differ in multiple ways, making it impossible to tell whether one of the variables was the causal one. In such a case, a child might observe a golf ball rolled down a steep ramp with a rough surface, and the comparison case would involve a rubber ball rolled down a shallow ramp with a smooth surface. If the two conditions led to a different outcome, it would be impossible to know why, because several variables had been varied simultaneously. The 7- to 10-year-old participants in the study were shown several examples of both confounded comparisons like these and other comparisons where extraneous variation was controlled. In each case, the participant was asked to decide whether the comparison was a "good test" or a "bad test." In a training condition, participants were provided explicit feedback after each trial; the experimenter also explained why the test either was or was not flawed. Chen and Klahr (1999) reported that not only were they able to improve children's abilities to judge the informativeness of experiments and to make inferences based on them; in addition, the older children were able to transfer the strategies they had learned to novel contexts, even after a 7-month delay. Moreover, Klahr and Nigam (2004) demonstrated that children who were taught these strategies were able to use them to evaluate science fair posters a week afterward.

The instruction designed by Chen and Klahr (1999) was tightly focused on the logic of control of variables. Although science-like materials (ramps, springs, and sinking objects) were used, the logic would have been precisely the same if the tasks had borne no relationship

whatsoever to science topics. Hence, this body of work is a particularly clear example of science-as-reasoning.

Practices of Investigation as a Route to Developing Reasoning

Like Chen and Klahr (1999), Kathleen Metz (2004) emphasized developing skills and strategies important to the conduct of scientific inquiry. However, in Metz's classroom investigations, the focus on domain-general forms of reasoning was not pursued at the expense of domain-specific conceptual knowledge. Instead, children received repeated opportunities to plan, conduct, and revise related *programs* of research in the service of developing coherent conceptual structures concerning important biological ideas such as behavior and adaptation. In this sense, the practices of children were similar to those of scientists.

For nearly a decade, Metz has been pursuing classroom design research with the ultimate goal of maximizing children's capability to conduct independent inquiry. A main conjecture of her research is that the learning of skills and knowledge is best supported in contexts that maintain the integrity of the original goal-focused enterprise where those skills and knowledge originated. Therefore, research methods and strategies should be introduced to students not as disembodied skills, but as tools for pursuing real questions that children pose in domains where they have opportunities to develop significant content knowledge.

In Metz's work, children are deeply immersed in one discipline, often for a year or longer. Metz (2004) makes the case that students should concentrate intensively in a relatively small number of domains, rather than learning a little bit about a wide variety of topics. After all, one cannot conduct inquiry in a field in which one knows nothing, so a curriculum that emphasizes breadth over depth is not a good one for supporting inquiry. Properly supported, the development of content knowledge and the development of scientific reasoning should bootstrap each other.

Inquiry depends on students being able to generate fruitful questions, acquiring a repertoire of appropriate methods for investigating those questions, and developing a sense of the forms and qualities of evidence (and counterevidence) that can inform the answers. Consistent with this view, Metz's participants study one scientific domain at a time—such as animal behavior, ornithology, botany, or ecology—for an extended period in which they repeatedly encounter the core ideas of the domain in multiple contexts. Initial investigations are carefully structured and scaffolded; subsequent inquiries are planned and conducted by the children themselves, who are increasingly given independent responsibility for the progress and evaluation of the scientific work.

For example, as an introduction to animal behavior, students in second and fourth/fifth grade began by conducting observations of a rodent confined to a small space in the center of the classroom. The fact that every child was observing the same animal meant that, inevitably, they selected different behaviors to describe, interpreted the behaviors differently, or failed to record them in a common form. These occurrences motivated debates about the need for standard ways to observe and also provoked awareness of the fact that under some conditions (such as loud talking), observation can change the behavior of the organism being observed. Children typically attribute intentions and thoughts to animals, so the observations also produced a forum for discussing the difference between observations and inferences, a distinction that Metz considered fundamental for subsequent work. In small teams, children recorded and displayed their data, and the different data displays generated a reason to talk about how data representations of different design communicate different information.

After these initial observations, students were reorganized into pairs and each pair was given their own organism to study, in this case, one or more crickets. Crickets available for observation varied both between species and within species (gender, age, etc.), raising questions about relationships between these variables and observable animal behaviors (such as chirping, fighting, or eating). Various forms of controls and research methods specific to the domains of study (e.g., time sampling as a technique commonly used in animal behavior) were introduced. To pursue the goal of building a rich knowledge base that could inform inquiry, students supplemented their direct observations with reading material, videotapes, and other media. The research teams independently generated questions about the crickets, and then the whole class compiled their questions and categorized them on a number of dimensions, including whether the questions were amenable to empirical inquiry ("Is this a question that you can collect data on?"). In some cases, students explicitly noted differences between forms of thinking in everyday contexts, as contrasted with their use in science.

For example, students concluded that in science, they might not always achieve consensus and that this was acceptable if they had good justifications for failing to agree. Students learned to recognize and mark sources of uncertainty in their developing knowledge.

Initial investigations with crickets were planned by the whole class working together and then were conducted individually by pairs working as research teams. The teacher assisted in recording and categorizing questions, summarizing observations, and developing a table that displayed classes of questions that might be investigated and methods appropriate for doing so. For subsequent investigations, both the direction and the procedures were increasingly ceded to the students. Finally, using the previously developed list of heuristics for evaluating potential questions and the class-generated list of domain-specific methods for investigating questions, each team planned and conducted its own investigation. The investigations culminated with a poster presentation in which each team presented its question, methods, and findings.

The notion of inquiry exemplified in these sequences contrasts sharply with the typical cookbook laboratory exercises in which students carefully carry out step-by-step procedures, and also with hands-on science activities and kits in which a preordained course of investigation is followed. Students in Metz's classrooms have much more (although not boundless) freedom to select their own question to pursue. This means that it is essential for curriculum designers to identify domains of study that support a wide variety of student questions, all of which, however, must be very likely to lead students directly into confrontation with one or more important scientific ideas. An important topic of Metz's research is to identify domains that have these properties. Metz's approach differs, as well, from those advocated by Chen and Klahr (1999) or D. Kuhn (1989), in which the content and surface features of inquiry tasks are considered secondary and the emphasis is on repeated practice at making logical judgments about problems with varying surface features that preserve a common underlying structure.

Metz (2000) identified five different aspects of children's knowledge that were the primary foci of this instruction: children's conceptual knowledge of the domains under investigation, their understanding of the enterprise of empirical inquiry, their knowledge of domain-specific methodologies, data representation and analysis, and tools. Careful study of the progress of children's investigations, coupled with postinvestigation interviews of the children's research teams (Metz, 2004), provided information about children's achievement of these goals. Findings were reported for one class of second graders and one class of mixed fourth/fifth graders, both from a public elementary school in a rural area, who had participated in the first iteration of the animal behavior curriculum.

All 10 of the second-grade research teams and 14 fourth/fifth-grade teams formulated both a researchable question and a method for investigating their question, although one second-grade team initially chose a research method that was not appropriate. Most of the second graders and about half of the fourth graders relied on the class-generated heuristics for identifying a good question. Interestingly, about half of the older teams pursued questions about social behavior, although none of the younger children did. The majority of the younger children conducted studies of the effect of some variable on cricket behavior by comparing the behavior of the crickets under different conditions. In sum, children showed considerable competence at taking charge of their investigations, even coming up with sophisticated proposals for controlling extraneous variation that seem surprising, given the previous literature about children's spontaneous performance on problems that require them to produce or evaluate comparisons that involve controls (Chen & Klahr, 1999; D. Kuhn et al., 1988).

After instruction, each team was individually interviewed about their conceptualization of their question and the method used to investigate it, their findings, and whether they could think of a way to increase their confidence level in the findings. In addition, each team was asked whether they could think of any way to improve the study. In her analysis of these interviews, Metz (2004) paid particular attention to how children conceptualized the sources of uncertainty in their study and the strategies they pursued in trying to resolve the uncertainty. A few of the younger children apparently held the simple idea that the point of inquiry is to produced a desired outcome, so that what was uncertain was how to make the experiment "work," a notion that has shown up repeatedly in previous research with preadolescent children (i.e., Schauble, 1990; Tschirgi, 1980). About 25% of the children focused primarily on the possibilities of uncertainty in their data that were due to imprecision of their instruments or experimenter error. About 15% of the children (approximately equal percentages in both grades) described themselves as uncer

tain about the generalizability of the trend in their data. The most frequent reasons given for this uncertainty were that the study was conducted within a limited range of experimental conditions or that the variability of the crickets made children uncertain that results achieved with some crickets would apply to others. Nearly 40% of the second graders and 25% of the older children were uncertain that their theory was adequate to account for the trend that they observed in the data. Finally, the most common source of uncertainty was attributed to the trend identified in the data (over 40% of the second graders and 85% of the fourth/fifth graders). In most (but not all) of these cases, children were able to propose at least one strategy to resolve the uncertainty.

In sum, the participants seemed to understand the problematic nature of knowledge in several respects: that uncertainty can enter the data generation process in a variety of ways, that what they know about their research question is mediated by the study they conducted and its inherent flaws and uncertainties, and in general, that the relationship between the world and the scientist's knowledge of it is far from straightforward, but rather complex and interpretation laden. Metz (2004, p. 282) concluded, "At least by the second-grade level, the decontextualization and decomposition of the elementary science curriculum appear to be more a function of curricular traditions than developmental need."

Instruction that is organized around self-directed investigations needs to maintain the right balance between investigation skill and the development of conceptual knowledge. But it is not always a simple matter to find that balance. In practice, teachers must be skillful to negotiate the tension between these two components and must continually work against the tendency for one to fade into the background as the other takes center stage. Metz advised extended study within a coherent domain of knowledge as a way of balancing the focus on methodology with a corresponding emphasis on the development of a rich knowledge base. Because children's knowledge built cumulatively over weeks and months, their repeated opportunities to conduct and interpret investigations not only familiarized them with a repertoire of methodologies, but also provided opportunities to construct expertise in a bounded but complex domain of investigation.

In summary, Metz's approach to supporting development of scientific reasoning has a methodological bent: It places its bets on introducing children to methods commonly employed by scientists, but it does so in contexts of prolonged investigation of a rich content domain. It borrows from studies of scientific practice to instantiate aspects of scientific community. Questions and investigations have both a self-directed and a communal nature. What we know less about from these studies is the nature of the conceptions children are developing about the domains under investigation. Clearly, they are developing methodological commitments akin to those of scientists. But do their evolving understandings of crickets serve as gateways to larger conceptual structures in biological sciences? And if so, how? These questions have been more explicitly addressed in research guided by science as theory development.

The Development of Theories

We contrast two programs of research, both of which are centered in theory change but that make different commitments to origins and analysis of what develops. The first, Intentional Conceptual Change, draws from science education and is informed by a view of science as a process of conceptual change. The second, Pathways to Science, as its name suggests, draws from studies of early origins of children's theories about nature and seeks to capitalize on these origins to create developmentally appropriate education.

Intentional Conceptual Change

For many years, Sister Gertrude Hennessey was the sole science teacher for grades 1 though 7 in a small parochial school in Wisconsin. As a result, she had the unusual opportunity to think about the goals and trajectory for students' scientific reasoning across all those grades of schooling. Fortunately, she had both the educational background and the wisdom to capitalize on this opportunity to pursue long-term development (she holds degrees in biology and science education). Hennessey not only planned the course of instruction and taught her students daily; she also kept detailed records and videotapes of her students' learning and conducted regular interviews of individuals, small groups, and intact classes. She pursued a structured approach to science instruction that made students' thinking visible and therefore accessible to her observation. From time to time, she collaborated with university researchers from both developmental psychology and science education to conduct cross-sectional and longitudinal studies of student learning (e.g., Beeth & Hewson, 1999; Smith, Maclin, Houghton, & Hennessey, 2000).

Hennessey (2002) regarded science learning primarily as conceptual change. However, in pursuing this characterization of science, she drew primarily on the field of science education rather than psychology. She was particularly influenced by the work of Posner and his colleagues (1982) and later revisions by Hewson and Hewson (1992), who pursued what they called a conceptual change model (CCM) to account for how students' mental representations of the world might shift from initial, naive notions to the conventionally accepted explanations of science. The CCM described conceptual change as a process by which a concept might be replaced by another, modified, or simply dropped. Critical to the conceptual change model is the assumption that the relative overall *status* of a concept for a particular learner determines whether the concept will be maintained or changed when an alternative is under consideration. Status refers to how the concept is evaluated relative to a consistent set of criteria. How an individual applies those criteria depends on his or her prior knowledge, motivation or stakes in both the new concept and those that it may replace, and ontological and epistemological commitments. Specifically, the evaluative criteria associated with status include the learner's evaluation of the concept's *intelligibility* (how comprehensible is it?), *plausibility* (is it believable?), and *fruitfulness* (how useful is it for getting things done in the world or for motivating new investigations?). A fourth factor, not directly included in status but important nonetheless in whether a concept is maintained or changed, is *conceptual coherence,* whether and how the new concept fits or fails to fit into the preexisting network of related knowledge. Hewson and Hewson used the analogy of a "conceptual ecology" to refer to the balanced interrelationships among beliefs. As in ecological systems in biology, the metaphor of conceptual ecology emphasizes the importance of interdependencies. Changing one concept is very difficult or impossible to do without changing others that are closely related. According to Hewson and Hewson, each concept occupies a niche within its conceptual ecology. Concepts, like organisms, may compete for survival within a niche. However, a concept is unlikely to be discarded or changed unless the individual becomes dissatisfied with it. Therefore, helping children clearly articulate the beliefs that they hold and, in some cases, helping them notice the inconsistencies or insufficiencies of those beliefs are reasonable strategies for a teacher who hopes to help children make conceptual progress toward accepted scientific theories.

Hennessey's instructional approach was to explicitly teach students the evaluative criteria in the Conceptual Change Model, starting with the earliest grades of instruction. Her emphasis was not primarily on learning what each criterion meant in a disembodied way, but on putting the criteria to use in the context of building their own explanations for scientific phenomena and deciding among competing explanations produced by other members of the class.

Hennessey placed a great deal of importance on students' developing metacognition, hence the emphasis on scientific reasoning as *intentional* conceptual change. Becoming aware of one's own theories and explicitly evaluating them against the conceptions proposed by peers was fundamental to her goals for students. Notice, however, that in contrast to a more general emphasis on self-regulation and self-evaluation, Hennessey's classrooms were focused on a more restricted sense of metacognition, one tightly tied to the CCM epistemology of science. Hennessey was adamant that her interest in improving metacognitive understanding was not a general, all-purpose goal for students conceived as transferable across content and disciplines. Rather, student metacognition was pursued in the service of achieving domain-specific conceptual change. Moreover, neither metacognitive development nor conceptual change was regarded primarily as an end in itself, but rather both were considered to be ways of helping students achieve the more fundamental goal of engaging with deep, domain-specific ideas in science.

Given this strong emphasis on epistemology of science, it is not surprising that Hennessey's instruction was frequently based on direct experience with the natural world. Students frequently began a unit of instruction by directly exploring a phenomenon carefully chosen to provoke surprise (again, the emphasis on anomaly in theory change), given students' likely prior beliefs and assumptions. Students worked with the phenomena, typically in a laboratory or field setting, recording questions that came up in the context of their explorations. Then, as in Metz's (2004) work, they planned and carried out investigations to answer their questions. These were more likely to be investigations with physical materials than "research" in books or online. In the course of these investigations, students were encouraged to represent their ideas in a variety of formats (charts, graphs, diagrams) and to compare their ideas with those being developed by other students. The emphasis was on first clarifying one's own ideas and then, after evaluating theories against the evidence and

against competing theories posed by others, evaluating and revising those ideas to account for anomalies experienced in the course of ongoing investigations.

In sum, the view of science portrayed in Hennessey's program was that science is a matter of developing and building progressively more adequate theories about the world. Moreover, what develops is not only the scientific theories but, equally important, students' critical standards for defending, adapting, or replacing those theories. Although Hennessey did not discuss development in depth in her published articles, she clearly had ideas about the general course of development of these metacognitive criteria. At each grade, students were expected to build on accomplishments in earlier grades, constructing a progressively more sophisticated capability to reflect on and evaluate their own theories and those of their classmates. Her goals for first graders were modest, focused primarily on helping students become adept at stating their own beliefs and providing reasons for them. By the fourth grade, students were expected to understand and apply all four criteria of intelligibility, plausibility, fruitfulness, and conceptual coherence as they evaluated their evolving beliefs. By sixth grade, students were also monitoring the beliefs of others, especially their peers, and considering the fit of competing explanations to patterns of evidence.

Hennessey's sixth graders, who at that point had received a total of a half-dozen years of instruction under her tutelage, were interviewed with an instrument previously developed by Carey and colleagues (Carey, Evans, Honda, Jay, & Unger, 1989) to ascertain their understanding of the nature of science. In this study, their performance was compared to a demographically similar group of sixth graders who were taught with a more traditional elementary program. The Nature of Science Interview (Carey et al., 1989) was designed to roughly classify students' responses with respect to their conceptual grasp of the epistemology of science. Level 1 ideas, compatible with what Carey and Smith (1993) called a *knowledge unproblematic epistemology,* reflect a belief that knowledge is certain and unproblematically true. It is a relatively simple matter to know what is true; one simply has to look (or be told). Responses classified as Level 2 reflect an understanding that scientists are concerned with explanation and testing, but nonetheless, knowledge is still regarded as true, certain, and discernible. Level 3 responses, in contrast, explicitly note that knowledge is tentative, changeable, and significant only within an interpretive framework.

In previously published research involving Massachusetts public school students (Carey et al., 1989), all seventh graders had provided interview responses that were classified as Level 1. In contrast, 83% of the students in Hennessey's classroom produced responses that were classified at least as Level 2. Smith and her colleagues found four clusters of issues that differentiated Hennessey's students from those in the comparison classroom. First, when asked about the *goals of science,* the Intentional Conceptual Change students said that scientists are involved in understanding and developing ideas. In contrast, the comparison students mentioned simply doing things and gathering information. The two classes also differed on the *type of questions that scientists ask.* Hennessey's students more frequently described questions about explanations and theories, whereas the majority of the comparison students' examples were about procedures (how to do things) or questions that Smith et al. (2000) referred to as "journalistic" (identifying who, what, where, when). When asked about the *nature and purpose of experiments,* Hennessey's students were likely to highlight testing a particular idea or to refer to the role of experiments in developing theories. The comparison students, in contrast, referred to experiments as a way to try things out or to find (unproblematic) answers to questions. Finally, when students were asked *what causes scientists to change their ideas,* many of the Intentional Conceptual Change students responded that scientists change their ideas when they are able to develop a better explanation, or pointed out in other ways that change is a response to complex evidence. In contrast, the dominant answer provided by the traditional students was that scientists decide to either keep or throw out an idea after one simple observation or experiment. Only a third of the comparison students spontaneously noted that changing a scientific idea requires hard work or careful thought.

By and large, Hennessey's students had not yet achieved a sophisticated Level 3 view that included either the logic of hypothesis testing or an acknowledgment of how framework theories entail coherent principles that shape the development of hypotheses. Yet, these findings from the Nature of Science Interviews suggested that most of her students had achieved an understanding of the epistemology of science that is quite unusual for their grade. Indeed, Smith et al. (2000) reported that they found these sixth graders' replies to be similar or superior to responses typically given by 11th graders.

Although Hennessey does not explicitly mention it, there are close conceptual ties between her Intentional

Conceptual Change project and developmental research on children's criteria for evaluating theories. Samarapungavan (1992) investigated first graders', third graders', and fifth graders' criteria for scientific rationality in a study in which children observed a phenomenon and then were asked to select which of two explanations accounted better for their observations. The pairs of explanations were constructed to be identical in surface features, but to contrast on one of four "metaconceptual criteria," as she called them. These criteria, which seem quite similar to the criteria that Hennessey emphasized, included such issues as *range of explanation* (how much of the observational data does the theory account for?), *non-ad hocness* (is the theory simple or does it include a number of added-on assumptions that are not testable?), *consistency with empirical evidence,* and *logical consistency* (internal consistency, lacking mutually contradictory claims). Samarapungavan found that even the youngest children in her sample preferred theories that met these criteria when they were choosing between competing theories that were consistent with their own prior beliefs. Even the first graders preferred the empirically and logically consistent theory to theories that were inconsistent. These children also preferred theories that could account for a broader range of observations. On the other hand, when the theory of broader range contradicted their prior beliefs, they were less likely to favor it. The most difficult criterion was the one Samarapungavan called ad hocness. Only the 11-year-olds reliably rejected theories that were overelaborated with special conditions or auxiliary hypotheses that could not be directly tested.

In her interpretation of these results, Samarapungavan (1992) cautioned that she thought of these criteria as heuristic only, not as definitive of the value of competing explanations. In her view, any of these criteria could legitimately be overridden by a more important concern, that is, whether the content or meaning of the new idea being considered was compatible with existing scientific ideas held with a reasonable amount of confidence (Hennessey's conceptual coherence criterion). Therefore, like Hennessey, Samarapungavan also gave highest priority to the fit between a concept and other related knowledge, or, as Hewson and Hewson (1992) might describe it, how and where a concept fit into the individual's "conceptual ecology."

Samarapungavan (1992) also pointed out the importance of understanding that although children may use these criteria in a simple forced-choice task, this does not mean they have mastered them or even that they were consciously aware of them. Students in her studies merely chose between two options and were never asked, for example, to formulate an explanation on their own. In some cases, students' choices were consistent with one of the criteria, but they did not explicitly mention the criterion in their justification for that preference. Samarapungavan suggested that children hold some of these criteria only implicitly. Therefore, she recommended, it would be helpful to highlight these "metaconceptual dimensions" in science instruction to foster awareness of them and support their systematic use.

We turn now to a more direct kind of connection between developmental research and an educational program. In this case, the educational intervention followed directly from a major trend in developmental research. This is not very surprising, given that one of the program developers is a prominent developmental scholar who conducts research on the origins of children's concepts and theories. Preschool Pathways to Science, which we describe shortly, resulted from a collaboration between developmental researchers and educators.

Early Resources for Scientific Thinking

Shortly after the seminal work of Jean Piaget became widely known in the United States, scholars began to investigate further his findings and conclusions, especially his claims that infants and young children literally do not possess the same forms of logic that adults do. Piaget's theory held that logic must be painstakingly constructed anew by each individual as he or she grapples with the regularities of objects, space, time, and cause that (in Piaget's view) necessarily structure our experience with the world and our evolving conceptual systems. These concepts and more complex forms of intelligence are developed gradually as each person adapts to the structure of the world through his or her actions upon it.

These claims inspired a flurry of interest in identifying more precisely the cognitive resources of infants and young children. Gelman and Baillargeon (1983) summarized the research on the development of Piagetian concepts and concluded that the evidence was not consistent with the idea that there are major, domain-general qualitative shifts in children's reasoning. Rather, Gelman and Baillargeon interpreted this research as suggesting that both the nature and the development of cognitive abilities are domain specific. Moreover, given the robustness and regularity with which some domain-

specific concepts emerge, it may well be that their development is governed by mechanisms that are genetically governed. As they remarked, "One lesson of modern research in child psychology is that accounts of how development proceeds can no longer ignore the possibility that at least some of the structures that underlie our systems of knowledge are innate" (p. 220).

The notion that infants may enter the world with well-formed knowledge in some domains—or at least, may be especially prepared for ready learning in them—was influenced and informed by related work in the field of ethology. For any animal, some things are very easy to learn, whereas others are very difficult. For example, as Gallistel, Brown, Carey, Gelman, and Keil (1991) point out, pigeons learn relatively easily to peck a key to obtain food but find it difficult to learn to peck a key to avoid receiving a shock. In contrast, they easily learn to flap their wings to avoid a shock. Like pigeons, humans also seem to be genetically prepared for some forms of learning. A frequently cited example of preferential learning is the relative ease with which most infants learn their native language. Moreover, babies learn language in a remarkably orderly way; both the sequence and timing of the emergence of language components are quite consistent across children and across cultures.

Not only do babies learn certain things with relative ease; they also seem to arrive in the world already possessing forms of knowledge that are relatively complex. In contrast to Piaget, who believed that infants' knowledge of objects developed very slowly over the first months of life, most contemporary developmentalists now believe that babies' conceptions of objects are much like those of adults. Recall that Piaget believed that young infants do not initially integrate information that comes from different sensory modalities, so that the appearance of a bouncing ball, the sound it makes as it bounces, and the way it feels when grasped may not be perceived as related aspects of a single, intact object. Yet, recent research suggests that infants are born prepared to process a world full of three-dimensional objects and that their perception of these objects is amodal (i.e., knowledge that comes in from different sensory modalities is integrated in a common mental representation: Children perceive objects, rather than uncoordinated sights, sounds, and tactile sensations). The child's mental representations of the world are interrelated from the very beginning; they are rich and complex and support all kinds of inferences and predictions about the appearance, motion, and qualities of objects.

Early findings along these lines, coupled with the invention of new technologies for studying cognition in preverbal children, have resulted in an explosion of research on the cognitive capacities of increasingly younger children and infants. Researchers have produced surprising new knowledge about infants' capabilities that was previously unforeseen. For example, even infants in the first year of life seem to know that two objects cannot simultaneously occupy the same place (Baillargeon, 1987). They directly perceive causality in displays in which one object seems to bump into and propel another (Leslie, 1984). They know about the continued existence of an object even when it is hidden from view after being observed (Baillargeon & Graber, 1988). On the other hand, they do not apparently expect unsupported objects to fall (Baillargeon & Hanko-Summers, 1990; Hood, Carey, & Prasada, 2000). At this point, it is not settled whether (and if so, which types of) this infant knowledge is intact at birth, develops as a result of innate predispositions to attend to some things at the expense of others, or emerges as the result of general learning mechanisms, perhaps operating under constraints.

Knowledge about objects and motion supports predictions and expectations, delineates the kinds of events and evidence that will be salient to the perceiver, and provides constraints on the kinds of inferences that are made. Moreover, children's knowledge about objects is not simply a list or collection of ideas; it appears to be organized in a tight network of interrelated concepts that are internally structured. Knowledge about objects also participates in wider knowledge structures, such as the coherent system of ontological classification that children develop (Gopnik & Meltzoff, 1997; Keil, 1992). Because the knowledge of objects (and certain other fundamental domains) appears to be structured in these ways, some researchers have argued that at least certain classes of infant knowledge can be appropriately described as early "theories," which serve the function of organizing both past experience and the generation of new knowledge. The so-called theory theorists emphasize that even babies' mental representations are structured, abstract, and complex. Therefore, although babies' theories may differ in content from those of adult scientists, the theories of both groups nonetheless share important defining properties. Moreover, these theories may be revisable as experience strengthens them or requires their elaboration or adaptation (Gopnik & Meltzoff, 1997). Revisability, of course, is an attribute also characteristic of the theories of adult scientists.

Debate continues about how "theory like" these mental structures are and what forms of early knowledge can be presumed to share these theory-like properties. Frequently mentioned candidates include children's "theories" of physical objects and their interactions, biology and living things, number/quantity, and the nature of human mental life.

Attempts to identify and characterize these "core theories" and to understand their character, their origins, and the mechanisms of their development currently account for considerable research activity in the field of cognitive development. The resulting domain-specific accounts of the ways that children's theories emerge and develop over early years of life have now come to the awareness of science educators, who, for their part, have tended to pay close attention to the influence of naive theories, but only later in the life span. The "misconceptions" literature in science education has been conducted mainly with students at high school or university age. Hundreds of studies have now amply demonstrated that even after succeeding at high levels in school science instruction, students often continue to cling to naive preconceptions about the way the world works. In many cases, these preconceptions are at direct odds with the implications of the science that students have just "mastered."

The growing research base about the origins of young children's theories, considered against this context of older students' failures to deeply understand the scientific theories they have been taught, suggests that it may be valuable to seek and develop potential links between children's unschooled theories of the world and the concepts and theories introduced in school science.

Preschool Pathways to Science

This concern is reflected in Preschool Pathways to Science, a program for prekindergarten children (Gelman & Brenneman, 2004). In this program, instruction is organized around core concepts, such as *biological change,* that are central in children's naive theories and also seem to hold the potential to serve as a firm foundation for acquiring important disciplinary understanding in science. The goal of instruction is the development of conceptual knowledge—not isolated definitions, but systems of concepts that are linked into the kind of rich, interconnected knowledge structures described in the research on core theories. The instruction also includes a focus on communication, including language and other forms of representation, such as writing, drawing, mapping, and charting. Students are encouraged to learn and use precise vocabulary, such as *observe, predict,* and *check,* that makes processes of their inquiry more visible to them and, hence, more open to inspection and self-evaluation.

Children's science work in this program is designed to first capitalize on and then extend children's initial theories about the world. One example described by Gelman and Brenneman (2004) involves a series of investigations about the distinction between what one knows and how one knows, a distinction that the "theory of mind" research identifies as difficult (and not only for young children). The teacher began with a discussion of the five senses, discussing what could be learned about an apple via each of the senses. Students were encouraged to record their observations and eventually to make predictions about things that could not be observed (such as the appearance of the inside of the apple, or the number of seeds). Children checked their predictions by opening the apple and making new observations. As a general principle, learning how to "talk science" (Lemke, 1990) and do science always occurs in the context of learning scientific concepts. Children develop their conceptual knowledge by doing science, and scientific processes and tools are regarded not as disembodied skills, but as a means to learn more about the domain at hand.

The emphasis throughout is on strengthening deep conceptual connections by revisiting the central concepts in a domain via a variety of activities and contexts. Because relevant prior knowledge enhances learning, the topics and concepts in the curriculum deliberately build on domains in which children already have relevant knowledge, such as the core theories about biology and physical properties of objects that have been identified by the theory theorists. Some of the concepts that teachers have developed from these starting points include change (biological, chemical, physical), insides and outsides of objects and organisms, relationships between form and function, and systems and interactions (Gelman & Brenneman, 2004), all topics in which children's early intuitions provide potential starting points for instruction. In each case, the goal is to capitalize on a child's-eye view of a topic, building on these early intuitions by providing additional illustrations, elaborations, and, in some instances, counterexamples that can challenge children's initial mental schemas.

Strengths of this program include that the classroom work is innovative and well connected to a solid research

base. As in most of the educational interventions we have described to this point, an important principle is to build deep knowledge within a few content domains, rather than to sample broadly. Children's prior knowledge is to be identified and harnessed, not dismissed or overridden with correct conventional explanations. At the same time, capitalizing on children's intuitive ideas does not mean stopping there. In all cases, the point is to build on these ideas along what Gelman and Brenneman (2004) refer to as a "learning path." Particular emphasis is paid to explicitly marking for children the forms of thinking that are valued and helping underline distinctions between these and everyday kinds of thinking. Specialized vocabulary is cultivated to assist in achieving this goal.

In spite of the many strengths of this program, its long-term outcomes are as yet unknown. Perhaps because of the commitment to tying education tightly to existing research, the developmental trajectory of the program is somewhat restricted, in that it does not extend beyond first grade (although Gelman has been involved in science programs for high school students that share some similarities with this general approach; see, e.g., Gelman, Romo, & Francis, 2002). A critical next step is to first conceptualize and then test empirically the central thesis of the approach: how the extended elaboration of a few central ideas can pay off in the long term with a deeper understanding of scientific ideas that have traditionally been challenging for students to learn. In short, what does the learning pathway look like farther down the road? Conducting conceptual analyses of the links between early theories and later learning can provide first hints, but testing these ideas will require longitudinal study. At this point, little is known about how those relationships are ideally expected to develop or how consistently they can be supported across years of education. Understanding these important questions may require extending the learning research on this program into elementary school, possibly beyond.

Learning to Participate in Scientific Practice

As we mentioned earlier, all long-term investigations of development make commitments to initiating students in forms of practice. In this section, we review programs in which this orientation served as the overriding rationale, although each program clearly also draws from developmental studies conducted with images of theory change or reasoning-skills/heuristics in mind. We contrast a founder program of work, Fostering Communities of Learners (Brown & Campione, 1994, 1996), which explicitly designed instruction to mimic the workings of a scientific community, with later programs that placed primary bets elsewhere, on supporting students' efforts to participate in practices of invention and revision of models of nature.

A Landmark in Developmental Science Education: Fostering Communities of Learners

One of the first attempts to implement and test a long-term developmental view of science education was the Fostering Communities of Learners (FCL) project, directed by Ann L. Brown and Joseph Campione over the course of a decade and a half in the 1980s and 1990s. This project was influential as an educational approach and as a way of conducting developmental research, although the influence on the field of development has been less pervasive than that on education. This work was pioneering in many ways, and all of the long-term projects described in this chapter have been influenced by it, in spite of some differences of opinion about goals and approaches.

In FCL, Brown and Campione (1994, 1996) sought to identify and test "developmental corridors," that is, pathways of the typical development of student knowledge from intuitive ideas to understanding of deep principles in the domains of investigation (like Gelman and Brenneman's, 2004, learning pathways). The shape and direction of these corridors was viewed as being determined by interactions among the capabilities and prior knowledge of students, the forms of teaching and support provided to learners, and the content and structure of the discipline being taught. These pathways were conceived both as conjectured trajectories for instruction (considered as continually revisable, based on emerging results) and as typical patterns of student change. The image of science was one of social community, where theories were developed and subjected to test according to criteria developed within that community. The vision of community was not insular: It included textual accounts and interactions with domain experts, including demonstration lessons. FCL was one of the first developmentally inspired projects in science learning that took seriously the importance of students' learning histories within the content domain of science.

Brown's earlier work in developmental psychology played an important role in the design of FCL. Indeed,

as early as 1978, Brown was foreshadowing a key assumption of FCL:

> Our estimates of a child's competencies are sometimes dramatically changed if we consider them in naturally occurring situations. If, therefore, we are in the business of delineating the cognitive competencies of the 4-year-old, we will have a distorted picture if we see the 4-year-old only in a laboratory setting. (Brown & DeLoache, 1978, p. 27)

FCL was entirely consistent with Brown's early emphasis on studying cognitive functioning in the contexts where thinking is naturally put to work. Brown (1992) conceived of school as a place where learning and development could be studied in their interaction. As Vygotsky argued (see Brown & Reeves, 1987), development and learning are related in close and complex ways, a view that contrasts with the typical assumption that development precedes learning and acts as a constraint on it. Children are smarter in contexts where being smart has a function, is expected, and is supported; understanding development relies on opportunities to study it in contexts of that kind. These assumptions led Brown out of the psychological laboratory and into the business of engineering contexts that nurture development and, therefore, produce opportunities to observe and understand it. Although this was by no means the first design study, because of Brown's prominence in developmental psychology, it was the first to become widely known to scholars in that field.

Brown's investigations of memory development in the 1970s were also influential in the direction taken in FCL. Her specific interest in metacognition and self-regulation, a topic where her research was especially influential, foreshadowed the role of metacognition as a pervading theme in FCL. In FCL classrooms, the overriding goal was to progressively turn over to students the responsibility for both the progress and the evaluation of their own learning and to help them construct the tools for managing this responsibility. This goal was pursued in a number of specific ways and was a prominent concern motivating everything from the activity structures in the classroom to the forms of discourse that were favored and supported. Much of the class's learning occurred in small research groups organized and directed by the students themselves. Students, rather than the teacher, were the ones to decide both who contributed to class discussions and the order of participation. Students learned to talk to, convince, and challenge each other rather than the teacher. The teacher, in turn, guided the topic selection and student work in instructionally fruitful directions and worked to build a sense of accountability, both to one's fellow students and to other audiences of a variety of kinds (students were regularly responsible for making presentations, preparing teaching materials for younger children and reports directed to classmates, and communicating with scientists from outside the classroom). Standards for evaluation of classroom work were consensually developed, publicly shared, and, as far as possible, transparent.

A recurring activity structure in the FCL classrooms was reciprocal teaching, a reading comprehension program that Brown had developed in collaboration with Anne Marie Palincsar (Palincsar & Brown, 1984). Reciprocal teaching was another means of placing self-regulation front and center in students' learning, in this case, for understanding information presented in textual form. In reciprocal teaching, students acquired, practiced, and eventually mastered the kinds of comprehension strategies that more expert readers use spontaneously. Students first learned to imitate strategies modeled by the teacher and eventually, with assistance, began to take over key roles themselves. For example, readers might be asked to provide a summary, ask a clarifying question, or make an inference on the basis of the given information. As students became more expert, the teacher progressively ceded responsibility to student group leaders for these kinds of functions; eventually, students read together in small groups and group members negotiated meaning. The studies on reciprocal teaching documented impressive and lasting gains in the reading comprehension of even struggling readers.

FCL teachers relied heavily on reciprocal teaching to carry out the central activity in FCL classrooms, namely, the conduct and sharing of research in cycles that Brown and Campione (1996) referred to as "research-share-perform." The research conducted in these classrooms primarily involved reading, analyzing, and compiling texts of various kinds (written, electronic, or video). Products of the children's research were also typically in the form of text or talk; they might be posters, public presentations, written reports, or teaching materials intended for younger children. The heavy emphasis on reading, analyzing, integrating, and preparing written information was consistent with Brown's earlier work in reading comprehension with reciprocal teaching and also with the general emphasis on metacognition and self-regulation that permeated FCL.

Typically, a research cycle began with the teacher introducing an important disciplinary theme (e.g., biolog-

ical adaptation). These themes and topics were identified by the project team, which included domain experts, as being fruitful for supporting deep understanding of important disciplinary ideas and productive for focusing the research of student teams. Topics were introduced with an "initiating event," such as a compelling story or a video, which provided a jumping-off point for students' questions and interests. Students would next convene in a whole-class discussion to generate a list of questions that the story, video, or classroom visit raised for them. The teacher categorized and guided questioning with an eye to ensuring that the important themes identified by the project team were represented in the questions that were subsequently investigated. Small teams of students would adopt one of the questions to "research." Commonly, an overarching theme like "food webs and chains" would be divided among the students, so that members of each research group became specialists in a single part of the problem. In the example explained in Brown and Campione (1996), the students studying food webs convened in specialty groups studying photosynthesis, energy exchange, competition, consumers, and decomposition. In another classroom, the same topic was subdivided in a different way, each group studying food webs in a different kind of ecosystem: the rain forest, grasslands, oceans, fresh water bodies, or deserts.

Over the course of these investigations, students were encouraged to develop expertise and knowledge in their own interest areas, to the point where some students became class experts whose knowledge exceeded that of most of the adults. For example, one student might become acknowledged for computer expertise, another for drawing and graphics, and a third for personal expertise in a related subject matter, for instance, a child who had sickle cell disease brought related personal biological knowledge to bear in the classroom investigations. This phenomenon, which Brown and Campione (1996) referred to as "majoring," was explicitly encouraged. In contrast to typical classrooms, where the goal is for all students to know the same things at approximately the same time, teachers in FCL classrooms explicitly encouraged variability, both in what individual students knew and in the distribution of knowledge across groups.

For an extended period (typically weeks or even months), students worked in their research teams to identify and consult a variety of text and electronic resources to assist them in coming to an answer to their question. From time to time, the research teams would form "jigsaw" groups composed of one "expert" from each of the subtopic specialty teams. Within the jigsaw groups, children taught each other about their own area of expertise and attempted to coordinate their disparate knowledge into a more integrated view of the problem. Often, a culminating "consequential event" (such as a performance, design task—e.g., "Design an animal of the future"—report, or parent visit) was planned to provide the motivation for this integrative work. Occasionally, outside experts (scientists, animal care professionals) would visit the classroom to conduct "benchmark lessons" in which they introduced new disciplinary concepts or modeled thinking from a disciplinary perspective. In "cross-talk" sessions, students convened in whole-class discussions to get preliminary feedback on their progress well before the consequential event, so that they might undertake corrective action or additional investigation, if it was considered warranted. In the class discussions, all assertions that students made were considered open to legitimate challenge from any group member. Students readily learned that they were expected to be able to produce evidence and refer to at least one identifiable source to back up a contested claim. Hence, the norms in the classroom included the idea that sources were to be recruited to support arguments whose purpose was to decide among alternative explanations.

For Brown and Campione (1996) and their colleagues, designing appropriate measures was a central challenge in conducting the research. Developmental researchers have considerable experience with interviews and pre- and posttests of conceptual knowledge, but these are not usually designed to track the development of deep forms of content knowledge that emerge over an extended period of time in ways that vary considerably from student to student. Moreover, in addition to the conceptual structures of science that were the targets of instruction, the FCL program had a broader set of learning goals. For example, researchers constructed ways to track changes in students' ability to read, comprehend, and integrate textual information. They attempted to demonstrate increasing sophistication in classroom performances that are not typically assessed, such as children's scientific reasoning in their groups and whole-class discussions. In this case, they classified the forms of classroom talk that students produced and sought to observe changes in frequency of use and levels of analogies, causal explanations, uses of evidence, argumentation, and predictions.

Beyond pioneering the FCL program and conducting research on students' cognitive development, Brown and

Campione (1996) were also concerned with being able to capture the spirit of the program in a set of design principles that would serve to explain the mechanisms that sustained ongoing implementations and therefore to inform the spread of the program to new sites. This concern for principled explanation may partly have been motivated by Brown's experiences with reciprocal teaching. She noted that a weakness of reciprocal teaching and other strategy training programs is the danger that teachers and students may focus too literally on the processes of learning to the neglect of the underlying goal that motivated them. As Brown and Campione put it, "Without adherence to first principles, surface procedures tend to be adopted, adapted, and ritualized in such a way that they cease to serve the 'thinking' function they were originally designed to foster" (p. 291). In the case of reciprocal teaching, Brown and Campione observed that in its widespread dissemination, teachers sometimes focused too much on surface procedures, such as summarizing or questioning, that were not deployed for the original purpose of helping students learn to read for understanding. Sometimes these strategies were even practiced outside of the context of reading actual texts and were introduced as rituals rather than reflective strategies. It is as if the husk of the intervention had been communicated, but the germ had been left out.

Perhaps as a result of these earlier experiences, Brown and Campione struggled repeatedly through the 1990s to encapsulate and refine the design principles that motivated the FCL intervention in a way that would help the field understand both what FCL actually looked like in practice and how those systems of activity followed from their particular commitments to learning theory. For example, their 1996 chapter delineates 37 principles under six major headings: systems and cycles (a description of the recurrent activity structures utilized in FCL), metacognitive environment, discourse, deep content knowledge, distributed expertise, instruction and assessment, and community features. With some variation and adaptation, many of these features have been preserved in educational interventions that followed FCL.

With respect to science learning specifically, FCL was a sustained classroom project that attempted to identify "big ideas" in science that children might learn cumulatively, and to try to understand how those ideas might be developmentally constructed, given appropriate forms of instructional assistance. In spite of its stature and influence, two questions about FCL remain open. The first concerns the utility of principles as a way to both describe and spread new educational programs. We do not doubt that principles may help readers understand the basis for the particulars of the intervention, but we do question their sufficiency for supporting the replication and adaptation of an intervention in a new site. As yet, little is known about the content and form of knowledge that are necessary and sufficient for catalyzing and sustaining changes in teaching practices. The Schools for Thought experiment, which attempted to capitalize on what was learned through FCL and two other successful classroom-based research projects, did not generate the results and sustainability that participants had hoped (Lamon et al., 1996). Participants in this work, including Brown and Campione, found that their principles were highly meaningful to those who had generated them, but were apparently open to all kinds of interpretations to outsiders who had not shared in the background experiences that motivated the principles in the first place. Principles seemed common sense after the fact, but as a means for prescribing what to do, they did not sufficiently constrain a designer's choices. For example, although one might agree in practice that it is a good idea to encourage shared discourse and common knowledge among students (one of the FCL principles), accepting the principle unfortunately provides no guidance about how to follow it or how one could know if the goal had been satisfactorily achieved.

A second major question about FCL is whether it is a good idea for school science to be so exclusively focused on the reading and integration of textual information. Certainly reading is an important way to build knowledge in science, but arguably, students should also experience direct forms of inquiry with the natural world. Ironically, the domain-general nature of FCL activity structures and goals—something that Brown and Campione probably considered a strength—may also entail a weakness from the perspective of a particular discipline. The activities and goals in the FCL classrooms would probably apply equally well to the learning of history or literature. However, one might legitimately wonder whether learning *about* science is sufficient for coming to appreciate its epistemology. One might legitimately take the position that students should also get some experience *doing* science. Indeed, Palincsar and Magnussun (2001) have subsequently developed an approach that blends textual instruction, which they call "secondhand investigations," with direct or firsthand investigations. In their educational approach, young stu-

dents read from carefully fabricated journals that explicitly display the thinking of a scientist, who explains in the text how she conceptualized a scientific problem, used graphs and other representational devices to interpret data, or otherwise made her thinking visible so that young children could model it in their own parallel investigations with physical materials.

The Development of Model-Based Reasoning

Philosophers of science have pointed out that the central activity of science is the generation and test of models (Giere, 1988; Hesse, 1974). In fact, Giere argues that all that distinguishes scientific explanation from everyday explanation is that the former is constructed with models that have been developed in the sciences: "Little can be learned . . . about science that could not be learned more directly by examining the nature of scientific models and how they are developed" (p. 105).

Until recently, modeling practices have taken a peripheral place, at best, in school science. Even in model-populated disciplines such as physics, students' modeling activity is typically restricted to applying models developed previously by scientists, perhaps to solve textbook problems or to analyze a situation presented in a laboratory. In school, the word "model" usually denotes a noun, the product of the modeling enterprise, rather than a verb describing the practice of science. Students tend to be interpreters and users of models, but they do not generate and test them. Recently, however, scientists, mathematicians, and educators have been impressed with the potential of new computer tools to put modeling within the reach of school students. Although many investigations of modeling in mathematics and science education are focused relatively tightly on the acquisition of a specific body of disciplinary knowledge, they have also led to a more general interest in the early origins and subsequent development of model-based reasoning—a more general capability and propensity to play what Hestenes (1992) referred to as "the modeling game." In the following section, we describe two classroom-based programs that seek to foster students' capabilities to generate and test models of scientific phenomena, one at the middle school level and the second in elementary grades. The intent of these programs is to help students develop along two tracks simultaneously. First, students develop conceptual understanding of the specific scientific ideas in the domain of study. That is, students come to understand particular models and eventually to acquire a repertoire of models usable across a variety of situations. Over a longer time span, the focus is on their understanding of modeling as a key epistemology of science.

Briefly, by modeling, we refer to the construction and test of representations that serve as analogues to systems in the real world. These representations can be of many forms, including physical models, computer programs, mathematical equations, or propositions. Objects and relations in the model are interpreted as representing theoretically important objects and relations in the represented world. A key hurdle for students is to understand that models are not copies; they are deliberate simplifications. Error is a component of all models, and the precision required of a model depends on the purpose for its current use. The two instructional programs that we describe take different approaches to the forms of models that they regard as central, so we will defer further discussion about the nature of models until the examples are introduced.

Causal Models: Understandings of Consequence. Perhaps the most ubiquitous and general kind of structural relationship that can be captured in a model is the relationship of cause and effect. Causal models are ubiquitous in science, so the value of understanding the kinds of causal models that people can learn and the sources of learning difficulty seems straightforward (White, 1993). An extensive literature on the development of causal reasoning, conducted during the 1980s, suggests that even preschool children are adept at using a variety of cues from the environment to identify the cause of an event from a set of potential candidates. Among these cues are temporal contiguity, spatial contiguity, consistent covariation between the candidate cause and the effect, and mechanism, that is, whether there is a plausible mechanism that would account for A causing B (Leslie, 1984; Shultz, 1982).

Recently, Gopnik and her colleagues (Gopnik & Sobel, 2000; Gopnik, Sobel, Schulz, & Glymour, 2001) conducted a series of investigations with children as young as 2 years in an attempt to identify both *how* young children learn about new causal relations and whether these learning systems are domain specific or applied across different domains of knowledge, such as biological or physical systems. The strategy was to observe online as children went about learning a causal relation that they had not previously encountered or been taught. In one series of studies, children were introduced to the "blicket detector," a machine that lights up

and plays music when (and only when) "blickets" are placed on it. Participants were shown several small blocks and told that one or more of them were blickets. Children were asked to identify which of the blocks were the blickets, either by observing patterns of placement and the resulting outcomes and then drawing a conclusion based on those observations or, in some studies, by taking direct action themselves to place blocks on the blicket detector. Across trials within a study and across studies, the patterns of evidence that children observed became increasingly complex, ultimately including multiple causes and probabilistic relationships. In most cases, even the 2-year-olds made correct conclusions about causality by observing patterns of contingency, although these young children did not perform as well as older preschoolers on tasks in which two additive causes were required to set off the blicket detector. Children demonstrated their reasoning in multiple forms, suggesting that they genuinely were reasoning about causes, not simply making judgments of association. These forms included causal conclusions and justifications made on the basis of observation, predictions about novel events on the basis of earlier learning, and direct production of requested outcomes. Moreover, children seemed to use similar kinds of causal learning principles across different content domains of knowledge.

Gopnik and her colleagues (2001) conjectured that data-driven formal learning procedures like these might be used in conjunction with innate, domain-specific causal schemas like those described in the prior section on the "theory theory." She proposed that both kinds of causal reasoning are important and serve complementary and useful roles in children's developing knowledge. The innate theories determine what features the child is likely to attend to and, therefore, what the data-driven procedures will operate on. In turn, the formal causal learning mechanisms provide a means by which initial theories can be modified or extended, as well as a way to learn new information not implicated in a core theory (Gopnik et al., 2001). Both kinds of knowledge are fundamentally important in determining the course of learning.

Research with young children (Bullock, Gelman, & Baillargeon, 1982; Gopnik et al., 2001; Shultz, 1982) emphasizes their competence at reasoning about causally complex situations. However, the developmental literature also tells another story that seems difficult to reconcile with these findings. As often occurs in developmental psychology, findings of early competence stand side by side with studies that emphasize the reasoning flaws and biases shown by adults in situations that are described in similar ways. In this case, the developmental literature seems to conclude that very young children understand causality, whereas adults do not. For example, research conducted by D. Kuhn and her associates (D. Kuhn, 1989; Kuhn et al., 1988, 1992, 1995) demonstrates that adults frequently show characteristic flaws in reasoning about multivariable causal situations. Indeed, they make many of the same errors that children do: generating experiments that are not valid tests, interpreting evidence that is flawed or insufficient, avoiding evidence that challenges their prior theories, and failing to systematically search the space of possibilities, entertain alternative interpretations of data, or rely on evidence rather than mere examples.

Perkins and Grotzer (2000), who direct the Understandings of Consequence project, suggested that the difficulties many students have in learning science concepts stem from differences in the ways that students and scientists think about cause and effect. Nonscientists, they argue, hold a few simplistic causal structures into which all new information gets assimilated. (Similar arguments have been made by Chi, 1992.) Most of the time, these simple causal structures do a perfectly adequate job of supporting our actions and interpretations in the world, and these are the relationships that young children appear to master easily. However, when less familiar forms of cause are involved, as is often the case in science, these structures can be misleading. In contrast to novices, scientists entertain a wide array of causal structures, which vary in complexity. Perkins and Grotzer attempted to identify the features that account for this complexity and to summarize them in a taxonomy that permits estimating the difficulty of any particular causal model with respect to these features.

The taxonomy describes four aspects of causal structures: mechanism, interaction pattern, probability, and agency. Each of these varies across several levels of complexity (and, by implication, difficulty of learning). Perkins and Grotzer (2000) propose that any model or explanation can be identified on the taxonomy with respect to its hypothetical difficulty level by locating it within these four dimensions. For example, a model may vary with respect to sophistication in the level of *mechanism* that it ascribes to the phenomenon being modeled. Very simple models rely on surface generalizations or explanations at the same level of description as the

events being explained. At the more sophisticated end of the spectrum, a model may appeal to analogical mapping or underlying mechanisms, including properties, entities, and rules that account for the situation at an underlying level of description. Similarly, simple *interaction patterns* include those that appeal in a straightforward way to one thing acting on another, via pushes, pulls, supports, resistances, and so on. The entities on this level seem similar to the simple schemas that diSessa (1993) referred to as phenomenological primitives, that is, schemas at a midlevel of abstraction that are automatically activated to support interpretations of physical events. According to diSessa, these interpretations seem self-evident and do not require justification; instead, people simply "recognize" an event as belonging to one class or another. At more complex levels, students may entertain mediating causes, interactive causality, feedback loops, or constraint-based systems. The dimension of *probability* specifies whether a particular explanation is deterministic or appeals to chance, chaotic systems, or fundamental uncertainty. The final dimension concerns the perspective taken on *agency:* Does the model assume that a central agent is the causal actor, or are other, more complex possibilities considered, such as additive causes, long causal chains, self-organizing systems, or emergent properties? At this point in the research, the taxonomy should probably be considered hypothetical; the dimensions of complexity were derived via rational analysis rather than empirical test. Moreover, the taxonomy appears to capture only *order* of complexity, not *degree;* there is no claim that difficulty level increases in measurable quantitative steps from the least to the most complex level of each dimension. Also, it is not clear how to cumulate these dimensions to make a judgment about the overall complexity of a model. The best use of the taxonomy at this time seems to be heuristic, and the authors do not comment on whether they consider it to have scale properties.

Along with their analysis of models and model explanations, Perkins and Grotzer (2000) have also developed an analysis of what they call *epistemological moves* toward better models. These are the cognitive behaviors with respect to modeling that they find worthy of encouraging in students. They include seeking a model with no gaps or missing parts, putting the model at risk by actively seeking counterevidence or contrasting cases, detecting flawed evidence, and entertaining reasonable criteria for revising or replacing the model in the face of different forms of counterevidence. These epistemological moves are similar to the criteria for changing theories that were delineated in the Conceptual Change model, described earlier. Presumably, acquiring and using these epistemological moves increases the likelihood that students will come to understand and, when appropriate, apply the most appropriate causal schema from their repertoire.

Initial research findings suggest that students who participated in activities that emphasized the underlying causal structure of a scientific topic and participated in direct discussions of these causal relationships performed better on measures of conceptual understanding of that topic than did students who worked on similar units that did not directly emphasize causal relationships (Grotzer, 2000; Perkins & Grotzer, 2000). However, the project had ambitions beyond simply boosting conceptual understanding domain by domain. In particular, the goal was that students who learned about causal structures in one topic (e.g., density) would transfer those structures to other topics (e.g., pressure) when it was appropriate to do so, and that importing the new causal structures would provide a firmer base for understanding the new material. Research to this point (Grotzer, 2003) suggests that there is some limited transfer of this kind from one topic to another when the causal structure in both tasks is isomorphic. However, the researchers found no evidence of spontaneous transfer when the causal structures between the two topics were not isomorphic. In other words, so far there is no evidence that students have acquired a general propensity to search among candidates for an appropriate causal model and then try to use it to understand novel cases. Grotzer (2003) observed that situation-specific default concepts like diSessa's (1993) phenomenological primitives seemed to interfere with transfer of the appropriate causal relationships. The investigators are now seeking to enhance the metacognitive aspects of the instruction in an effort to learn whether more explicit reflection on the nature and uses of causal models might help improve the transfer of causal structures between science contexts.

An attractive feature of causal models is that they have both a domain-general aspect, derived from the general structure of the causal relationship that is expressed, and a domain-specific aspect, in that the relationship represents structure in a particular domain or situation (Gopnik et al., 2001). Because of this integrative quality, modeling approaches at least hold the potential of avoiding the process/content or syntax/

substance dichotomies that sometimes plague science education (and psychological accounts of scientific reasoning). The causal modeling approach being developed by Perkins and Grotzer (2000) may be described as a top-down modeling approach. Through rational analysis, the investigators first attempted to exhaustively describe the landscape of kinds of causal models, and content domains for study were apparently selected because they exemplified one or more of these target forms of causal reasoning. It is not clear whether considerations of conceptual development guided the selection of domain topics beyond a commitment to generating opportunities for the acquisition, transfer, or comparison of causal models. Therefore, in this program, development of scientific conceptual knowledge is probably a more important focus within units rather than across domains. Over years of a student's education, the acquisition of a repertoire of causal schemas takes priority as an educational objective over the development of any particular conceptual knowledge base.

Of course, when scientists construct, test, and revise models, they do so in the service of contributing to a base of knowledge within a coherent content domain. The final classroom research program we describe, our own, aims to open the activity of modeling to school students. It integrates Perkins and Grotzer's (2000) emphasis on refining a repertoire of structural analytical tools with the focus on conceptual development within a coherent domain that is favored by investigators like Gelman and Brenneman (2004) and Metz (2004).

Modeling Nature. The kinds of models that scientists construct vary widely, both within and across disciplines. Nevertheless, the rhetoric and practice of science are governed by efforts to invent, revise, and contest models. We (Lehrer & Schauble, 2005) have been investigating the implications of this view of science for the education of students in elementary and middle school grades. Our primary interest was not just on students' understanding of models per se, but, more specifically, on their understanding of modeling. To provide a context where the development of model-based reasoning could be studied, participating teachers worked collaboratively and systematically to build on young children's interests and abilities in representing aspects of the world in all kinds of ways—via language, drawings, physical models, maps and globes, rules that capture regularities and patterns—and to provide effective forms of instructional support, building on children's initial modeling attempts to help them achieve a progressively more sophisticated grasp of science. Early emphasis on representational form, especially on purposes and uses, was derived from developmental studies that suggested a rich repertoire of such resources (e.g., Karmiloff-Smith, 1992) and from social studies of science, which indicated their critical role in model building (e.g., Latour, 1999). We were especially interested in those forms of representation that would help children "mathematize" (Kline, 1980) natural phenomena, such as growth or relations between structure and function. By mathematizing, we mean the common scientific practice of quantifying or visualizing phenomena geometrically (or both). Privileging mathematics meant introducing mathematics to elementary children that went beyond arithmetic to include space and geometry, measurement, and data/uncertainty (e.g., Lehrer & Chazan, 1998; Lehrer & Schauble, 2002). The focus of the research that was coordinated with this instructional agenda was on the early emergence and subsequent development of model-based reasoning. A secondary agenda concerned students' conceptual development in target forms of mathematics and science.

The developmental literature illustrates that there are myriad ways in which even preschool children come to regard one thing as representing another. This representational capacity provides roots for the development of a modeling epistemology. For example, long before they arrive at school, children have some appreciation of the representational qualities of pictures, scale models, and video representations (DeLoache, 2004; DeLoache, Pierroutsakos, & Uttal, 2003; Troseth, 2003; Troseth & DeLoache, 1998; Troseth, Pierroutsakos, & DeLoache, 2004). In pretend play, children treat objects as stand-ins for others (a block stands in for a teacup, a banana for a telephone), yet they still understand that the object has not really changed its original identity, character, or function (Leslie, 1987). Later in school, they will capitalize on very similar understandings to use counters for "direct modeling" to solve simple early arithmetic problems that involve grouping and separating.

However important, these early symbolic capacities do not yet capture all the key aspects of a scientific modeling epistemology. Although they certainly know the difference between a model and its referent, children do not usually self-consciously think about the separation of the model and the modeled world. Consequently, they often show a preference for copies over true models, because they tend to resist symbolic depic-

tions that leave out information, even if the information is not important to the current theoretical purposes (Grosslight, Unger, Jay, & Smith, 1991; Lehrer & Schauble, 2000b). For example, children using paper strips to represent the height of plants may insist on the strips being colored green (like the plant stems) and demand that each strip be adorned with a flower (Lehrer & Schauble, 2002). Students are unlikely to spontaneously consider issues of precision and error of a representation or the implications of deviations between the model and the modeled world in light of current goals (although they certainly have intuitions that are helpful as starting points; see Masnick & Klahr, 2003; Petrosino, Lehrer, & Schauble, 2003). Having identified a way to represent one or more aspects of the world, they may be unable to entertain the possibility of alternatives. Indeed, the search for and evaluation of rival models in evaluating alternative hypotheses is a form of argument that does not typically emerge spontaneously (Driver, Leach, Millar, & Scott, 1996; Grosslight et al., 1991).

In addition to these general symbolic capacities, the development of specific representational forms and notations is also a critical part of being able to enter what Hestenes (1992) referred to as the "modeling game." Representational tools such as graphs, tables, computer programs, and mathematical expressions do not simply communicate thought; they also shape it (Olson, 1994), so acquiring a vocabulary of inscriptions and notations and a critical understanding of their design qualities was considered essential. Accordingly, helping students develop their metarepresentational competence (diSessa, Hammer, Sherin, & Kolpakowski, 1991) was a central target of both instruction and the related research.

Particular emphasis was placed on mathematics as a tool that both describes the world and serves as a resource for meaning making (Lehrer, Schauble, Strom, & Pligge, 2001; E. Penner & Lehrer, 2000). Often, science educators delay the mathematization of scientific ideas, believing that students should first develop a qualitative analysis of the science underlying the phenomenon, and that too early attention to mathematical description may encourage an emphasis on computation rather than understanding. This assumes, however, that students have no history of learning mathematics as a sense-making enterprise. Experience and research suggest that this need not be the case. With good instruction even young students can meaningfully consider the epistemic grounds of generalization and even proof (Lampert, 2001; Lehrer et al., 1998; Lehrer & Lesh, 2003). These epistemic considerations often arise when children investigate the mathematics of shape and form, measurement, and data. Therefore, developing and testing appropriate inroads to these new mathematical ideas was an important part of the program (e.g., Lehrer & Chazan, 1998; Lehrer, Jacobson, Kemeny, & Strom, 1999; Lehrer & Romberg, 1996; Lehrer & Schauble, 2000c, 2005). If they are lacking these mathematical resources, it is unlikely that students' conjectures can be held accountable in any meaningful way to data, which has mathematical qualities that need to be appreciated if their interpretations are to be disciplined. The aim was to develop students' mathematical understanding to the point where it would be sufficient to support description and systematization of the natural world—the heart of modeling.

In the science class, we attempted to orient instruction around a cumulative focus on important core themes, such as growth and diversity, behavior, and structure and function, as described in national science standards (National Research Council, 1996). Themes were selected in part for their centrality to science disciplines, but also for their potential for engaging students in the progressive mathematization of nature (e.g., Kline, 1980). Central concepts such as diversity and structure derive their power from the models that instantiate them, so to fulfill the promise of the "big ideas" outlined in national standards, students must realize these ideals as models. Moreover, models are not simply constructed; equally important, they must be mobilized—that is, put to work—to support socially grounded arguments about the nature of physical reality (Bazerman, 1988; Latour & Woolgar, 1979; Pickering, 1995). Achieving these goals with school students meant identifying forms of modeling that are well aligned with children's development.

We concluded that in children's instruction, it is advisable to begin with models that resemble their target systems (i.e., the phenomena being described or explained) in ways that can be easily detected, because resemblance helps children make and preserve the mappings between models and their referents (Brown, 1990; Lehrer & Schauble, 2000c). For example, when first graders were given a variety of materials from a hardware store and asked to construct a device that "works like your elbow," initial models were guided by a concern for copying perceptually salient features (Grosslight et al., 1991). Most of the children insisted on using round foam balls to simulate the "bumps" in their

elbow joints and Popsicle sticks to simulate fingers (D. E. Penner, Giles, Lehrer, & Schauble, 1997). However, this beginning concern with "looks like" lost importance over multiple revisions of the models, which eventually began to focus on "works like": relations among and functions of components in the target system, in this case, ways of constraining the motion of the elbow. Consistent with the emphasis on mathematics in children's modeling activity, third graders went on to mathematically explore relationships between the position of a load and the point of attachment of the tendon in a more complex elbow model (D. E. Penner, Lehrer, & Schauble, 1998).

Modeling is a form of disciplinary argument, one that students learn to participate in over a long and extended period of practice and only with good teaching assistance. Lehrer and Schauble (2005) argued that acquiring disciplinary forms of argument requires emphasizing students' long-term development of central conceptual and epistemic structures, not the acquisition of nuggets of instruction that are delivered within brief periods. Decisions about what is taught should be informed by a long-term view, one that regards learning as a historical activity in which current learning builds from and on learning achieved in earlier weeks, months, and years. Therefore, the research focused on identifying and empirically testing science themes that provide easy entry for young children, while supporting plenty of conceptual challenge for students in the upper grades. Identifying mathematical and scientific models and concepts that could potentially serve as a core and then working with teachers to investigate the potential of these ideas across grades of schooling constituted an important part of the design research agenda.

An example is the theme of growth and change. Students in primary grades represented the growth of flowering bulbs planted under different conditions (in soil or water), using paper strips to depict the heights of plant stems at different points in the growth cycle (Lehrer, Carpenter, et al., 2000). Depiction of height required a transformation in children's thinking from considering the plant as an intact whole to thinking of it as a set of attributes, height being the most salient. Representing and comparing heights required working out standard ways of measuring and a firm understanding of the mathematics of measure, which was developed systematically during this investigation. Indeed, it is worth noting that understanding an attribute and understanding how to measure it are related ideas, regardless of the grade of the "scientist." When children raised the question *how much faster* one plant grew than another, their attention turned from comparing final heights to noting successive differences in the lengths of the strips from day to day. These questions relied on the arithmetic of comparative difference, a form of mathematics within their grasp. They noted that the amaryllis grew faster at the beginning of the life cycle and then slowed, whereas the paperwhite narcissus grew very slowly at the beginning and then "catched up."

In the third grade, students investigated change of Wisconsin Fast Plants™ in a variety of ways (Lehrer, Schauble, Carpenter, & Penner, 2000). (Wisconsin Fast Plants, or *brassica rapta,* complete an entire life cycle in about 40 days, making it feasible to use them in population studies or other classroom investigations that require comparisons of groups of plants that can be readily grown within one school semester.) They developed pressed plant silhouette graphs that recorded changes in the plants over time, coordinate graphs that showed relations between plant height and time, rectangles that represented relationships between plant height and canopy "width," and three-dimensional prisms and cylinders to capture changes in plant volume. These diverse representations raised new questions about the plants. Students wondered whether the growth of roots and shoots were the "same" or "different." They concluded that the rates of growth were different at similar points in the life cycle, but that the general shape of growth (S-shaped logistic curves) was similar. Why, students wondered, might the growth of different plant parts have the same form? When was growth the fastest, and when the slowest, and what features in the plants were changing in ways that might account for this? Teachers played a central role in helping students compare and evaluate their questions, produce and contrast different kinds of representational displays, and generate evidence-based claims. Although it is not possible to include detailed information here about the data on teachers' professional development and changing teaching practices, these were necessary conditions for the student learning that was observed (more information on this aspect of the program is provided in Lehrer & Schauble, 2000c, 2005).

In the fifth grade, students compared populations of plants and reasoned about features of distributions of the plant measurements to decide whether growth factors such as fertilizer and amount of light were affecting variables such as height and reproductive capacity (i.e., number of seeds and seed pods) of the plants (Lehrer & Schauble, 2004). Features of distributions such as typicality and spread were investigated thoroughly, and dif-

ferent representations of these statistics were invented and explored. Sampling experiments based on the students' measurements of plant height at a particular day of growth supported discussions about typical plant height and its variability under different numbers of samples and samples of different sizes. Children learned to read the shape of different distributions as signatures of growth processes. For example, a distribution with a left wall was interpreted as representing the plants early in their life cycle because, as one child explained, "You can't get any shorter than zero mm."

As these examples illustrate, at each grade children's representational repertoires were systematically stretched, making it possible to expand their knowledge about growth and change in new ways. In turn, as their knowledge grew, there was change in children's considerations about what might next be worthy of investigation.

Lehrer and Schauble (2000c, 2003, 2005) reported observing characteristic shifts in understanding of modeling over the span of the elementary school grades, from an early emphasis on literal depictional forms to representations that were progressively more symbolic and mathematically powerful. Diversity in representational and mathematical resources both accompanied and produced conceptual change. As children developed and used new mathematical means for characterizing growth, they understood biological change in increasingly dynamic ways. For example, once students understood the mathematics of ratio and changing ratios, they began to conceive of growth not as simple linear increase, but as a patterned rate of change. These transitions in conception and inscription appeared to support each other, and they opened up new lines of inquiry. Children wondered whether plant growth was like animal growth, and whether the growth of yeast and bacteria on a petri dish would show a pattern like the growth of a single plant. These forms of conceptual development required a context in which teachers systematically supported a restricted set of central ideas, building successively on earlier concepts over grades of schooling.

Learning research was conducted to investigate the development and use of a variety of mathematical and scientific models. One strategy was to conduct detailed studies of student thinking in the context of, or immediately following, particular units of study. The purpose of these investigations was to learn whether and how students developed new models, to identify the variability in student understanding of the mathematical and scientific concepts at hand, and to document how students appropriately applied mathematical concepts learned in one context to novel situations. For example, in one study, students explored the mathematics of ratio via geometry by investigating the properties of families of similar rectangles (Lehrer & Schauble, 2001). Subsequently, while investigating properties of materials, they spontaneously wondered whether materials might also come in families, a reference to whether there might be constant ratios between volume and weight for objects made of Styrofoam, wood, Teflon, and brass. Pursuing this question led to an extended investigation of the properties of coordinate graphs and linear relationships as models (the plots of weight by volume seemed nearly linear, but many of the points did not lie directly on the line). Lehrer and Schauble conducted numerous classroom investigations of student model-based reasoning in the context of instruction in mathematics (e.g., data modeling, classification, distribution, similarity) and science (e.g., growth, diversity, motion, density). Details of this work are reported in a variety of publications (Horvath & Lehrer, 1998; Lehrer, Carpenter, et al., 2000; Lehrer & Schauble, 2005; Lehrer, Schauble, & Petrosino, 2001; Lehrer, Schauble, Strom, et al., 2001; D. E. Penner et al., 1997, 1998). Most of these investigations were cross-sectional; they either focused on students within a classroom or classrooms at the same grade or drew comparisons of the performance of students in different grades at the same time point.

In addition to these within-grades and between-grades studies, longitudinal investigations were conducted to confirm whether and, if so, how students' understanding of mathematics was growing systematically over years of instruction, because mathematics was the primary tool employed for modeling. Because students were learning forms of mathematics that are not typically taught in elementary grades or measured by current standardized assessments, the project team created a series of standardized measures to assess student achievement, organized into a 3-hour test that could be administered to groups of students. There were two forms for this instrument, one for the primary grades and the other for upper elementary grades. Each form was revised every year, although a core pool of items was administered each year to all students. Several released items from the National Assessment of Educational Progress were included to benchmark student achievement to national performance. The results, reported in detail in Lehrer and Schauble (2005), found gains in student learning that were reliable at each grade from grades 1 through 5 (effect sizes ranged from 0.43 to 0.72). The average gain scores indicated substantial

growth in student understanding, and the gains were widespread (i.e., not confined to selected strata of students). Moreover, on the nationally benchmarked items, students in the early grades outperformed those from much higher grades in the national sample.

Of course, there is much yet to be learned. One issue is the relationship between mathematics and science. We generally first introduce students to mathematics, so that they have opportunities to explore and understand mathematical structure before these structures are employed to model nature. We are concerned that if we introduce only the mathematics that students need to model a particular system, then much else about the mathematics will be lost (e.g., its more general, systematic quality). However, this approach clearly contrasts with curricular approaches that emphasize integrating mathematics and science. A related issue is how students view epistemologies within each discipline. For example, in some of our classroom studies, we have noted children drawing clear distinctions between mathematical (e.g., general by definition) and scientific (e.g., general by model) senses of generalization (Lehrer & Schauble, 2000a). How these epistemologies unfold over time is not yet understood.

Summary: Classroom Design Studies

In this section, we reviewed seven extended programs of classroom research in which researchers studied the development of student thinking in contexts that were engineered to support it. Although these are by no means the only developmentally informed investigations of scientific thinking in classrooms, they do represent a range of visions about what scientific literacy should entail. Each vision was either consistent with or directly informed by related research in cognitive development. Many, although not all, of the scholars who conducted this work also articulated an explicit perspective on the relationship between learning and development.

In the chapter's introduction, we claimed that new answers to the questions "What develops?" and "What is development?" were being raised within this niche of classroom-based developmental research. We next briefly summarize what these investigations, taken as a group, suggest about potential answers to these two questions.

With respect to their views of science and science literacy, all of the investigators reviewed in this section acknowledge the complexity and variability of science. The focus on what develops is necessarily much broader than in typical studies of learning and thinking, which appropriately tend instead to focus tightly on particular skills or concepts. This broader focus is necessary, of course, for seeking to characterize and understand development that occurs only over years of education. The wider perspectives taken here may also be useful for considering the implications of more traditional research on scientific thinking with respect to the goals of education. For example, consider Chen and Klahr's (1999) research on the control of variables strategy in juxtaposition to Metz's (2004) broader agenda of helping students conduct self-initiated and self-regulated inquiry. Both studies share a focus on helping children understand the logic and methods of research, yet they do not come to the same conclusion about what should happen in classrooms. Indeed, one of the unresolved issues in science education is this disagreement about whether children should first explicitly be taught strategies and procedures for conducting inquiry and then later learn to apply them, or whether they should learn these strategies and methods in contexts of their use, so that they are situated within a larger, coherent process of inquiry. This question takes on special poignancy when the children are struggling students or come from cultures where they have had less exposure to forms of thinking valued in school. Lee and Fradd (1996) have argued that in these situations, it is important to directly instruct children first on processes and strategies of inquiry, so that they do not come to science instruction with a disadvantage. In contrast, Warren and Rosebery (1996) have emphasized the many points of contact between everyday thinking and scientific thinking, which seem to hold for all children, even those whose first language may not be English and whose first culture may not be Anglo-European. In their view, with sensitive instruction children are quite capable of sophisticated forms of inquiry, and the evidence seems to bear out these claims. It may be, however, that the dispute is more apparent than real. The need to be explicit and clear about the forms of argument and evidence valued in science is widely accepted, and there is plenty of evidence that this need is not restricted to students who are struggling. The reason for contrasting Chen and Klahr's position with Metz's is not to suggest that one conclusion necessarily is associated with psychological research and the other with classroom research. It is to make the more general point that in many cases, taking the wider view that an educational perspective demands, leads to a realignment of what is valued, so that design researchers are not simply involved in bringing together

in one site interventions that have individually been more thoroughly studied in psychology laboratories.

Whether one thinks it is more useful for purposes of instruction to highlight science as building knowledge or theories, conducting investigations, or generating and testing models, these are probably best regarded as partially overlapping rather than mutually exclusive views of science and science literacy. They may lead to somewhat different commitments with respect to choices of topics of study or classroom activities. Regardless, theory change, inquiry, and modeling are mutually reinforcing. Therefore, any well-formulated program may focus on all of these goals, even though the relative emphasis or proportions of time spent on each may differ. Similarly, there may be differences in what teachers are oriented toward in professional development, perhaps leading to discernibly different results in teachers' practices and student learning. At this point, we do not know.

There is both a normative (What *should* students be learning?) and empirical (How *does* development typically unfold?) aspect to these guiding perspectives. Taking a longer-term developmental view raises questions about what educators should be trying to achieve in the long term and also about the instructional pathways that can best lead students toward these goals from their current conceptual resources. Ideas about instructional pathways should be conceived as rational analyses that require empirical testing. It is impossible to know in advance how students' cognition is likely to develop given the right kinds of instructional support, partly because we cannot know in advance which kinds of instruction are optimal and partly because our initial views of students' capabilities almost always are distorted by knowing the way they usually perform under typical (or lacking) instructional conditions (Brown & Campione, 1996).

For the most part, the research reviewed in this chapter reflects a preference for students doing science over simply learning final-form science concepts. This preference is due not to a naive belief that knowledge is somehow better if it is reinvented by students, but to a commitment to providing opportunities for students to experience one of civilization's most powerful forms of epistemology. We would probably agree that all students should learn to write to some level of fluency, even though few will eventually become employed as professional authors. Similarly, all students should get a taste of doing science, and those opportunities should not be restricted to those bound for careers in science or technology.

The emphasis on doing science, however, does not imply that nobody cares if students learn any scientific knowledge. Without exception, the emphasis in the programs we have reviewed is on doing science for the purpose of building rich, elaborated bases of knowledge. That is why the programs reviewed in this section value extended study within a bounded content domain over broad sampling of science topics. Focusing deeply in a domain provides a base from which students can develop criteria for evaluating their changing theories about the domain and also provides the foundation of knowledge necessary for inquiry to be both fruitful and meaningful. Not all of these researchers, however, have a clear vision of how science content knowledge is expected to cumulate over a student's education, or even whether having such a vision is considered important. Some investigators (e.g., Metz and Gelman) expect that students will sequentially investigate domains of study in depth, one at a time, but they do not say much about what space of domains needs to be visited by the time a student leaves elementary education. Lehrer and Schauble seek scientific and mathematical themes, such as growth, structure and function, and behavior that can connect inquiry across years of schooling. These themes serve as the criteria for selecting specific topics of study. However, Lehrer and Schauble (2005) argue that it is necessary to empirically test conjectures about the themes that best permit easy entry to younger or less sophisticated students and, at the same time, provide abundant curricular challenge for those who are more knowledgeable. Hennessey and Grotzer and Perkins appear to be focused primarily at a more domain-general level on causal schemas and criteria for conceptual change. Presumably, domain knowledge is selected for its exemplification of the variety of causal schemas that students need to learn about or its potential to highlight criteria for theory change.

To varying degrees, all of these investigators place instructional emphasis on one or another form of metacognition. That said, what is meant by metacognition varies somewhat from program to program, and the actual cognitive processes involved may have little or nothing in common. Brown and Campione generally encouraged students to assume responsibility for their own learning, a goal that Metz also adopted but applied in a more focused way to student planning and conduct of empirical investigations. As we have seen, Hennessey wanted students to understand and apply specific evaluative criteria to their own theories and the theories of classmates. This is a view of metacognition that seems more closely related to the one articulated by Grotzer

and Perkins, who expected students to notice and describe the causal structure underlying content domains whose surface features varied. Lehrer and Schauble regarded metacognition somewhat differently, as learning to use varying forms of representation that allow one to literally grasp thought, and as putting these representations to use in the service of arguments about qualities of natural systems.

There was widespread agreement on the importance of data representation and other forms of symbolization. Many of these researchers endorsed the value of capitalizing on the variability in students' invented representations. Repeatedly producing, critiquing, and revising representations helps students appreciate the uses and purposes for inscriptions, what they communicate, and the design trade-offs entailed in their construction. In traditional classrooms, students are taught conventional forms for graphing, making tables, drawing maps, and the like, as context-free tools. They may be given a variety of problems to practice on, but these are regarded merely as contexts to serve the primary goal of learning how to construct and use the inscription in its conventional form. In contrast, a theme common to the programs we reviewed is tying education about forms and uses of representation and inscription to contexts of their use. Other tools as well, from scientific instruments to rulers, are introduced when students have encountered a problem that the tool would be helpful in addressing.

Views on the nature of development emphasize continuity from children's early intuitions and theories to their instruction in conventional theories of science disciplines. In distinction to the misconceptions literature in science education, which tends to draw sharp contrasts between students' conceptions and those of experts, these investigators see early theory building as a resource for rather than a barrier to instruction. Attention to the features of learning contexts that optimize development is considered an essential part of an account of development for these researchers, although the extent to which they focus on cataloguing these features varies somewhat. Brown and Campione, with their 39 principles, are probably most exhaustive in their attempt to specify the features that account for developmental change in a classroom context. Across the six projects, a range of features was proposed that varied from the kinds of tasks presented to students (all of the researchers in this section) to the forms of activity repeatedly engaged, the classroom norms, and the kinds of evidence and argument that characterize classroom discourse.

We have briefly described seven classroom design studies organized around investigation of the development of some aspect of scientific thinking. Each was grounded in a particular vision of what develops in scientific thinking and literacy, and each provides at least initial data about the learning potential of the program. However, at this point in time, none of these projects has secured a base of longitudinal research that is extensive enough or has been sustained for a long enough period to permit clear comparisons about the long-term educational consequences of pursuing one design rather than another. We still know little about what we might expect of a student who participates in one of these programs for an extended period. What capabilities or propensities would this student develop, and what forms of practice would he or she master that graduates from the other programs might not? From a design perspective, the point of having longitudinal comparative data would be not to find out which approach is best, in the simple sense of winning a horse race, but to better understand the characteristic profile of strengths and weaknesses of each, so that choices about educational directions can be informed by their fit to more clearly articulated values. Do some of these programs provide a smoother transition to becoming a generally literate citizen, whereas others provide a better pathway to the professional practice of science? Do some do a better job than others of providing foundational tools that will pay off consistently over the scope of a child's education? What does each approach emphasize, and what does it tend to move to the background?

We now know something about how education starts off under these approaches and a little about how it proceeds, but we know little or nothing about how it ends up many years down the road. In the final section of the chapter, we seek to understand what it takes to build and sustain conditions that permit the acquisition of comparative data of this sort. This question is pursued in the context of discussing the implementation challenges of conducting classroom design research.

IMPLEMENTATION ISSUES IN DESIGN STUDIES: WHY AREN'T WE MAKING FASTER PROGRESS?

Why is it so difficult to conduct the kind of longitudinal, comparative work that can inform educational decisions

about science literacy in a systematic, scientific way? There are both conceptual and logistical challenges to developing and refining educational programs that are informed by developmental theory and research, sustaining those programs in ways that preserve and extend their educational integrity, and assessing learning in organizational systems that are both highly changeable and politically sensitive. Rather than discussing these implementation issues in general, we will view them through the lens of our own work. As explained earlier, information about these matters is seldom openly discussed. Therefore, we resort here to our own experience, trusting that it is more common than uncommon.

Challenge 1: Developing and Refining the Design

Although previous and concurrent research can be of help in identifying likely starting points for children's learning, learning research insufficiently constrains educational design. A significant amount of conceptual and empirical work is required to develop and refine an educational design that can foster long-term development. The more extensive the target of educational concern, the more conceptual and empirical work is required to cash in, test, and revise the elements of the educational design. Careful consideration of what is to be done, day by day, does not follow obviously and smoothly from a few key principles or even from hypothetical prospective trajectories of student learning. For instance, deciding that we intend to support the development of model-based reasoning in children, that we will seek to build on early origins of this form of thinking, and that we will systematically provide mathematical resources, representational tools, and appropriate classroom norms still leaves us with the need to make day-to-day decisions about how to accomplish these goals. If the means are wrong, it will not matter if the principles are right.

The instantiation of an educational design routinely requires the revision of initial plans and assumptions. Students have a way of getting stuck on forms of learning that seem relatively straightforward until one tries to help children achieve them, or to the contrary, of readily producing forms of thinking that seemed unlikely on first consideration. At key points during instruction, it is necessary to be able to predict the near landscape of educational possibilities most likely to unfold and to foretell the consequences of following one or another path through this landscape (Lehrer & Schauble, 2001). Developing this kind of knowledge requires replicating the "same" lesson sequences—while exploring key variants—on multiple occasions and often at different grades. Cross-grade study helps us better understand both what is developing and the likely pathways of development.

For example, we deliberately adopted a developmental focus with the previously described study of data classification (Lehrer & Schauble, 2000b), in which children developed models to predict the age of the artists of a series of self-portraits. This investigation was conducted in grades 1, 4, and 5. The first graders readily classified the portraits by the presumed grade of the artist and identified the features that they felt differentiated the pictures drawn by kindergarteners ("dinosaur" hair, no feet) from those drawn by fifth graders ("lots of detail," all five fingers). However, their classification systems were merely post hoc descriptions applied to decisions they had already made via casual inspection. Tellingly, they did not use their feature lists to make predictions about a set of novel portraits. Therefore, to the first graders, the lists did not really serve as models at all. In contrast, the fourth graders did develop models and apply them to support predictions, but it took many attempts to use the models and rounds of subsequent revision before students came to prefer models that did not include extraneous detail. These fourth graders struggled with the idea that a model that did not include all discernible information about a portrait might be preferable to one that did. Fifth graders not only eliminated features that were not predictive from their models; they even developed quantitative estimates of the predictive power of their features ("A portrait drawn by a fifth grader is twice as likely to have eyelashes as it is to have shoes with shoelaces"; "Two thirds of the time, a fourth-grade portrait will include eyelashes").

To the extent feasible, we replicate instructional sequences to understand more about what is repeatable, what varies, and what routes development typically takes. Our purpose is to achieve a clearer understanding of what constitutes the intervention. That is, what is essential to produce desired outcomes and what is peripheral? What variations in features still produce similar results, and what forms of variation fundamentally change the character of the outcomes? What is the permissible window of variability of each key feature within which we would judge that the intervention maintains its integrity? Failing to understand these issues,

we believe, accounts for much of the difficulty experienced in attempting to "scale up" educational interventions—much of the time, what is being "scaled" is only dimly understood. For this reason, we seek to understand the generalization (and generativity) of a pathway of learning by investigating how lesson sequences play out with a variety of different student and teacher populations. We attempt to replicate within and across grades in a participating school, across schools in a district, and across sites. Portions of our work have been replicated in both suburban and urban school districts in the upper Midwest, in Phoenix, Arizona, and currently in Nashville, Tennessee. Yet replicating educational interventions that extend over several years is a very slow process, one that should be pursued before comparative trials are undertaken. At a minimum, they involve considerable challenges in assisting teachers' professional development to a level where the intervention can be reliably produced. Treating a program as if it were transparent to teachers is an invitation to the kinds of lethal mutations discussed earlier.

Challenge 2: Implementing and Sustaining the Program and Its Integrity

So far, we have been discussing the conceptual challenges involved in identifying the defining features of an educational program. There are equally daunting logistical challenges, which require solutions that are every bit as intellectually demanding. These solutions are costly in terms of both researcher time and resources, and our training typically does not equip us to address problems of this kind. First among the implementation challenges is the difficulty of marshaling and maintaining capacity to do this kind of work within our own organizational setting, in this case, the university.

The education of graduate students poses challenges. Rather than introducing students in a gentle way to well-understood and routine procedures, we must help graduate students learn within and make productive contributions to an enterprise that is under continual evolution. One is always updating newcomers of all kinds (staff as well as students) to an ongoing effort that existed before they came and will extend beyond their tenure. Participants at all levels need to continually recalculate the relationships between the part of the project in which their contributions are made and the larger enterprise in which it resides. These features of the research sometimes generate difficulties for the indoctrination and socialization of new students into this form of research.

Classroom design research requires interdisciplinary teams and multiple forms of talent that are unlikely to reside within one individual. We have found it helpful to form collaborations with individuals from other disciplines: in-service teachers and school administrators, of course, but also biologists, mathematicians, and psychometricians. Identifying and coordinating multiple participants and forms of expertise over extended time periods is a goal that does not always align well with the expectations of university promotion and tenure committees, resources and cycles of funding agencies, or colleagues' existing disciplinary allegiances. We have needed to play multiple roles ourselves, including educator, professional development provider, and community politician, in addition to education researcher.

Sometimes these roles involve managing contingencies as they emerge and cannot be identified in principle beforehand. For example, our decade-long program of work in a school district was preceded by a decade of work that one of us conducted in classrooms in that district. This earlier work involved coming to be seen as a member of the school district by teachers, administrators, and parents. It entailed countless conversations "in the cracks" that gradually built trust, so that stakeholders, especially teachers, did not perceive research as something done to them and their children. Some of these events might be viewed as extraordinary, even bizarre, from some perspectives. For example, one parent was concerned that the screen image of the Logo programming language might be a form of idolatry prohibited by her religion. Concerns like this were not anticipated by the researcher but, nonetheless, had to be addressed in ways that preserved the integrity of all concerned. An outcome of this previous work was increased capacity for teacher leadership, so that teachers were prepared to build on the changes they had already begun. This preparation served as an essential foundation to the research we described; without it, it is highly unlikely that we would have been able to achieve significant levels of student learning within a 3-year period. Hence, this history proved relevant to the conduct of the research program, but it also raised the problem of identifying which aspects of history should be judged relevant when reporting current design research.

Schools, of course, are daunting organizations in which to pursue research, especially if they are organized around an educational change agenda. This is par-

ticularly the case in today's climate of politicized education. The leadership in most districts is unstable, schools are vulnerable to all sorts of competing political pressures, and their goals and activities are publicly contested. In our work, we have struggled with an array of havoc-producing events, including the resignation of a supportive superintendent, a shift in the school board's political affiliation, the serious illness of a teacher-leader's young child, internal disagreements among the faculty (e.g., over whether to pursue *looping*, in which teachers graduate with their students across one or more grades, or multiage classes). We are confident that other classroom researchers have similar tales to tell. The legitimate agendas of schools often inadvertently put them at cross-purposes to the goals of the research. At one site, we were making good progress at consolidating a cross-grade team of like-minded teachers who had worked for several years together on professional development oriented around the study of student learning. Over the years, the group had achieved strong community affiliation and had amassed impressive technical knowledge about the development of student thinking, achievements that were central to our shared goal of supporting a systematic and consistent approach to mathematics and science education. However, this district was one of the fastest growing in the state. As the district expanded, it became necessary to build a new elementary school. To our dismay and that of the teachers, administrators moved several of the participating teachers to the new school to colonize the reform in this new site. Although the intentions were noble—administrators hoped to see these new forms of teaching spread more widely—the result was the disruption of the cross-grade community and our capability to follow students longitudinally across grades in which the experimental instruction was being implemented. Even when radical changes of this kind are not occurring, the degree of teacher and student mobility that is typical of American schools makes longitudinal research difficult to sustain.

Within the past several years, we have found the politics of education to be especially disruptive to any agenda that includes systematic capacity building. Lack of consensus over the role and form of education leaves teachers highly vulnerable to disagreements about standards, testing, curriculum, grading, student grouping, and almost every other aspect of education. It is not uncommon for the major focus of a district's educational effort to shift suddenly in response to a biannual school board election or the arrival of a new member of the administrative staff. Mandatory testing is now highly consequential for both students and teachers, yet national and state tests lag behind curricular innovation. Hence, research and development aimed at upping the ante for what is taught and learned may not show up on widely accepted measures. Under these circumstances, it is difficult to maintain the sustained focus required to effect educational improvement.

Sometimes logistical and conceptual difficulties become intertwined, for example, the problem of deciding whether the educational program has, in fact, been implemented. All change in schools is uneven, and at any point in time it is far from complete, even if the change has been supported or even mandated by district leadership. Some teachers are early adopters who become essential to the maintenance of the program; others hang on the periphery. Some are enthusiastic about the program but never achieve more than a superficial understanding of it; some resist in active or passive ways. This unevenness of implementation poses problems for the research, especially if the design includes comparison between schools or classrooms that are and are not considered participants. How much and what kinds of participation make a teacher a participant?

In sum, design researchers do not just need to address the conceptual and measurement problems involved in changing and studying the long-term development of learning. In addition, they must cultivate and maintain relationships with the research site, a role that usually includes providing the forms of professional development that support desired forms of teaching and learning. (Professional development that produces generative change in teachers' practice is a difficult and important goal to which an entire base of literature is devoted. See, e.g., Grossman, 1990; Palincsar, Magnussun, Marano, Ford, & Brown, 1998.) Researchers must assist the participating site in managing change, a process that is not always comfortable and that may perturb roles and identities for some individuals. Developing a test bed for extended research is a full-time job in itself. The effort invested in this enterprise means that it is not feasible to step away from site activity to spend a year in uninterrupted analysis of data. One cannot wave goodbye to a school that has come to depend on your support, leaving teachers and students with a promise that you will return when the sabbatical is completed or the book written. Although change may become self-sustaining over time, it is impossible to predict in advance when this will occur, as the organization and constraints of

schooling are powerful forces that operate continually to push teaching and learning back into their more conventional forms. As Spillane (2000) and others have demonstrated, educational reforms usually get assimilated into the patterns of knowledge and practice that preexist in a school, with the result that they are often distorted and rendered sterile.

Challenge 3: Assessing Learning

In these classroom investigations, it is necessary to coordinate fine-grained studies of change in individual students (to identify typical strategies and typical forms of change over time) with coarser-grained measures of achievement in groups of students. The finer-grained studies are required to learn more about the development of scientific thinking that is taken by researchers to be the desirable core of scientific literacy, whether the focus is on change in theories, in students' capabilities to conduct self-regulated investigations, or to engage in modeling practices. As suggested earlier, studies of development that span multiple years pose significant measurement problems because at the outset, little solid evidence exists about how thinking develops when it is systematically supported in an educational context. Therefore, it is unclear when one should look for expected benchmark changes. Coarser-grained studies of student achievement must simultaneously address educators' and parents' concerns about performance on assessments that are consequential with respect to progress, graduation, and college, and at the same time must be sensitive to the goals that are specific to the design.

In our work, we found that developing, revising, and retuning the achievement measures constituted a psychometric project of considerable scope. First, there were no measures of long-term development for the forms of thinking we wished to study (e.g., students' representational competence, spatial visualization, data interpretation, statistical reasoning). Therefore, we developed and/or borrowed items based on our own and others' previous research and initial conjectures about likely forms and rates of student learning. In advance, we were not always able to accurately foretell when it would be reasonable to expect particular benchmark changes. As the educational design unfolded, it was frequently necessary to recalibrate the measures, leading to some undesirable shift from year to year in the data we could collect. Other data collection problems followed from student mobility, the bane of longitudinal designs. Students who studied in collaborating classrooms for 2 contiguous years constituted a reasonably large proportion of our sample, but the proportions of those in project classrooms for 3 years in a row or longer dropped considerably.

There were design issues that followed from the problem of how to identify a fair comparison. The difficulties of accounting for teacher effects and differences in student populations are well established in education research, and these are certainly contributors to the complications of understanding variation in the design, as described earlier. But these difficulties are not just logistical; they are also conceptual. We do not favor control groups that do not control for anything in particular, and moreover, we felt it unlikely that we could persuade teachers in comparison classrooms to spend 3 hours per year testing students on difficult forms of mathematics that they had never studied. Rather than setting up straw-person comparisons of experimental classrooms with those that pursue business as usual, we feel that much more could be learned if the field would pursue a collaborative assessment strategy. Specifically, we hope that in the near future it will be possible to compare the development of student thinking across a few key design studies that vary in interesting ways. The overall strategy would be to develop and use a negotiated common bank of items to assess the learning of students enrolled in different research programs. Because each lead researcher could identify the features theoretically considered central to his or her intervention, the results of such a comparison would be more informative than a typical experimental versus traditional instruction comparison. Presumably, the results would show characteristically different patterns of strengths and weaknesses associated with identifiable instructional approaches. In our opinion, this kind of comparison is a potentially powerful strategy for better understanding the developmental affordances of different designs. We might find, for example, that some approaches produce impressive results in the short term, but others do a far better job over the long haul of producing and sustaining valued outcomes.

Challenge 4: Explaining Contingency

Although design studies offer new opportunities for educational inquiry, they differ from more traditional kinds of study in their purpose, scope, and form of explanation. Like evolutionary biologists and practitioners in some other disciplines (see, e.g., Rudolph & Stewart,

1998), researchers engaged in explaining extended interrelationships between instruction and learning need to account for phenomena that are contingent and historical. Because classroom learning has this character, an important goal for research is to identify and explain the contingencies that the design accounts for—in other words, the patterns of learning and change that, broadly speaking, can reliably be expected to emerge if the design is instantiated. These contingencies need to be teased out from the broad array of features that are not accounted for in the explanatory structure (Lehrer & Schauble, 2001).

One way that researchers address this problem is to generate a set of conjectures that, collectively, take the form of a learning trajectory or pathway. Collectively, these conjectures form a hypothesized sequence or route, one that describes our best-informed guesses of how students typically progress along the path from less expert to more expert forms of thinking. The sequence is conjectural because design studies are typically employed to investigate the teaching and learning of unexplored or underexplored content. For that reason, one cannot be confident that the trajectory will play out as foreseen. Although less detailed and broader in scope, a learning trajectory is a little like an instructional task analysis in cognitive psychology. Its purpose is to guide the overall direction of instruction in domains in which little research currently exists to inform teaching and learning. Accordingly, a learning trajectory embodies one's best bets (informed by research, general knowledge of children's thinking, and reconceptualization of central ideas in the relevant domain) about how development is likely to occur. Of course, as instruction based on a hypothetical learning trajectory is instantiated, the trajectory needs to be revised in real time, in response to what one is learning in the classroom.

Although this brief description captures the general purposes and processes of design studies, there is some danger to taking the analogy too literally. The metaphors of "developmental corridors" and "learning trajectories" do not foreground contingency and variability, which we have argued are very important to understand. What comes to mind when one thinks of "corridor" is an invariant and circumscribed path from a particular beginning place to a known goal. Thinking of development as a path supports the sense of going from somewhere to somewhere else but does not capture the kind of variability in student thinking and performance that often serves as a fundamental mechanism of change. For this reason, it may be more accurate to conceive of development as an ecology that emerges in interactions determined (in part) by the learning opportunities and constraints of tasks, semiotic means (e.g., tools, systems of inscription), recurrent activity structures, and the ways teachers or other members of the community recruit, select, and enhance the contributions of participants (see Lehrer, Strom, & Confrey, 2002).

From this perspective, corridors or trajectories are retrospective accounts of particular realizations of this prospective space of interaction. Designing for education must encourage emergence and variability or else risk pruning the potential for development to sanctioned pathways. Faced with such complexity, educators can choose the path for students and use teaching assistance primarily to minimize straying from the predetermined route. Or instead, one can foster and encourage variability in student thinking and then capitalize on the local opportunities that emerge from it. In that case, the design problem is to craft situations and tasks that are most likely to produce forms of variability that are rich with instructional potential. Of course, one needs an overall vision of where instruction is headed, but that vision can be an elastic one, modifiable at all points by an ongoing assessment of what next move best capitalizes on the contingencies that emerge in the classroom. We argue that this approach is best for capitalizing on students' cognitive resources and performances, but it admittedly makes it more difficult to explain conceptual change. If one reconceives of variability not as error or noise but as grist for development (Siegler, 1996), then documenting and accounting for contingency become an essential part of the research enterprise.

For purposes of tractability, we often ignore these contingencies; indeed, much research is designed so that we can safely do so. But explaining learning entails explaining a phenomenon that is fundamentally historical. Students come to classrooms with learning histories, and moreover, teachers seek to build on those histories. If they succeed, those histories coalesce into enduring propensities and capabilities of the kind that we sometimes call "development." Effective learning does not simply cumulate; instead, later learning transforms what we knew earlier on. Understanding development means understanding those histories, not just their shape, but also their causes. Indeed, the internal psychological characteristics of the learner are important mechanisms, but to understand how scientific thinking and scientific literacy take shape, instruction and other

forms of assistance must also be accounted for. One cannot understand these forms of development without understanding the means by which they are supported. In that sense, an account of development is an account of its history. As Gopnik and Metlzoff (1997) explain:

> Like Darwinian biology, the view presented here suggests that explanations in cognitive science will often be historical and contingent. If we want to say why we have a conceptual structure of a particular kind, we will typically not be able to reduce that structure to some set of first principles. Rather, we will need to trace the historical route that led from our innate theories to the theory we currently hold. On this view, all of cognitive science would be developmental. (p. 218)

Recognizing contingency is an important first step. Developing sound models of history is an enduring challenge.

REFERENCES

American Association for the Advancement of Science (1964). *Science: A process approach.* Annapolis Junction, MD: Author.

Baillargeon, R. (1987). Young infants' reasoning about the physical and spatial characteristics of a hidden object. *Cognitive Development, 2*(4), 179–200.

Baillargeon, R., & Graber, M. (1988). Evidence of location memory in 8-month-old infants in a nonsearch AB task. *Developmental Psychology, 24*(4), 502–511.

Baillargeon, R., & Hanko-Summers, S. (1990). Is the top object adequately supported by the bottom object? Young infants' understanding of support relations. *Cognitive Development, 5*(1), 29–54.

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science.* Madison: University of Wisconsin Press.

Beeth, M. E., & Hewson, P. W. (1999). Learning goals in an exemplary science teacher's practice: Cognitive and social factors in teaching for conceptual change. *Science Education, 83,* 738–760.

Bransford, J. D., Vye, N., Kinzer, C., & Risko, V. (1990). Teaching thinking and content knowledge: Toward an integrated approach. In B. F. Jomes & L. Idol (Eds.), *Symbolizing, communicating, and mathematizing: Perspectives on discourse, tools, and instructional design* (pp. 275–324). Mahwah, NJ: Erlbaum.

Brown, A. L. (1990). Domain-specific principles affect learning and transfer in children. *Cognitive Science, 14*(1), 107–133.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in evaluating complex interventions in classroom settings. *Journal of the Learning Sciences, 2,* 141–178.

Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229–270). Cambridge, MA: MIT Press.

Brown, A. L., & Campione, J. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289–326). Hillsdale, NJ: Erlbaum.

Brown, A. L., & DeLoache, J. (1978). Skills, plans, and self-regulation. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 3–36). Hillsdale, NJ: Erlbaum.

Brown, A. L., & Reeves, R. (1987). Bandwidths of competence: The role of supportive contexts in learning and development. In L. Liben (Ed.), *Development and learning: Conflict or congruence?* (pp. 173–221). Hillsdale, NJ: Erlbaum.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York: Academic Press.

Bullock, M., & Ziegler, A. (1999). Scientific reasoning: Developmental and individual differences. In F.E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich longitudinal study* (pp. 38–54). Cambridge, England: Cambridge University Press.

Carey, S. (1985a). Are children fundamentally different kinds of thinkers than adults? In S. Chipman, J. Segal, & R. Glaser (Eds.), *Thinking and learning skills* (Vol. 2, pp. 485–518). Hillsdale, NJ: Erlbaum.

Carey, S. (1985b). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). An experiment is when you try it and see if it works: A study of grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education, 11,* 514–529.

Carey, S., & Smith, C. (1993). On understanding the nature of scientific knowledge. *Educational Psychologist, 28,* 235–251.

Case, R., & Griffin, S. (1990). Child cognitive development: The role of central conceptual structures in the development of scientific and social thought. In E. A. Hauert (Ed.), *Developmental psychology: Cognitive, perceptuo-motor, and neurological perspectives* (pp. 193–230). Amsterdam, The Netherlands: Elsevier.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120.

Chi, M. T. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. Giere (Ed.), *Cognitive models of science: Minnesota studies in the philosophy of science* (pp. 129–186). Minneapolis: University of Minnesota Press.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in education. *Educational Researcher, 32*(1), 9–13.

DeLoache, J. S. (2004). Becoming symbol-minded. *Trends in Cognitive Sciences, 8*(2), 66–70.

DeLoache, J. S., Pierroutsakos, S. L., & Uttal, D. H. (2003). The origins of pictorial competence. *Current Directions in Psychological Science, 12*(4), 114–118.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*(2/3), 105–225.

diSessa, A. A. (2002). Students' criteria for representational adequacy. In K. Gravemeijer, R. Lehrer, B. van Oers, & L. Verschaffel (Eds.), *Symbolizing, modeling and tool use in mathematics education* (pp. 105–129). Dordrecht, The Netherlands: Kluwer Press.

diSessa, A. A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction, 22,* 293–331.

diSessa, A., Hammer, D., Sherin, B., & Kolpakowski, T. (1991). Inventing graphing: Meta-representational expertise in children. *Journal of Mathematical Behavior, 10,* 117–160.

Dow, P. B. (1991). *Schoolhouse politics: Lessons from the Sputnik era.* Cambridge, MA: Harvard University Press.

Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young people's images of science.* Buckingham, England: Open University Press.

Dunbar, K. (1993). Scientific reasoning strategies for concept discovery in a complex domain. *Cognitive Science, 17*(3), 397–434.

Dunbar, K. (1998). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. F. Davidson (Eds.), *The nature of insight* (pp. 265–395). Cambridge, MA: MIT Press.

Duschl, R. A. (1990). *Restructuring science education: The importance of theories and their developments.* New York: Teachers College Press.

Fay, A., & Klahr, D. (1996). Knowing about guessing and guessing about knowing: Preschoolers' understanding of indeterminacy. *Child Development, 67,* 689–716.

Ford, M. J. (2004). *The game, the pieces, and the players: Coherent dimensions of transfer from alternative instructional portrayals of experimentation.* Manuscript submitted for publication.

Galison, P., & Assmus, A. (1989). Artificial clouds, real particles. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The uses of experiment: Studies on the natural sciences* (pp. 225–274). Cambridge, England: Cambridge University Press.

Gallistel, C. R., Brown, A. L., Carey, S., Gelman, R., & Keil, F. C. (1991). Lessons from animal learning for the study of cognitive development. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind* (pp. 3–36). Hillsdale, NJ: Erlbaum.

Gee, J. P., & Green, J. (1998). Discourse analysis, learning, and social practice: A methodological study. *Review of Research in Education, 23,* 119–169.

Gelman, R., & Baillargeon, R. (1983). A review of some Piagetian concepts. In J. H. Flavell & E. M. Markman (Eds.), *Cognitive development* (pp. 167–230). New York: Wiley.

Gelman, R., & Brenneman, K. (2004). Science learning pathways for young children. *Early Childhood Research Quarterly, 19*(1), 150–158.

Gelman, R., Romo, L., & Francis, W. S. (2002). Notebooks as windows on learning: The case of a science-into-ESL program. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 269–293). Cambridge, England: Cambridge University Press.

Giere, R. N. (1988). *Explaining science: A cognitive approach.* Chicago: University of Chicago Press.

Gooding, D. (1989). "Magnetic curves" and the magnetic field: Experimentation and representation in the history of a theory. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The uses of experiment: Studies on the natural sciences* (pp. 183–223). Cambridge, England: Cambridge University Press.

Gooding, D. (1990). *Experiment and the making of meaning.* Dordrecht, The Netherlands: Kluwer Press.

Goodwin, C. (1994). Professional vision. *American Anthropologist, 96*(3), 606–633.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71*(5), 1205–1222.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5), 620–629.

Greeno, J., & Hall, R. (1997, January). Practicing representation: Learning with and about representational forms. *Phi Delta Kappan,* 361–367.

Grosslight, L., Unger, C., Jay, E., & Smith, C. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching, 28,* 799–822.

Grossman, P. (1990). *The making of a teacher: Teacher knowledge and teacher education.* New York: Teachers College Press.

Grotzer, T. A. (2000, April). *How conceptual leaps in understanding the nature of causality can limit learning: An example from electrical concepts.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Grotzer, T. A. (2003, March). *Transferring structural knowledge about the nature of causality: An empirical test of three levels of transfer.* Paper presented at the annual meeting of the National Association of Research in Science Teaching, Philadelphia, PA.

Hennessey, M. G. (2002). Metacognitive aspects of students' reflective discourse: Implications for intentional conceptual change teaching and learning. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 103–132). Mahwah, NJ: Erlbaum.

Hesse, M. B. (1974). *The structure of scientific inference.* Berkeley: University of California Press.

Hestenes, D. (1992). Modeling games in the Newtonian world. *American Journal of Physics, 60*(8), 732–748.

Hewson, P. W., & Hewson, M. G. (1992). The status of students' conceptions. In R. Duit, F. Goldburg, & H. Niedderer (Eds.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 59–73). Kiel, Germany: University of Kiel, Institute for Science Education.

Hood, B., Carey, S., & Prasada, S. (2000). Predicting the outcomes of physical events: Two-year-olds fail to reveal knowledge of solidity and support. *Child Development, 71*(6), 1540–1554.

Horvath, J., & Lehrer, R. (1998). A model-based perspective on the development of children's understanding of chance and uncertainty. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K to 12* (pp. 121–148). Mahwah, NJ: Erlbaum.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.

Kaiser, D. (2000). Stick-figure realism: Conventions, reification, and the persistence of Feynman diagrams, 1948–1964. *Representations, 70,* 49–86.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science.* Cambridge, MA: MIT Press.

Keil, F. C. (1992). The origins of an autonomous biology. In M. R. Gunnan & M. Maratsos (Eds.), *Minnesota Symposia on Child Psychology: Modularity and constraints on language and cognition* (pp. 103–137). Hillsdale, NJ: Erlbaum.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes.* Cambridge, MA: MIT Press.

Klahr, D., & Dunbar, K. (1988). Dual search space during scientific reasoning. *Cognitive Science, 12,*1–55.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology, 25*(1), 111–146.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction. *Psychological Science, 15*(10), 661–667.

Kline, M. (1980). *Mathematics: The loss of certainty*. Oxford: Oxford University Press.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*(4), 674–689.

Kuhn, D., Amsel, E. D., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. New York: Academic Press.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction, 18*(4), 495–523.

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, 60*(4, Serial No. 245), 1–128.

Kuhn, D., & Phelps, E. (1982). The development of problem solving strategies. In H. Reese (Ed.), *Advances in child development and behavior* (Vol. 17, pp. 1–44). New York: Academic Press.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction, 9*(4), 285–327.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lamon, M., Secules, T., Petrosino, A. J., Hackett, R., Bransford, J. D., & Goldman, S. R. (1996). Schools for Thought: Overview of the project and lessons learned from one of the sites. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 243–288). Mahwah, NJ: Erlbaum.

Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven: Yale University Press.

Latour, B. (1990). Drawing things together. In M. Lynch & S. Woolgar (Eds.), *Representation in scientific practice* (pp. 19–68). Cambridge, MA: MIT Press.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. London: Cambridge University Press.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The construction of scientific facts*. Princeton, NJ: Princeton University Press.

Lee, O., & Fradd, S. H. (1996). Literacy skills in science learning among linguistically diverse students. *Science Education, 80*(6), 651–671.

Lehrer, R., Carpenter, S., Schauble, L., & Putz, A. (2000). Designing classrooms that support inquiry. In J. Minstrell & E. van Zee (Eds.), *Inquiring into inquiry learning and teaching in science* (pp. 80–99). Washington, DC: American Association for the Advancement of Science.

Lehrer, R., & Chazan, D. (Eds.). (1998). *Designing learning environments for developing understanding of geometry and space*. Mahwah, NJ: Erlbaum.

Lehrer, R., Jacobson, C., Kemeny, V., & Strom, D. (1999). Building on children's intuitions to develop mathematical understanding of space. In E. Fennema & T. A. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 63–87). Mahwah, NJ: Erlbaum.

Lehrer, R., Jacobson, C., Thoyre, G., Kemeny, V., Danneker, D., Horvath, J., et al. (1998). Developing understanding of space and geometry in the primary grades. In R. Lehrer & D. Chazan (Eds.), *Designing learning environments for developing understanding of geometry and space* (pp. 169–200). Mahwah, NJ: Erlbaum.

Lehrer, R., & Lesh, R. (2003). Mathematical learning. In W. Reynolds & G. Miller (Eds.), *Handbook of psychology: Vol. 7. Educational psychology* (pp. 357–391). Hoboken, NJ: Wiley.

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction, 14*(1), 69–108.

Lehrer, R., & Schauble, L. (2000a). The development of model-based reasoning. *Journal of Applied Developmental Psychology, 21*(1), 39–48.

Lehrer, R., & Schauble, L. (2000b). Inventing data structures for representational purposes: Elementary grade students' classification models. *Mathematical Thinking and Learning, 2*(1/2), 51–74.

Lehrer, R., & Schauble, L. (2000c). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in instructional psychology: Vol. 5. Educational design and cognitive science* (pp. 101–159). Mahwah, NJ: Erlbaum.

Lehrer, R., & Schauble, L. (2001, April). *Accounting for contingency in design experiments*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Lehrer, R., & Schauble, L. (2002). Symbolic communication in mathematics and science: Co-constituting inscription and thought. In E. D. Amsel & J. Byrnes (Eds.), *The development of symbolic communication* (pp. 167–192). Mahwah, NJ: Erlbaum.

Lehrer, R., & Schauble, L. (2003). Origins and evolution of model-based reasoning in mathematics and science. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: A models and modeling perspective on mathematics problem-solving, learning, and teaching* (pp. 59–70). Mahwah, NJ: Erlbaum.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal, 41*(3), 635–679.

Lehrer, R., & Schauble, L. (2005). Developing modeling and argument in the elementary grades. In T. Romberg & T. P. Carpenter (Eds.), *Understanding mathematics and science matters* (pp. 29–53). Mahwah, NJ: Erlbaum.

Lehrer, R., Schauble, L., Carpenter, S., & Penner, D. E. (2000). The inter-related development of inscriptions and conceptual understanding. In P. Cobb, E. Yackel, & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design* (pp. 325–360). Mahwah, NJ: Erlbaum.

Lehrer, R., Schauble, L., & Petrosino, A. J. (2001). Reconsidering the role of experiment in science education. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 251–278). Mahwah, NJ: Erlbaum.

Lehrer, R., Schauble, L., Strom, D., & Pligge, M. (2001). Similarity of form and substance: Modeling material kind. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 39–74). Mahwah, NJ: Erlbaum.

Lehrer, R., Strom, D., & Confrey, J. (2002). Grounding metaphors and inscriptional resonance: Children's emerging understanding of mathematical similarity. *Cognition and Instruction, 20*, 359–398.

Lemke, J. L. (1990). *Talking science: Language, learning and values*. Norwood, NJ: Ablex.

Leslie, A. M. (1984). Spatiotemporal continuity and the perception of causality in infants. *Perception, 13*, 287–305.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review, 94*(4), 412–426.

Lynch, M. (1990). The externalized retina: Selection and mathematization in the visual documentation of objects in the life sciences. In M. Lynch & S. Woolgar (Eds.), *Representation in scientific practice* (pp. 153–186). Cambridge, MA: MIT Press.

Masnick, A. M., & Klahr, D. (2003). Error matters: An initial exploration of elementary school children's understanding of experimental error. *Journal of Cognition and Development, 4*(1), 67–98.

Metz, K. E. (1995). Re-assessment of developmental assumptions in children's science instruction. *Review of Educational Research, 65*(2), 93–127.

Metz, K. E. (2000). Young children's inquiry in biology: Building the knowledge bases to empower independent inquiry. In J. Minstrell & E. H. van Zee (Eds.), *Inquiring into inquiry learning and teaching in science* (pp. 371–404). Washington, DC: American Association for the Advancement of Science.

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction, 22*(2), 219–290.

Minstrell, J., & van Zee, E. (Eds.). (2000). *Inquiring into inquiry teaching and learning in science*. Washington, DC: American Association for the Advancement of Science.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

Olson, D. R. (1994). *The world on paper: The conceptual and cognitive implications of writing and reading*. New York: Cambridge University Press.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and monitoring activities. *Cognition and Instruction, 1*(2), 117–175.

Palincsar, A. S., & Magnussun, S. (2001). The interplay of first-hand and second-hand investigations to model and support the development of scientific knowledge and reasoning. In D. Klahr & S. Carver (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 151–193). Mahwah, NJ: Erlbaum.

Palincsar, A. S., Magnussun, S. J., Marano, N., Ford, D., & Brown, N. (1998). Designing a community of practice: Principles and practices of the GIsML community. *Teaching and Teacher Education, 14*, 5–19.

Penner, D. E., Giles, N. D., Lehrer, R., & Schauble, L. (1997). Building functional models: Designing an elbow. *Journal of Research in Science Teaching, 34*(2), 125–143.

Penner, D. E., & Klahr, D. (1996). The interaction of domain-specific knowledge and domain-general discovery strategies: A study with sinking objects. *Child Development, 67*, 2709–2727.

Penner, D. E., Lehrer, R., & Schauble, L. (1998). From physical models to biomechanics: A design-based modeling approach. *Journal of the Learning Sciences, 7*(3/4), 429–449.

Penner, E., & Lehrer, R. (2000). The shape of fairness. *Teaching Children Mathematics, 7*(4), 210–214.

Perkins, D. N., & Grotzer, T. A. (2000, April). *Models and moves: Focusing on dimensions of causal complexity to achieve deeper scientific understanding*. Paper presented at the annual conference of the American Educational Research Association, New Orleans, LA.

Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning, 5*(2/3), 131–156.

Piaget, J. (1962). *Play, dreams and imitation in childhood*. New York: Norton.

Pickering, A. (1989). Living in the material world: On realism and experimental practice. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The uses of experiment: Studies on the natural sciences* (pp. 275–297). Cambridge, England: Cambridge University Press.

Pickering, A. (1995). *The mangle of practice: Time, agency, and science*. Chicago: University of Chicago Press.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Towards a theory of conceptual change. *Science Education, 66*(2), 211–227.

Roth, W., & McGinn, M. K. (1998). Inscriptions: Toward a theory of representing as social practice. *Review of Educational Research, 68*(1), 35–59.

Rudolph, J. L., & Stewart, J. H. (1998). Evolution and the nature of science: On the historical discord and its implications for education. *Journal of Research in Science Teaching, 35*(10), 1069–1089.

Samarapungavan, A. (1992). Children's judgments in theory-choice tasks: Scientific rationality in childhood. *Cognition, 45*(1), 1–32.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*(1), 31–57.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*(1), 102–119.

Schwab, J. (1962). The teaching of science as enquiry. In J. Schwab & P. Brandwein (Eds.), *The teaching of science* (pp. 1–103). Cambridge, MA: Harvard University Press.

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air pump*. Princeton, NJ: Princeton University Press.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Research Council.

Shultz, T. R. (1982). Casual reasoning in the social and nonsocial realms. *Canadian Journal of Behavioral Science, 14*, 307–322.

Sibum, H. O. (2004). What kind of science is experimental physics? *Science, 306*(5693), 60–61.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Sloane, F., & Gorard, S. (2003). Exploring modeling aspects of design experiments. *Educational Researcher, 32*(1), 29–31.

Smith, C. L., Maclin, D., Houghton, C., & Hennessey, M. G. (2000). Sixth grade students' epistemologies of science: Experiences on epistemological development. *Cognition and Instruction, 18*(3), 349–422.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62*, 753–766.

Spillane, J. (2000). Cognition and policy implementation: District policymakers and the reform of mathematics education. *Cognition and Instruction, 18*, 141–179.

Strauss, S. (1998). Cognitive development and science education: Toward a middle level model. In W. Damon, I. Sigel, & K. A. Renninger (Eds.), *Handbook of child psychology* (Vol. 4, pp. 357–399). New York: Wiley.

Troseth, G. L. (2003). Getting a clear picture: Young children's understanding of a televised image. *Developmental Science, 6*(3), 247–253.

Troseth, G. L., & DeLoache, J. S. (1998). The medium can obscure the message: Young children's understanding of video. *Child Development, 69*, 950–965.

Troseth, G. L., Pierroutsakos, S. L., & DeLoache, J. S. (2004). From the innocent to the intelligent eye: The early development of pictorial competence. In R. Kail (Ed.), *Advances in child development and behavior* (Vol. 32, pp. 1–35). New York: Academic Press.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development, 51,* 1–10.

Van Valkenburgh, B., Wang, X., & Damuth, J. (2004). Cope's rule, hypercarnivory, and extinction in North American canids. *Science, 306*(5693), 101–104.

Warren, B., & Rosebery, A. S. (1996). This question is just too, too easy! Students' perspectives on accountability in science. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 97–126). Mahwah, NJ: Erlbaum.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*(4), 129–140.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20,* 273–281.

White, B. Y. (1993). Intermediate causal models: A missing link for successful science education. *Cognition and Instruction, 10*(1), 1–100.