Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Developmental Review 27 (2007) 172–223
DEVELOPMENTAL REVIEW
www.elsevier.com/locate/dr

The development of scientific thinking skills in elementary and middle school [☆]

Corinne Zimmerman

Department of Psychology, Illinois State University, Campus Box 4620, Normal, IL 61790, USA

Received 23 January 2006; revised 18 December 2006
Available online 9 March 2007

Abstract

The goal of this article is to provide an integrative review of research that has been conducted on the development of children's scientific reasoning. Broadly defined, scientific thinking includes the skills involved in inquiry, experimentation, evidence evaluation, and inference that are done in the service of *conceptual change* or scientific *understanding*. Therefore, the focus is on the thinking and reasoning skills that support the formation and modification of concepts and theories about the natural and social world. Recent trends include a focus on definitional, methodological and conceptual issues regarding what is normative and authentic in the context of the science lab and the science classroom, an increased focus on metacognitive and metastrategic skills, and explorations of different types of instructional and practice opportunities that are required for the development, consolidation and subsequent transfer of such skills.

Keywords: Scientific thinking; Scientific reasoning; Evidence evaluation; Experimentation; Investigation; Inquiry; Cognitive development

Children's scientific thinking has been of interest to both psychologists and educators. Developmental psychologists have been interested in scientific thinking because it is a fruitful area for studying conceptual formation and change, the development of reasoning and problem solving, and the trajectory of the skills required to coordinate a complex set of cognitive and metacognitive abilities. Educators and educational psychologists have

[☆] I thank Carol Smith, David Klahr, Barbara Koslowski, Deanna Kuhn, Amy Masnick and two anonymous reviewers for helpful comments on previous drafts of this paper.

E-mail address: czimmer@ilstu.edu

0273-2297/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.dr.2006.12.001

shared this interest, but with the additional goal of determining the best methods for improving learning and instruction in science education. Research by developmental and educational researchers, therefore, should and can be mutually informative.

In an earlier review (Zimmerman, 2000), I pointed to the need for an increase in research at the intersection of cognitive development and science education, and that such synergistic research could help children to become better science students and scientifically literate adults. In the intervening years, there is evidence that educators and curriculum designers have been influenced by laboratory research on children's thinking. Concurrently, cognitive and developmental researchers have become aware of the objectives of educators and updated science education standards which recommend a focus on investigation and inquiry at all educational levels (e.g., American Association for the Advancement of Science, 1990, 1993; National Research Council, 1996, 2000) and have used such knowledge in guiding research in both the lab and the classroom. Such a synergistic research strategy is especially important in light of current political and educational climate calling for "scientifically based research" and "evidence-based strategies" to support educational reforms (Klahr & Li, 2005; Li, Klahr, & Siler, 2006).

Scientific thinking is defined as the application of the methods or principles of scientific inquiry to reasoning or problem-solving situations, and involves the skills implicated in generating, testing and revising theories, and in the case of fully developed skills, to reflect on the process of knowledge acquisition and change (Koslowski, 1996; Kuhn & Franklin, 2006; Wilkening & Sodian, 2005). Participants engage in some or all of the components of scientific inquiry, such as designing experiments, evaluating evidence and making inferences in the service of forming and/or revising theories[1] about the phenomenon under investigation.

My primary objective is to summarize research findings on the development of scientific thinking, with a particular focus on studies that target elementary- and middle-school students. To preview, sufficient research has been compiled to corroborate the claim that investigation skills and relevant domain knowledge "bootstrap" one another, such that there is an interdependent relationship that underlies the development of scientific thinking. However, as is the case for intellectual skills in general, the development of the component skills of scientific thinking "cannot be counted on to routinely develop" (Kuhn & Franklin, 2006, p. 974). That is, even though young children demonstrate many of the requisite skills needed to engage in scientific thinking, there are also conditions under which adults do not show full proficiency. Although there is a long developmental trajectory, research has been aimed at identified how these thinking skills can be promoted by determining the types of educational interventions (e.g., amount of structure, amount of support, emphasis on strategic or metastrategic skills) that will contribute most to learning, retention and transfer. Research has identified what children are capable of with minimal support, but is moving in the direction of ascertaining what children are capable of, and

[1] Although there are many definitions of and disagreements about what counts as *theory*, this term will be used in an approach-neutral way to refer to an "empirical claim." This usage is consistent with Kuhn and Pearsall (2000) who outline four possible uses of the term theory or "theoretical claim," which range from least stringent such as *category* and *event claims* (e.g., "this plant died") to most stringent such as *causal* or *explanatory* claims which include an explanation of why the claim is correct (e.g., "this plant died because of inadequate sunlight"). The commonality among theoretical claim types is that "although they differ in complexity, each . . . is potentially falsifiable by empirical evidence" (p. 117).

when, under conditions of practice, instruction and scaffolding. These basic findings will inform the development of educational opportunities that neither underestimate nor overestimate children's abilities to extract meaningful experiences from inquiry-based science classes. The goal of the present review is to reiterate that research by cognitive developmental and educational psychologists can inform efforts to reform science education and, potentially, teacher preparation.

Overview

The current review will focus on the reasoning and thinking skills involved in students' scientific inquiry, such as hypothesis generation, experimental design, evidence evaluation and drawing inferences. Klahr's (2000; 2005a; Klahr & Dunbar, 1988) *Scientific Discovery as Dual Search* (SDDS) model will serve as the general framework for organizing the main empirical findings to be discussed. The SDDS framework captures the complexity and the cyclical nature of the process of scientific discovery (see Klahr, 2000; for a detailed discussion). Thus one can use the top level categories of the model to organize the extensive literature on scientific reasoning by focusing on the three major cognitive components of scientific discovery: Searching for hypotheses, searching for experiments (i.e., data or evidence from experiments or investigations more generally), and evidence evaluation. The studies to be reviewed involve one or more of these three processes.

The review of the literature will include (a) research on experimentation skills; (b) research on evidence evaluation skills; and (c) research that takes an integrated approach. In these integrative investigations, participants actively engage in all aspects of the scientific discovery process so that researchers can track the development of conceptual knowledge and reasoning strategies. Such studies typically include methodologies in which participants engage in either partially guided or self-directed experimentation for either a single session or over the course of several weeks. Studies that focus specifically on conceptual development in various scientific domains (e.g., physics, biology) will not be discussed here.[2]

In the final section of the paper, I will provide a general summary and highlight the consistent findings and limitations of the body of work that addresses the dual purposes of understanding cognitive development and the application of such knowledge to the improvement of formal and informal educational settings.

Research focusing on experimental design skills

Experimentation is an *ill-defined* problem for most children and adults (Schauble & Glaser, 1990). The goal of an experiment is to test a hypothesis against an alternative, whether it is a specific competing hypothesis or the complement of the hypothesis under

[2] Reviews and collections of work on domain-specific concepts can be found in Carey (1985, 2000), Gelman (1996), Gentner and Stevens (1983), Hirschfeld and Gelman (1994), Keil (1989), Pfundt and Duit (1988), Sperber, Premack, and Premack (1995), and Wellman and Gelman (1992, 1998). Additional approaches to the study of scientific reasoning also exist, such as the study of explanation (e.g., Keil, 1989; Rozenblit & Keil, 2002) and students' epistemologies of science (e.g., diSessa, 1993; Smith, Maclin, Houghton, & Hennessey, 2000), but space limitations also preclude a thorough treatment of these topics, despite their obvious importance for a full understanding of scientific thinking.

consideration (Simon, 1989). Experimentation can serve to generate observations (expected as well as unanticipated or "anomalous" data) in order to induce a hypothesis to account for the pattern of data produced (discovery context) or to test the tenability of an existing hypothesis under consideration (confirmation/verification context). The isolation and control of variables and the systematic combination of variables are particular skills that have been investigated. The control of variables is a basic, domain-general strategy[3] that allows valid inferences and is an important strategic acquisition because it constrains the search of possible experiments (Klahr, 2000). In addition to being essential for *investigation*, producing unconfounded experiments yield evidence that is interpretable and therefore facilitates *inferential* skills. Confounded experiments yield indeterminate evidence, thereby making valid inferences and subsequent knowledge gain impossible.

One approach to examining experimentation skills involves minimizing the role of prior knowledge to focus specifically on the strategies that can be used regardless of the content to which they are applied. For example, building on the research tradition of Piaget (e.g., Inhelder & Piaget, 1958), Siegler and Liebert (1975) examined the acquisition of experimental design skills by fifth- and eighth-grade children on a task for which domain-specific knowledge could not be used. The problem involved determining how to make an electric train run by finding a particular configuration of four on/off switches. The train was actually controlled by a secret switch so that the discovery of the correct solution could be postponed until all 16 combinations were generated.

Siegler and Liebert (1975) used two instructional conditions and a control condition. In the *conceptual framework* condition, children were taught about factors, levels, and tree diagrams. In the *conceptual framework plus analogs* condition, children were also given practice and help representing all possible solutions to a problem with a tree diagram. All students were provided with paper and pencil to keep track of their findings. Few students in the control condition (0% of fifth graders and 10% of eighth graders) were successful in producing the complete set of 16 factorial combinations. Students exposed to 20–25 minutes of instruction were more successful in the case of eighth graders (50% produced all combinations) but not fifth graders (0%). In contrast, in the *conceptual framework plus analogs* condition which included 20–25 min of instruction and practice, the majority of fifth graders (70%) and all eighth graders were able to engage in the manipulation and isolation of variables necessary for success on this task.

An equally important finding was that, in addition to instructional condition and age, *record keeping* was a significant mediating factor for success in producing the complete combinatorial solution. The eighth-graders were more aware of their memory limitations, as most kept records (90–100% in the instructional conditions). The fifth-graders were less likely to anticipate the need for records. Those who did rely on memory aids were more likely to produce the complete factorial combination.

An analogous knowledge-lean task is the colorless liquid task originally used by Inhelder and Piaget (1958). Kuhn and Phelps (1982) presented four different flasks of colorless fluid to fourth- and fifth-graders. The researcher demonstrated that by adding several drops of a fifth fluid, one particular combination of fluids changed color. On subsequent weekly sessions, the children's task was to determine which fluid or combinations of fluids

[3] Note that I do not intend the term "domain general" to mean "content free." I use the term to describe a skill or strategy that can be applied to many different content areas. Some research, however, has used the approach of minimizing the role of content knowledge in order to focus on strategy development.

would reproduce the effect. The specific goal was provided to students and domain knowledge of fluids (e.g., color or smell) could not be used to identify likely hypotheses, so success depended on the ability to isolate and control variables to determine which colorless fluid was causally related to the outcome (i.e., producing a cloudy or red mixture).

Neither specific instruction nor feedback was provided—the only feedback students received was the effects of their experiments (i.e., a mixture changing color or not). Although an interviewer asked questions in order to interpret what students were doing, reinforcement was not provided and solutions or strategies were not suggested. Experimentation strategies could be classified as one of three types of genuine (or valid) experimentation (e.g., conducted for the purpose of testing a hypothesis, use of controlled tests) or one of three types of pseudo-experimentation (e.g., uncontrolled, no rationale for the selection of materials). Inferences could also be coded as valid (i.e., based on a controlled comparison) or invalid (e.g., based on intuition, uncontrolled tests, insufficient evidence, etc.).

In an initial study (11 weeks) and a replication (13 weeks), approximately half of the students went on to master the task, and showed consistent use of efficient and valid inference and experimentation strategies.[4] However, an abrupt change from invalid to valid strategies was not common. Rather, there was the gradual attainment of stable valid strategies by some students (typically around weeks 5–7). Students who were ultimately successful showed a relatively frequent use of genuine experimentation strategies (60–100%) prior to stabilization (unsuccessful students used genuine experimentation only 9–45% of the time). Experimentation coded as genuine included the characteristic of "planfulness," meaning it was conducted with a purpose in mind, including the possibility of alternative outcomes (i.e., producing or not producing the effect). Planful experimentation was common among successful students, leading Kuhn and Phelps to speculate that students who eventually discarded invalid strategies attained some level of metastrategic understanding – that is, they began to understand that the strategy worked, but also how and why it works and therefore was the best strategy to apply to the problem.

Tschirgi (1980) looked at how experimental design was related to hypothesis testing in "natural" problem situations. It was hypothesized the value of the outcome might be one factor that determines whether people seek either disconfirming or confirming evidence. Story problems were used in which two or three variables were involved in producing either a good or a bad outcome (e.g., baking a good cake). Adults and children in grades 2, 4, and 6 were asked to determine which levels of a variable to change and which to keep constant to produce a conclusive test of causality. In the cake scenario, for example, there were three variables: type of shortening (butter or margarine), type of sweetener (sugar or honey), and type of flour (white or wholewheat). Participants were told that a story character used margarine, honey, and wholewheat flour and believed that the honey was the responsible for the (good or bad) outcome. They were then asked how the character could prove this given three options: (a) baking another cake using the same sweetener (i.e., honey), but changing the shortening and flour (called the HOTAT strategy, for "Hold One Thing At a Time"); (b) using a different sweetener (i.e., sugar), but the same shortening and

[4] The Kuhn and Phelps (1982) study can be classified as a self-directed experimentation study because it includes a hands-on task and follows participants in a microgenetic context. Although it anticipates Kuhn's later work (e.g., Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Kuhn, Schauble, & Garcia-Mila, 1992) it is discussed here because the task used is relatively knowledge-lean and primary skills investigated include isolating, controlling and combining variables.

flour (called the VOTAT strategy, for “Vary One Thing At a Time” and which is the only strategy that results in an unconfounded experiment[5]); or (c) changing all the ingredients (i.e., butter, sugar, and white flour) (CA or “Change All”). Participants were told to pick the one *best* answer from the three choices provided for eight different problems (four good and four bad outcome).

Tschirgi (1980) found that the value of the outcome influenced the strategy for selecting an experiment to produce evidence. In all age groups, participants looked for confirmatory evidence when there was a “positive” outcome by selecting the HOTAT strategy for manipulating variables (choice *a* above) more frequently than VOTAT or CA. That is, when the outcome was positive, there was a tendency to hold the presumed causal variable constant in order to maintain the good result (consistent with a confounded experiment). In contrast, disconfirmatory evidence was selected when there was a “negative” outcome. The VOTAT strategy (choice *b* above) was chosen more frequently than HOTAT or CA, suggesting that participants were searching for the one variable to change to eliminate the bad result (consistent with the elements of a controlled experiment). The only developmental difference was that the second- and fourth-graders were more likely to select the Change All strategy, but more so for the bad outcomes (likely as a way to eliminate all possible offending variables). Tschirgi suggested that the results support a model of natural inductive logic that develops through everyday problem-solving experience with multivariable situations. That is, individuals base their choice of strategy on empirical foundations (e.g., reproducing positive effects and eliminating negative effects), not logical ones.

Zimmerman and Glaser (2001) investigated whether sixth-grade students were influenced by variations in cover story when designing an experiment about plants. The task followed a curriculum unit involving the design and execution of experiments with plants. Students were provided with a hypothesis to test, but did not conduct the experiment. All students who were asked to design an experiment to test the claim that “tap water is bad for plants” (i.e., a negative outcome) suggested a controlled design (i.e., only one variable was manipulated). The majority of students (79%) suggested the manipulation of the correct independent variable (i.e., water type) to test the claim directly. In contrast, students given the claim that “coffee grounds are good for plants” (i.e., a positive outcome) designed experiments to test the generality of the claim. Rather than testing the veracity of the claim, they designed experiments to determine which types of plants coffee grounds are good for, with only a quarter of students suggesting the correct variable to test (i.e., coffee grounds). Even with classroom experience, variations in the form of the hypothesis (positive/negative) affected students’ selection of a design in an open-ended task. Either cover story could have served as a plausible assessment task at the end of this curriculum unit, but the resulting information about what students learned would be quite different.

In the studies by Tschirgi (1980) and Zimmerman and Glaser (2001), students’ experimentation skills appear to be influenced by situational factors, such as whether the outcome can be interpreted as positive or negative. Under the conditions of a positive plausible outcome, individuals proceed as though they are certain about the causal status of a variable and the task is to demonstrate that the claim holds under a variety of conditions (e.g., to show that honey produces a good cake regardless of flour or shortening type). Normatively, the missing step is the initial confirmation of the claim in a controlled way—

[5] The VOTAT strategy is more recently referred to as the “control of variables” strategy or CVS.

showing that under some constant condition, for example, honey is better than sugar. Variations in mental models of experimentation and/or mental models of causality may underlie these performance variations (e.g., Grotzer, 2003; Kuhn, Black, Keselman, & Kaplan, 2000), and these issues will be addressed more fully in subsequent sections (to preview, the influence of perceived goal on experimentation strategy is a robust finding).

Bullock and Ziegler (1999) also used story problems that involved manipulating a number of variables to determine which were important in producing a particular outcome (e.g., the construction of the best plane or kite). In a longitudinal study to assess the logic of experimentation, children were tested once a year beginning in third grade with data reported through sixth grade. Children were able to produce contrastive tests, but it was not until fifth grade that the production of appropriate controlled tests was evident. By sixth grade, performance was equivalent to a comparison group of adults. In contrast, the ability to *choose* or recognize controlled tests showed a linear developmental trend, including appropriate verbal justification of that choice. That is, even when individuals could not spontaneously produce a controlled test, they were able to recognized one in contexts that involved producing a particular outcome.

Sodian, Zaitchik, and Carey (1991) investigated whether children in the early school years understand the difference between testing a hypothesis and producing an effect. Many of the tasks used previously involved producing an effect, thus it was not possible to compare performance under conditions of being instructed to test a hypothesis versus being instructed to produce an effect. Sodian et al. (1991) presented children in first and second grade with a story situation in which two brothers disagree about the size of a mouse (large versus small) in their home. Children were shown two boxes with different sized openings (or “mouse houses”) that contained food. In the *feed* condition, children were asked to select the house that should be used to make sure the mouse could eat the food, regardless of its size (i.e., to produce an effect/outcome). In the *find out* condition the children were to decide which house should be used to determine the size of the mouse (i.e., to test a hypothesis). If a child can distinguish between testing a hypothesis with an experiment and producing an effect (i.e., feeding the mouse), then different houses should be selected in the *feed* and *find out* conditions.

Over half of the first graders answered the series of questions correctly (with justifications) and 86% of the second graders correctly differentiated between conclusive and inconclusive tests. In a second experiment, story characters were trying to determine whether a pet aardvark had a good or poor sense of smell. Children were not presented with a forced choice between a conclusive and inconclusive test. Even with the more difficult task demands of generating, rather than selecting, a test of the hypothesis, spontaneous solutions were generated by about a quarter of the children in both grades. For example, some children suggested placing food very far away; if the aardvark has a good sense of smell, it will find the food. The results support the idea that children as young as 6 can distinguish between a conclusive and inconclusive experimental test of a simple hypothesis when provided with the two mutually exclusive and exhaustive hypotheses or experiments.

Summary of studies on experimentation skills

Under conditions in which producing an effect is not at issue, even children in the first grade understand what it means to test a hypothesis by conducting an experiment, and furthermore, that children as young as 6 can differentiate between a conclusive and an inconclusive experiment (Sodian et al., 1991). Such abilities are important early precursors. The systematic production of factorial combinations and the isolation and control of variables on multivariable tasks have been shown to emerge under conditions of practice or instruction. Without instruction, few fifth- or eighth-graders could produce the full set of possible combinations required to isolate the effects of any one variable (Siegler & Liebert, 1975). With brief instruction in variables and levels and practice with analogous problems the majority of fifth-graders and all eighth-graders were successful. An awareness of one's memory limitations and the need to keep records appears to emerge between the ages of 10 and 13 and was directly related to successful performance. Over the course of several weeks, fourth- and fifth-graders used a mix of valid and invalid experimentation strategies both within and across sessions (Kuhn & Phelps, 1982). Without any direct instruction but with frequent practice, half of the students were able consistently generate successful solutions and these students were more likely to employ valid experimentation strategies and to understand why such strategies were effective.

When the results of an experiment can be construed as either positive or negative, the experimental strategy employed or selected differed (Tschirgi, 1980; Zimmerman & Glaser 2001). Participants selected valid experimental tests when the hypothesized outcome was negative, but less valid strategies when the hypothesized outcome was positive. This finding suggests that domain knowledge may serve to draw attention to the functional effect of the experimental manipulation, and therefore influence the choice of experimental design. Strategies may reflect pragmatic goals of repeating positive effects and avoiding negative effects (and may perhaps be related to Simon's (1957) concept of *satisficing*). A second explanation may involve students' developing epistemologies and metacognitive understanding of the purposes of experimentation. For example, Carey, Evans, Honda, Jay, and Unger (1989) interviewed seventh graders about their understanding of the nature of science. Based on pre-instruction protocols, most students believed that "a scientist 'tries it to see if it works"' (p. 520) and that the goal of the scientist is, for example, to invent things or to cure disease. At this epistemological level, there is a pragmatic concern for particular valued outcomes. Moreover, students did not differentiate between producing a particular phenomenon and understanding a phenomenon (Carey et al., 1989).

The research described in this section was limited to studies focused on experimentation. The specific set of skills included the isolation and control of variables, producing the full set of factorial combinations in multivariable tasks, selecting an appropriate design or a conclusive test, generating experimental designs or conclusive tests, and record keeping. Although limited with respect to the full range of skills involved in scientific thinking, these studies provide a picture of the development of experimentation skills and the conditions under which more and less sophisticated use emerges. The findings from these studies anticipate those to be reviewed in subsequent sections. In particular, robust findings include inter- and intra-individual variability in strategy usage with the co-existence of more and less efficient strategies, the perceived goal of experimentation influencing strategy selection, and the importance of metacognitive awareness. A current practical and theoretical debate concerns the types of practice opportunities that students require to learn and consolidate such skills and the relative advantages of different forms of instructional intervention for different types of learners.

Research on evidence evaluation skills

The evaluation of evidence as bearing on the tenability of a theory is another important process skill that is necessary for scientific investigation. Kuhn (1989, 2002) has argued that the defining feature of scientific thinking is the set of skills involved in differentiating and coordinating theory and evidence. Most studies of students' ability to coordinate theory and evidence focus on what is best described as *inductive causal inference* (i.e., given a pattern of evidence, what inferences can be drawn?). The coordination of theory and evidence can also be studied with respect to its bearing on *epistemological understanding*. In Kuhn's numerous writings she has discussed theory-evidence coordination in both connotations. The implications of these two different connotations will be discussed in more detail below, as this issue represents one of the fundamental debates among researchers who study evidence evaluation.

In most studies examining the development of evidence evaluation skills, the evidence provided for participants to evaluate typically is in the form of *covariation* evidence. Hume (1988/1758) identified the covariation of perceptually salient events as one potential *cue* that two events are causally related. Even young children have a tendency to use the covariation of events (antecedent and outcome) as an indicator of causality (e.g., Gopnik, Sobel, Schulz, & Glymour, 2001; Inhelder & Piaget, 1958; Kelley, 1973; Schulz & Gopnik, 2004; Shultz, Fisher, Pratt, & Rulf, 1986; Shultz & Mendelson, 1975). Although covariation between events is a necessary but not sufficient cue for inferring a causal relationship, it is one of the bases for making inductive causal inferences.

In a simple covariation matrix, there are four possible combinations of the presence and absence of antecedent (or potential cause) and outcome. The most common type of evidence used in such tasks is data in the form of frequencies of the co-occurrence of events.[6] Participants are provided with data corresponding to the cells of a 2 × 2 contingency table in either tabular or pictorial form. The pattern could represent perfect covariation, partial (or imperfect) covariation, or no correlation between the two events. The task may require participants to evaluate a given hypothesis in light of the evidence (i.e., a deductive step) or to determine which hypothesis the pattern of data support (i.e., an inductive step). In either case, the focus is on the inferences that can be made on the basis of the *pattern of evidence*, while disregarding prior knowledge.

The evaluation of covariation matrices and data tables

Early work on covariation detection was conducted by Shaklee and her colleagues (e.g., Shaklee, Holt, Elek, & Hall, 1988; Shaklee & Mims, 1981; Shaklee & Paszek, 1985). Children (grades 2 through 8) and adults were presented with 2 × 2 covariation matrices. The data in the table represented two events that may or may not be related (e.g., healthy/sick plant and the presence/absence of bug spray). The task was to determine, given the pattern

[6] Some researchers have begun to explore how children and adults evaluate quantitative evidence. A growing area of educational and psychological research that intersects with the scientific thinking literature involves students' understanding of statistics and numerical data. At this time a thorough review of such research outside the scope of this paper. For examples of such work, see Lovett and Shah (Eds.) (in press) for the proceedings of the Carnegie Symposium, "Thinking with Data"; articles such as Petrosino, Lehrer, and Schauble (2003); and the collection of articles in Lajoie (1998) (Ed.).

of evidence, which hypothesis was supported (i.e., if the events are related or not). The most sophisticated strategy that participants used, even as adults, was to compare the *sums-of-diagonals*. The correct *conditional probability* rule was only used by a minority of participants, even at the college level. Adults could readily learn this rule, when instructed how to compare the relevant ratios. Children in grades 4 through 8 could be taught to use the *sums-of-diagonals* rule (Shaklee et al., 1988). Training success was apparent at a one-week delayed post-test. In many respects, the task in this form has more to do with mental arithmetic or naïve data analysis and less with identification of covariation between events (Holland, Holyoak, Nisbett, & Thagard, 1986). Shaklee's work, however, demonstrated that participants' judgments were rule-governed, and that information from all four cells was used but in a less than ideal manner.

Using data tables in which only two conditions were compared, Masnick and Morris (2002) examined how the characteristics of measurement data, such as sample size and variability within the data set (i.e., the magnitude of differences, relative size of data points within a data set, and the presence of outliers) influenced the conclusions drawn by third- and sixth-grade children and adults. Participants were shown pairs of data sets of differing samples sizes and variability characteristics with plausible cover stories (e.g., testing new sports equipment), and asked to indicate what conclusions could be drawn on the basis of the data sets (e.g., which type of golf ball travels farther?), including the reason for that conclusion. At all ages, participants were sensitive to the idea that one can be more confident of a conclusion that is based on a larger sample of observations. When asked to make decisions without the use of statistical tools, even third- and sixth-graders had rudimentary skills in detecting trends, overlapping data points, and the magnitude of differences. By sixth grade, participants had developing ideas about the importance of variability and the presence of outliers for drawing conclusions from data.

Coordinating theory with covariation evidence

Kuhn, Amsel, and O'Loughlin (1988) were responsible for pioneering work on the development of children and adults' evaluation of covariation evidence. Their primary motivation was to examine how participants reconcile prior beliefs about causal variables (but not causal mechanisms) with covariation evidence presented to them. Simple, everyday contexts were used rather than phenomena from specific scientific disciplines. In an initial theory interview, participants' beliefs about the causal status of various variables were ascertained. For example, adults, sixth- and ninth-graders were questioned about their beliefs concerning the types of foods that make a difference in whether a person caught a cold (35 foods in total). Four variables were selected based on the initial interview: two factors that the participant believed make a difference in catching colds (e.g., type of fruit, and type of cereal) and two factors that do not (e.g., type of potato, and type of condiment). This procedure allowed evidence to be manipulated to present evidence that *confirmed* one existing causal theory and one noncausal theory. Likewise, noncovariation evidence was presented that *disconfirmed* one previously-held causal theory and one noncausal theory.

Covariation data were presented sequentially and cumulatively. Participants were asked a series of questions about what the evidence showed for each of four variables. Responses were coded as either *evidence-based* or *theory-based*. Evidence-based responses made reference to the patterns of covariation or instances of data presented (i.e., the findings of the scientists) even if it meant ignoring reasonable background beliefs. For example, if shown a pattern in which type of cake covaried with getting colds, a participant who noted that the sick children ate chocolate cake and the healthy kids ate carrot cake would be coded as having made an evidence-based response. In contrast, theory-based responses made reference to prior beliefs or theories about why the scientists might have found that particular relationship. For example, a response that chocolate cake has "sugar and a lot of bad stuff in it" or that "less sugar means your blood pressure doesn't go up" (Kuhn, 1989, p. 676) would be coded as theory-based.

Through the series of studies, Kuhn et al. found certain patterns of responding. First, the skills involved in differentiating and coordinating theory and evidence, and bracketing prior belief while evaluating evidence, showed a monotonic developmental trend from middle childhood (grades 3 and 6) to adolescence (grade 9) to adulthood. These skills, however, do not develop to an optimum level even among adults. Even adults had a tendency to meld theory and evidence into a single representation of "the way things are." Second, participants had a variety of strategies for keeping theory and evidence in alignment when they were in fact discrepant. One tendency was to ignore, distort, or selectively attend to evidence that was inconsistent with a favored theory or with plausible background information. For example, the protocol from one ninth-grader demonstrated that upon repeated instances of covariation between type of breakfast roll and catching colds, he would not acknowledge this relationship: "They just taste different... the breakfast roll to me don't cause so much colds because they have pretty much the same thing inside [i.e., dough]" (Kuhn et al., p. 73, elaboration added).

A third tendency was to adjust a theory to fit the evidence. This practice is perfectly reasonable or even normative. What was non-normative was that this "strategy" was often outside an individual's conscious awareness. Participants were often unaware that they were modifying their theory. When asked to recall their original beliefs, some participants would report a theory consistent with the evidence that was presented, and not the theory as originally stated. An example of this is one ninth grader who did not believe type of condiment (mustard versus ketchup) was causally related to catching colds. With cumulative covariation evidence, he acknowledged the evidence and elaborated a theory based on the amount of ingredients or vitamins and the temperature of the food the condiment was served with to make sense of the data (Kuhn et al., p. 83). Kuhn argued that this tendency suggests that the individual's theory does not exist as an object of cognition. That is, a theory and the evidence for that theory are undifferentiated—they do not exist as separate cognitive (or metacognitive) entities. If they do not exist as separate entities, it is not possible to flexibly and consciously reflect on the relation of one to the other.

Kuhn's interpretation that children (and some adults) cannot distinguish belief or theory from evidence that confirms (or disconfirms) those beliefs has generated much research in response. Various researchers have questioned the interpretation and conclusions of Kuhn et al. (1988) on both methodological and conceptual grounds. Methodological considerations have focused on issues of task complexity and strength of prior beliefs. For example, using a simpler task and story problems for which children do not hold strong prior beliefs, Sodian et al. (1991) demonstrated that even first- and second-grade children can distinguish between the notions of "hypothesis" and "evidence" by selecting or generating a conclusive test of a simple hypothesis.

Ruffman, Perner, Olson, and Doherty (1993) examined 4- to 7-year-old children's abilities to form hypotheses on the basis of covariation evidence using less complex tasks with fewer factors to consider. When given only one potential cause (type of food) that covaried perfectly with an outcome (tooth loss), children as young as 6 could form the hypothesis that the factor is causally responsible based on perfect or partial covariation evidence. To rule out the possibility that children were simply describing a state of affairs, Ruffman et al. tested if 4- to 7-year-olds understood the predictive properties of a hypothesis and found that by age 7, children understood that a newly formed hypothesis (inferred from evidence) could be used to make predictions. Koerber, Sodian, Thoermer, and Nett (2005) used a variant of the Ruffman et al. task, finding further evidence that 5- and 6-year-olds could form beliefs based on evidence, and understand that story characters could change their beliefs based on covariation evidence. They concluded that performance on a "faked evidence task" showed that children did have a metaconceputal understanding that beliefs are formed based on evidence. Children this young, however, did have difficulty making inferences based on patterns of non-covariation evidence (Koerber et al., 2005).

Amsel and Brock (1996) examined whether children and adults evaluated covariation evidence independently of prior beliefs or not, using a task that was less complex and cognitively demanding. However, participants were selected only if they held strong prior beliefs. That is, participants believed that a relationship exists between the health of plants and the presence/absence of sunshine; and that no relationship exists between health of plants and the presence/absence of a charm (represented as a four-leaf clover). Children in 2nd/3rd grade, 6th/7th grade, college students, and non-college adults evaluated four data sets showing either perfect positive correlation or zero correlation that either confirmed or disconfirmed prior beliefs. Participants were asked whether (sun/no sun) or (charm/no charm) were causally related to plant health and to respond based only on the information given and not what they know about plants.

College adults were most like the "ideal reasoner" (i.e., someone whose causal certainty scores were based solely on the pattern of data, disregarding prior knowledge). Both groups of children made judgments consistent with prior beliefs, even when the evidence did not support those beliefs. For example, when the presence of a charm covaried with plant health, children's mean causal certainty was somewhere between "a little sure" and "pretty sure" that the charm was *not* causal. Likewise, children were "a little sure" that sunlight was causally related to plant health, even when the researcher-provided evidence was disconfirming. There was an age and education trend for the frequency of evidence-based justifications. When presented with evidence that disconfirmed prior beliefs, children from both grade levels tended to make causal judgments consistent with prior beliefs. When confronted with confirming evidence, however, both groups of children and adults made similar evidence-based judgments.

The studies discussed thus far address the issue of the conditions under which children are more or less proficient at coordinating theory and evidence. Such work was motivated by Kuhn's assertion that most children's and some adults' evaluation of evidence is done in a way that suggests they meld the two into one representation. When task demands are simplified such that a hypothesis can be induced from a pattern of evidence, children can detect those patterns and use the resultant hypothesis to make predictions (e.g., Ruffman et al., 1993). When a simple deduction is required (e.g., Sodian et al., 1991), children can differentiate between producing an effect and testing an idea. Other methodological variants, such as tasks complexity, the plausibility of factors, participants' method of responding (e.g., certainty judgments versus forced choice), and data coding (e.g., causal judgments and justifications assessed jointly or separately), can be used to demonstrate differences in children's performance on certain evidence evaluation tasks. These methodological variants have produced interesting findings of children's *performance* under different conditions, but they do not speak to the issue of the epistemological status of theory and evidence. Conceptual issues will be addressed next.

Covariation does not imply causation: conceptual issues

Koslowski (1996) has questioned conclusions about children and adults' ability to coordinate theory and evidence on conceptual grounds. The maxim "correlation does not imply causation" has been part of the required training of students in statistics, philosophy, science and social science (e.g., Stanovich, 1998, chap. 5). However, previous researchers utilized tasks in which correct performance has been operationalized as the identification of causal factors from covariation evidence while simultaneously suppressing prior knowledge—in particular, considerations of plausibility and causal mechanisms. In short, correct performance has been defined by the "bracketing" or disregarding prior knowledge: researcher-supplied evidence was taken to be superior to an individual's prior knowledge or theory in the process of evaluating evidence.

The issue of appealing to one or more *causal mechanisms* when evaluating covariation evidence is one of the conceptual issues raised by Koslowski (1996). One of the main concerns in scientific research is with the discovery of causes (Koslowski & Masnick, 2002). Psychologists who study scientific reasoning have been influenced by the philosophy of science, most notably the empiricist tradition that emphasizes the importance of observable events. In real scientific practice though, scientists are also concerned with *causal mechanism*, or the process by which a cause can bring about an effect. It is through a consideration of causal mechanism that we can determine which correlations between perceptually salient events should be taken seriously and which should be viewed as spurious. For example, it is through the identification of the *Escherichia coli* bacterium that we consider a causal relationship between hamburger consumption and illness or mortality.[7]

In the studies by Kuhn et al. (1988) and others (e.g., Amsel & Brock, 1996), correct performance entailed inferring causation from covariation evidence (or lack of a causal relationship from noncovariation evidence). Evidence-based justifications were considered superior to theory-based justifications. For example, a ninth grader was asked to generate the pattern of evidence that would occur if the color of a tennis ball influences the quality of serve, and placed 8 light-colored tennis balls in the "bad serve" basket and 8 dark-colored balls in the "good serve" basket (Kuhn et al., Study 4). When asked how this pattern of evidence proves that color makes a difference, the response was coded as theory-based: "These [dark in *Good* basket] are more visible in the air. You could see them better." (Kuhn et al., 1988, p. 170). Participants frequently needed to explain why the patterns of evidence were sensible or plausible. Kuhn asked "Why are they unable simply to acknowledge that the evidence shows covariation without needing first to explain why this is the outcome one should expect?" (p. 678). Kuhn argued that such responses indicate that participants

[7] Similarly, it is through the absence of a causal mechanism that we do not consider seriously the classic pedagogical example of a correlation between ice cream consumption and violent crime rate. We also use this pedagogical example to illustrate the importance of considering additional variables that may be responsible for both outcomes (i.e., high temperatures for this example). Koslowski and Masnick (2002) also used this example to illustrate that such a correlation could prompt further investigation if a link between fat consumption and testosterone production were found.

do not recognize theory and evidence as distinct. Koslowski (1996; Koslowski & Okagaki, 1986; Koslowski, Okagaki, Lorenz, & Umbach, 1989), in contrast, would suggest this tendency demonstrates that participants' naïve scientific theories incorporate information about both covariation and causal mechanism. In the case of theories about human or social events, Ahn, Kalish, Medin, and Gelman (1995) also presented evidence demonstrating that college students seek out and prefer information about causal mechanism over covariation when making causal attributions (e.g., determining the causes of an individual's behavior).

Koslowski (1996) presented a series of experiments to demonstrate the interdependence of theory and evidence in legitimate scientific reasoning. In most of these studies, participants (sixth graders, ninth graders, adults) considered information about mechanism when evaluating evidence in relation to a hypothesis about a causal relationship. For example, participants were shown either perfect or partial covariation between a target factor (e.g., a gasoline additive) and effect (e.g., improved gas mileage). Perfect correlation was rated as more likely to indicate causation than partial correlation. Participants were then told that a number of plausible mechanisms had been ruled out (e.g., the additive does not burn more efficiently or more cleanly). When asked to rate again how likely it was that the additive is causally responsible for improved gas mileage, the ratings for both perfect and partial covariation were lower for all age groups.

Koslowski (1996) has argued that another legitimate way that participants use prior knowledge during evidence evaluation tasks is with respect to a consideration for the *plausibility* of the covariation evidence. Ruffman et al. (1993), for example, deliberately chose factors that were all equally plausible. Correct performance in the Kuhn et al. (1988) tasks was defined by considering the researcher-supplied covariation evidence as more important than the implausible hypothesis it was intended to support (e.g., ball color as causally related to the quality of tennis serve). Ruffman et al. argued that revising prior beliefs (e.g., about the causal power of color) is more difficult than forming new theories when prior beliefs do not exist or are not held with conviction. Literature on inductive inference supports this claim (e.g., Holland et al., 1986).

In some situations, scientific progress occurs by taking seemingly implausible correlations seriously (Wolpert, 1993). Similarly, Koslowski argued that if people rely on covariation and mechanism information in an interdependent and judicious manner, then they should pay attention to implausible correlations (i.e., those with no apparent mechanism) when the implausible correlation occurs often. For example, the cause of Kawasaki's syndrome depended upon taking seriously the implausible correlation between the illness and having recently cleaned carpets (Koslowski, 1996). Similarly, Thagard (1998a) describes the case of researchers Warren and Marshall who proposed that peptic ulcers could be caused by a bacterium and their efforts to have their theory accepted by the medical community. The bacterial theory of ulcers was initially rejected as implausible, given the assumption that the stomach is too acidic to allow bacteria to survive.

When Koslowski (1996) presented participants with an implausible covariation (e.g., improved gas mileage and color of car), participants rated the causal status of the implausible cause (color) before and after learning about a possible way that the cause could bring about the effect (improved gas mileage). In this example, participants learned that the color of the car affects the driver's alertness (which affects driving quality, which in turn affects gas mileage). At all ages (sixth- and ninth-graders, adults), participants increase their causal ratings after learning about a possible mediating mechanism. The presence of a possible mechanism in addition to a large number of covariations (4 instances or more) was taken to indicate the possibility of a causal relationship for both plausible and implausible covariations.

In summary, the series of experiments presented by Koslowski (1996) as well as research from the conceptual development (e.g., Brewer & Samarapungavan, 1991; Murphy & Medin, 1985) and causal reasoning literatures (e.g., Cummins, 1995; Schulz & Gopnik, 2004; Shultz et al., 1986; White, 1988) supports the idea that both children and adults hold rich causal theories about "everyday" and scientific phenomena that include information about covariation, theoretically relevant causal mechanisms and possible alternative cause. Plausibility is a general constraint on the generation and modification of theories (Holland et al., 1986). Without such constraints, the countless number of possible correlations in a complex environment would be overwhelming.

Inductive causal inference versus epistemological understanding

Kuhn's assertion that some children and adults meld theory and evidence into one representation of "the way things are" has motivated a lot of empirical research to investigate how individuals coordinate theory and evidence. It is important to return to the two different connotations of theory-evidence coordination outlined at the beginning of this section. Kuhn's claim is not that individuals cannot coordinate theory and evidence (e.g., that one implies the other, or that one is consistent with the other). Rather, the claim is "about epistemological understanding, that is, about the failure to recognize theory and evidence as distinct epistemological categories" (Kuhn & Franklin, 2006, p. 983). That is, it is necessary to differentiate between the level of performing inductive causal inferences and that of an individual's metacognitive awareness that the two categories—theory versus evidence—are different kind of information that have unique properties. Most of the studies described thus far do not use tasks that provide evidence to address the distinction between the performance level and metacognitive understanding.

Even though much of the research on evidence evaluation (i.e., inductive causal inference) has not specifically addressed issues of students' epistemological understanding, it has done much to clarify assumptions about how correct performance on evidence evaluation tasks should be operationally defined—assumptions about performance that reflects a fundamental bias, and performance that reflects a consideration of plausibility, causal mechanism, and alternative causes, but that is still scientifically legitimate. For example, when evaluating evidence, it is considered scientifically legitimate to attend to theoretical considerations *and* patterns of evidence. Based on case studies in the history of science (e.g., Thagard, 1998a, 1998b; Tweney, 2001) there are times when it was important to take seriously information about plausibility and causal mechanism when evaluating evidence that required a major alteration to an existing theory or belief. In other cases, it is imperative that theory be held in abeyance to evaluate a pattern of evidence. Evidence can only be judged as plausible or implausible in relation to current knowledge, theory or belief.

Causal versus scientific reasoning

In a recent line of research, Kuhn and Dean (2004) compared the characteristics (e.g., dominant models, methodology) of evidence evaluation research in the scientific reasoning and causal inference literatures. There clearly is (or should be) some connection between scientific and causal reasoning, but these two bodies of work have developed somewhat independently. Researchers who study causal reasoning have had the goal of identifying the universal inference rules used to make judgments of causality from covariation evidence (e.g., by appealing to a causal mechanism). Most research aimed at identifying such rules (e.g., Cheng, 1997) has been conducted with college student populations. The few developmental studies that have been conducted have been used to conclude that causal inference rules emerge in childhood and remain established well into adulthood, with key developmental differences being adults' superior abilities to differentiate between causes and enabling conditions and to consider a greater amount of information when making judgments (e.g., Harris, German, & Mills, 1996). In contrast, research on scientific thinking has been developmental as a rule rather than as an exception. Rather than identification of universal rules, inter- and intra-individual differences have been explored in tasks that focus on both inductive and deductive inferences and for which determining both causal and non-causal factors is important.

Clearly, both groups of researchers are interested in the cognitive processes underlying causal judgments based on evidence. Kuhn and Dean (2004) summarize the key differences in methodologies used in these two lines of research. In causal inference research, college students complete single session paper-and-pencil tasks with investigator-selected evidence to evaluate and for which judgments take the form of probabilities. In scientific reasoning research, microgenetic studies are conducted with children, adolescents, and/or adults, and a real or virtual causal system is investigated. Judgements take the form of inferences of causality, non-causality or indeterminacy. Using these different methodologies, causal reasoning researchers have proposed universal rules that apply across individuals and contexts, whereas scientific reasoning researchers have proposed a long developmental trajectory of skills that vary as a function of the individual and the context.

To shed light on the conflicting findings and conclusions from two research literatures with such similar objectives, Kuhn and Dean (2004) used an experimental paradigm typical of causal inference studies, but which retains certain features of scientific reasoning tasks. Sixth-graders and adults were asked to evaluate evidence about a multivariable system (i.e., factors that influence the speed of a boat such as shape and depth of water) presented in a paper-and-pencil format, but for which judgements were deterministic (i.e., causal, non-causal, or indeterminate) rather than probabilistic. Variable levels and outcomes were presented pictorially with a sequential and cumulative presentation of investigator-selected evidence with intermittent and final prompts to participants to indicate which features were responsible for the outcome.

Kuhn and Dean (2004) found intra-individual variability in performance and developmental trends. During the course of accumulating evidence, both children and adults changed their minds about the causal status of particular variables. Although adults typically justified inferences of causality based on evidence, children were just as likely to appeal to theory as to evidence, or to a mix of the two. Such developmental trends and variability in performance suggest that the causal theories that individuals hold do not translate into universal inference rules, as suggested by theoretical accounts of causal reasoning (e.g., Cheng, 1997; Lien & Cheng, 2000). Moreover, if such universal rules are central to making inferences, then neither children nor adults would have changed their minds about the causal powers of a variable when contradictory evidence was presented, which was not the case. Kuhn and Dean concluded that a full account of the way in which people draw causal inferences from evidence must include an assortment of strategies and rules that vary in validity and efficiency rather than a stable set of inference rules.

Consistent with the idea that there is variability in how individuals evaluate and react to evidence, Chinn and Brewer (1998) developed a taxonomy of possible reactions to evidence that does not fit with one's current beliefs. Such "anomalous data" is frequently encountered by scientists, and has been used by science educators to promote conceptual change. The idea that anomalous evidence promotes conceptual change (in the scientist or the student) rests on a number of assumptions, including that individuals have beliefs about natural or social phenomena, that they are capable of noticing new evidence as inconsistent with those beliefs and as such calls those beliefs into question, and in some cases, beliefs will be altered in response to the new (anomalous) evidence (Chinn & Brewer, 1998).

Chinn and Brewer proposed eight possible responses to anomalous data. Individuals can (a) ignore the data, (b) reject the data (e.g., because of methodological error, measurement error, or bias); (c) acknowledge uncertainty about the validity of the data; (d) exclude the data as being irrelevant to the current theory; (e) hold the data in abeyance (i.e., withhold a judgment); (f) reinterpret the data as consistent with the initial theory; (g) accept the data and make peripheral change or minor modification to theory; (h) accept the data and change the theory. Examples of all of these responses were found in undergraduates' responses to data that contradicted theories to explain the mass extinction of dinosaurs and theories about whether dinosaurs were warm-blooded or cold-blooded. The development of this taxonomy lends support to the idea that data cannot and is not evaluated without context or in relation to belief and theory.

Evaluating anomalous evidence: instructional interventions and cognitive processes

In a series of studies, Chinn and Malhotra (2002a) examined fourth-, fifth-and sixth-graders' responses to data from experiments to determine if there are particular cognitive processes that interfere with conceptual change in response to evidence that is inconsistent with current belief (rather than apply the Chinn and Brewer taxonomy to children's responses). Experiments from physical science domains were selected in which the outcomes produced either ambiguous or unambiguous data, and for which the findings are considered counterintuitive for most children. For example, most children assume that a heavy object falls faster than a light object. When the two objects are dropped simultaneously, there is some ambiguity because it is difficult to observe both objects. Likewise, the landing position of an object dropped by a moving walker is ambiguous because the event occurs quickly. An example of counterintuitive but unambiguous evidence is the reaction temperature of baking soda added to vinegar. Children believe that either no change in temperature will occur, or that the fizzing causes an increase in temperature. Thermometers unambiguously show a temperature drop of about 4 degrees centigrade.

When examining anomalous evidence, difficulties may occur at one of four cognitive processes: observation, interpretation, generalization or retention (Chinn & Malhotra, 2002a). Prior belief may influence what is "observed," especially in the case of ambiguous data. At the interpretation stage, the resulting conclusion will be based on what was (or was not) observed (e.g., a child may or may not perceive two objects landing simultaneously). At the level of generalization, an individual may accept, for example, that these particular heavy and light objects fell at the same rate, but that it may not hold for other situations or objects. Prior beliefs may re-emerge even when conceptual change occurs, so retention of information could also prevent long-term belief change.

Chinn and Malhotra also investigated instructional interventions to determine if they would affect students' evaluation of anomalous data. In the third study, one group was instructed to *predict* the outcomes of three experiments that produce counterintuitive but unambiguous data (e.g., reaction temperature). A second group answered questions designed to promote unbiased observations and interpretations by *reflecting* on the data. A third group was provided with an *explanation* of what scientists expected to find and why. All students reported their prediction of the outcome, what they observed and their interpretation of the experiment. A generalization and retention test followed 9–10 days later. Fifth- and sixth-graders' performance was superior to fourth-graders. The explanation condition resulted in the best generalization and retention scores relative to the data-reflection and prediction conditions. Chinn and Malhotra suggest that the explanation-based intervention worked by influencing students' initial predictions. This correct prediction then influenced what was observed. A correct observation then led to correct interpretations and generalizations, which resulted in conceptual change that was retained. A similar pattern was found using interventions employing either full or reduced explanations prior to the evaluation of evidence.

The set of four experiments led Chinn and Malhotra (2002a) to conclude that children could change their beliefs based on anomalous or unexpected evidence, but only when they were capable of making the correct observations. Difficulty in making observations was found to be the main cognitive process responsible for impeding conceptual change (i.e., rather than interpretation, generalization or retention). Certain interventions, in particular those involving an explanation of what scientists expected to happen and why, were very effective in mediating conceptual change when encountering counterintuitive evidence. With particular scaffolds, children made observations independent of theory, and changed their beliefs based on observed evidence.

Chinn and Malhotra's (2002a) study is unique in the set of studies reviewed here with respect to the inclusion of instructional interventions, but also by the use of first-hand observations of evidence. Studies of student-initiated experimentation will be described next, but it is an interesting question if there are differences when evaluating evidence that is directly observable compared to second-hand evidence that is common in this type of research (e.g., Koslowski, 1996; Kuhn et al., 1988). Kuhn and Ho (1980) examined children's inferences from data they collected themselves (using the colorless fluids task) or from second-hand data already collected by another child. Children evaluating second-hand data did make progress, but not to the same extent and speed as children who conducted the experiments. Kuhn and Ho suggest that an "anticipatory scheme" that results from designing and generating data may be responsible for the differences in progress. This finding is consistent with the intervention used by Chinn and Malhotra (2002a) in which superior performance resulted from explanation-based instruction (i.e., explanations concerned what to anticipate) that influenced children's predictions, observations, inferences, and generalizations.

How evidence is evaluated: Chinn and Brewer's models-of-data theory

Having established that both children and adults have rich theories and beliefs, and that theory and evidence are used interdependently to make inductive causal inferences in a scientifically legitimate manner, the next issue that needs to be addressed is *how do people evaluate evidence*? Koslowski (1996) stressed the importance of the *interdependence* of theory and evidence, and that skilled individuals consider patterns of evidence in conjunction with information about potential causal mechanisms, alternate causes, and the issue of plausibility. Similarly, Chinn and Brewer (2001) proposed the *models-of-data* theory, in which they suggest individuals evaluate evidence by building a cognitive representation that incorporates both: "theories and data become intertwined in complex ways in models of data so that it is not always possible to say where one begins and the other ends" (p. 331). A research narrative or experiment can be represented as a cognitive model that is schematically similar to a semantic network (Chinn & Malhotra, 2002b). The construction of a cognitive model varies by individual, but integrates elements of the research, such as evidence, procedural details, and the theoretical explanation of the observed findings (which may include unobservable mechanisms such as molecules, electrons, enzymes or intentions and desires). Information and events can be linked by different kinds of connections, including causal, contrastive, analogical and inductive links.

Chinn and Brewer (2001) suggest that the cognitive model is then evaluated by considering the plausibility of these links. In addition to considering the links between, for example, data and theory, the model could also be evaluated by appealing to alternate causal mechanisms or alternate explanations. Essentially, an individual seeks to "undermine one or more of the links in the model" (p. 337). If no reasons to be critical can be identified, the individual may accept the new evidence and/or theoretical interpretation.

Models-of-data theory has some empirical support, based on undergraduates' evaluation of evidence in the form of detailed narratives of scientific research (e.g., evidence for whether dinosaurs were warm- or cold-blooded). The tenability of this theory awaits full empirical support, and it has yet to be tested with younger children, perhaps because Chinn and Brewer consider it to be a theory of how people evaluate data, rather than "evidence" in the more generic sense. The general descriptive account, however, may help interpret individual, developmental and task differences in evidence evaluation, especially with respect to how differences in *prior knowledge* could influence the process. For example, Chinn and Malhotra (2002a) noted that some researchers use task domains in which participants' beliefs are particularly "entrenched" or personally involving. For example, when given evidence that type of condiment or type of breakfast roll covaries with catching colds, it may be difficult forsake one's belief that the cold virus is implicated. Other researchers have used tasks with adolescents or adults in which, for example, religious or social beliefs must be questioned (e.g., Klaczynski, 2000; Klaczynski & Narasimham, 1998; MacCoun, 1998). Thus, the *strength* of prior beliefs, and the *personal relevance* of those beliefs may influence the evaluation of the cognitive model. When there is reason to disbelieve evidence (e.g., because it is inconsistent with prior belief), individuals will search harder for flaws (Kunda, 1990). As such, new evidence may not be compelling enough to find fault with the links in the cognitive model. In contrast, beliefs about simple empirical regularities may not be held with such conviction (e.g., the falling speed of heavy/light objects), making it easier to change a belief in response to evidence. Developmentally, adults are more likely than children to possess relevant knowledge, which provides more ammunition for evaluating the links in the cognitive model.

Summary: the development of evidence evaluation skills

The research described in this section was limited to studies in which there was a particular focus on evidence evaluation. The specific skills include the inductive skills implicated in generating a theory to account for a pattern of evidence, and general inference skills involved in reconciling existing beliefs with new evidence that either confirms or disconfirms those beliefs. Different types of tasks with different cover stories and cognitive demands show some of the ways in which individuals make appropriate or inappropriate connections between theory and evidence at a performance level. Given perfect or partial covariation between one potential cause and one effect, children as young as six could generate the hypothesis that the factor is causally responsible. When individuals hold strong prior beliefs, they respond differentially to evidence that confirms or disconfirms those beliefs. Children had difficulty evaluating evidence that disconfirms a prior belief.

With respect to justifying causal attributions, there is a general developmental trend in the use of evidence-based justifications (Amsel & Brock, 1996; Kuhn et al., 1988; Kuhn & Dean, 2004). As with experimentation skills, inter- and intra-individual variability in the use of strategies was found in the ways children and adults draw causal inferences from evidence (Kuhn & Dean, 2004). Children had some difficulties with first-hand observations (rather than researcher-supplied evidence). When children were capable of making the correct observations (which could be facilitated with instructional interventions), conceptual change was promoted (Chinn & Malhotra, 2002a). An effective way of promoting children's observational abilities was to explain what scientists expected to observe and why. Similarly, a general "anticipatory scheme" may be effective (Kuhn & Ho, 1980) at the observational or encoding stage of evidence evaluation. A robust mechanism found to be responsible for differences in cognitive development in general is *encoding* differences (Siegler & Alibali, 2005).

Research focused on evidence evaluation has done much to clarify how normative behavior should be defined relative to the way in which scientists coordinate theory and evidence (e.g., Koslowski, 1996). The nature and strength of prior knowledge, assessments of plausibility of theory and/or evidence, presence of or ability to generate causal mechanisms, and the number of instances are important factors that influence students' ability to make inductive causal inferences. *Models-of-data* is an additional descriptive account of the evidence evaluation process (Chinn & Brewer, 2001). Individuals are hypothesized to construct a cognitive model that includes information about and links between, for example, evidence, theory, mechanism, methods, and alternate causes. The evaluation process involves appraising the links (e.g., causal, inductive) between information and events in the model.

The coordination of theory and evidence may be thought of as corresponding to *inductive causal inference*, as consistent with the skills studied in much of the research reviewed here. The coordination of theory and evidence may also be thought of as an element of *epistemological understanding*. Recently, Chinn and Malhotra (2002b) outlined the characteristics of authentic science with respect to cognitive processes and epistemological understanding, and placed theory-evidence coordination in the subset of skills involving epistemological understanding, referring to "people's basic beliefs about what knowledge is and when it should be changed" (p. 187). Because theory-evidence coordination, at its core, potentially involves the changing of one's belief system or knowledge, Kuhn has argued that one of the key features that differentiate more and less proficient ability is the metacognitive control over the process.

One does not just change their mind in response to evidence—one understands why one has changed a belief. The mechanism for this developmental shift is an explicit recognition that theory and evidence have unique epistemological statuses.

Chinn and Brewer's (2001) hypothesis that individuals construct a cognitive model in which theory and evidence are "intertwined in complex ways" (p. 331) is reminiscent of Kuhn's interpretation that students seem to merge theory and evidence into one representation of "how things are." For example Kuhn (2002) has argued that the development of proficient scientific thinking involves the process of theory-evidence coordination becoming more *explicit, reflective* and *intentional*. This is where we see a second connotation of theory-evidence coordination as reflecting an individual's epistemological understanding. By invoking a cognitive model that includes both theory and evidence as initially intertwined, it is possible to see that with the development of metacognitive and metastrategic competence, how the epistemological status of evidence and theory will become more evident, and the process of knowledge change in response to evidence becomes increasingly within the student's control. A full account of developmental differences in scientific thinking will need to account for both *cognitive processes* (e.g., inductive inference, causal reasoning) and *epistemological understanding*.

Although only one study in this section explicitly explored the effect of instructional interventions, several instructional implications can be drawn. First, although scientific thinking in general and evidence evaluation in particular are complex cognitive skills, it is important to remember that basic cognitive processes are foundational—such as the encoding of information that will be reasoned about. Teachers who use anomalous evidence in the science classroom as a method to promote conceptual change (e.g., Echevarria, 2003) need to be aware that such information will be effective only if students correctly observe and encode it. Elements of students' prior knowledge (e.g., strength, type) will factor into the evidence evaluation process, and there may be inter and intra-individual differences that are evident as students develop inferential skills. The development of proficient evidence evaluation skills may require the co-development and educational support of epistemological understanding.

In the next set of studies to be reviewed, children are presented with tasks that require numerous cognitive skills and the coordination of inferential and investigative skills. Such studies address the issues of the interdependence of prior knowledge, experimentation strategies for generating and evaluating evidence, and the inference and evaluation skills that result in changes to existing knowledge.

Skip to p. 212

Integrated approaches to scientific reasoning: partially guided and self-directed experimentation

Research focusing on either experimentation or evidence evaluation skills has produced both an interesting set of findings and additional research questions. Understanding the development of scientific thinking would be incomplete without studies in which participants take part in all phases of scientific discovery. Rather than trying to control for prior knowledge by using knowledge-lean tasks or instructing participants to disregard prior knowledge, such research examines the "reciprocal influences of strategy on knowledge and knowledge on strategy" (Schauble, Glaser, Raghavan, & Reiner, 1991, p. 203). The co-development of domain-specific knowledge and domain-general strategies is examined as students engage in first-hand investigations in which they conduct experiments to discover and confirm the causal relations in multivariable systems, with minimal constraints imposed by the researcher.

In describing these studies, I will first provide an overview of common features as well as the types of task variants that have been used. I will then highlight the key findings with respect to developmental differences and address each of the performance indicators that map onto the main cognitive components of the SDDS model (hypothesis search, experimentation, evidence evaluation and knowledge change).[8] Characteristic or individual approaches will then be discussed. The findings of these developmental studies are suggestive of the competencies children have, and also the difficulties that can or should be targeted for scaffolding or instruction. As such, the last section will describe research addressing the effects of different instructional and practice interventions.

General features of integrated approaches

In *self-directed experimentation* (SDE) studies, individuals participate in all phases of the scientific investigation cycle (hypothesis generation and revision, experimentation, evidence evaluation). Participants explore and learn about a multivariable causal system through activities that are self-initiated. *Partially guided experimentation* studies include the features of the SDE approach but for the sake of experimental control, or tractability of data analysis, some guidance may be provided by the experimenter (e.g., which questions to address or hypotheses to test).[9] An experimenter may prompt a participant to, for example, explain a design, make an inference, or justify an inference in order to generate codeable responses.

There are two main types of multivariable systems. In the first type of system, participants are involved in a hands-on manipulation of a physical system, such as the ramps task (e.g., Masnick & Klahr, 2003; Klahr, Triona, & Williams, 2007) or the canal task (e.g., Gleason & Schauble, 2000). Although causal mechanisms typically are unobservable, other cues-to-causation are present such as contiguity in time and space, temporal priority, intended action, and generative transmission (e.g., Corrigan & Denton, 1996; Shultz et al., 1986; Sophian & Huber, 1984; White, 1988). The second type of system is a computer simulation. A variety of virtual environments have been created, in domains such as electric circuits (Schauble et al., 1992), genetics (Echevarria, 2003), earthquake or flooding risk (e.g., Keselman, 2003) as well as social science problems such as factors that affect TV enjoyment (e.g., Kuhn et al., 1995) or CD catalog sales (e.g., Dean & Kuhn, 2007; Kuhn, 2005b). For any given system, some variables are consistent with participants' prior beliefs and some are inconsistent. The starting point is the participant's own theory, so the course of theory revision can be tracked as participants evaluate self-generated experimental evidence that either confirms or disconfirms these prior beliefs.

[8] Many self-directed experimentation studies (some with microgenetic designs) have been conducted only with adult participants (e.g., Azmitia & Crowley, 2001; Dunbar, 1993; Okada & Simon, 1997; Schauble, Glaser, Duschl, & Schulze, 1995; Schauble et al., 1991; Schauble, Glaser, Raghavan, & Reiner, 1992; Swaak & de Jong, 2001) and have involved somewhat more sophisticated science domains (e.g., the mechanisms of gene reproduction, electricity) but these will not be reviewed here (but see Zimmerman, 2000).

[9] An analogy may be made between an unstructured interview and a semi-structured interview. For example, in Gleason and Schauble (2000), the researcher did not intervene, except if there were questions about the experimental apparatus. Tytler and Peterson (2004) allowed fairly free-form exploration of different science tasks. In the Masnick and Klahr (2003) study, in contrast, children were guided through some of the experimental trials to ensure consistency across participants.

Given the cyclical nature of the discovery process, analyzing the performance of participants as they explore a causal system results in a wealth of data (see Table 1). With respect to *hypothesis search*, participants' initial theories may be assessed, and if and when any changes occur during the course of the investigation (e.g., Schauble, 1996). Elements of initial and changing beliefs may also be noted (e.g., plausibility of hypotheses, mention of causal mechanisms). A related measure involves the assessment of comprehension or *knowledge change*. For example, seventh-graders knowledge of genetics was measured before and after three weeks of experimentation with genetics simulation software (Echevarria, 2003). Another measure of knowledge acquisition is the successful discovery of the causal/non-causal status of all variables in the multivariable system (e.g., Penner & Klahr, 1996a; Reid, Zhang, & Chen, 2003).

With respect to *conducting experiments*, there are a number of ways to code participants' strategies. The variables and levels that are selected (and when) indicate whether an individual is devoting more or less time to particular variables. A design can be coded as either controlled (CVS/VOTAT) or confounded (HOTAT/Change All). The size of the experiment space can be calculated for multivariable systems (e.g., the "canal task" with varying boat characteristics and canal depths can produce 24 unique combinations of variables) and so the percentage of experiment-space searched (including repetitions) can be measured. The use of data management (recording, consulting) is frequently noted (e.g., the percentage of experiments and outcomes recorded). Other measures include intended plans as well as predictions.

When *evaluating evidence*, the number and type of inferences are recorded. Inferences are coded as being causal (or "inclusion" inferences), non-causal, or indeterminate. Inferences can be coded as being either valid (i.e., based on sufficient evidence and a controlled design) or invalid, and justifications for inferences can be coded as being either evidence-based or theory-based. The number of valid inferences has become a common performance indicator (e.g., Gleason & Schauble, 2000; Keselman, 2003; Kuhn et al., 2000; Kuhn & Pearsall, 1998) because such inferences involve (a) the design of an unconfounded experiment, (b) the correct interpretation of the evidence, and (c) a conclusion that is correctly justified.

Table 1
Common measures used in self-directed and partially guided experimentation studies

Hypotheses space search	Experiment space search	Evidence evaluation	Other
Assessment of initial beliefs	Selection of variable and levels	Inferences/conclusions	Predictions
Change/no change in belief	Percent of E-space searched	Causal	Intentions/plans
Type of hypotheses selected (e.g., plausible, causal, single/multiple)	Strategies	Non-causal	Record keeping/record consulting
	CVS (VOTAT)	Indeterminate	Successful knowledge acquisition
	HOTAT	False inclusion	Status of variables
	Change all	Justifications	Conceptual understanding
		Theory-based	Transfer
		Evidence-based	Retention

Note: Task variants include (a) use of prompts (partially guided) versus minimal intervention (self-directed); (b) individual versus collaborative exploration, (c) science domain; (d) time on task; (d) real or virtual environments; (f) categorical versus continuous/quantitative outcomes; (g) task complexity (e.g., number of variables and levels); and (h) type of instructional intervention or practice.

An additional task variant is the length of time with the task. Many SDE studies use a microgentic method (e.g., Siegler & Crowley, 1991), which involves repeated exposure to the problem-solving environment, often over the course of weeks. In other cases, there is a single experimental session. Some studies include a delayed post-test to assess retention on the same task or transfer to an isomorphic or related task.

Developmental differences

In this section I will describe findings that characterize children's performance, and where possible, make comparisons among groups of children or between children and adults. Findings that relate to changes that occur within or across sessions will also be noted.

Searching hypothesis space: prior knowledge and the selection of hypotheses

When asked to engage in a scientific discovery task, both knowledge and problem-solving strategies are important. Individuals come to the task with existing conceptual knowledge of the task domain, or hypotheses are developed about how a system operates and knowledge changes during the course of investigation.

In multivariable systems, participants form hypotheses about the role of several variables on the outcome measure. Children often *proposed different hypotheses* than adults (Dunbar & Klahr, 1989) and younger children (age 10) often conduct experiments *without explicit hypotheses*, unlike 12–14 year olds (Penner & Klahr, 1996b). Success in SDE tasks is associated with a search for hypotheses to guide experimentation (Schauble & Glaser, 1990). Children tended to *focus on plausible hypotheses* and often go "stuck" focusing on a *single hypothesis* (e.g., Klahr, Fay, & Dunbar, 1993). Adults were more likely to consider multiple hypotheses (e.g., Dunbar & Klahr, 1989; Klahr et al., 1993). For both children and adults, the ability to consider many alternative hypotheses was a factor contributing to success.

Participants come to such tasks with prior beliefs (or developed them on the spot), and such beliefs influence the *choice* of which hypotheses to test, including which hypotheses were tested *first*, *repeatedly*, or received the most *time and attention* (e.g., Echevarria, 2003; Klahr et al., 1993; Penner & Klahr, 1996b; Schauble, 1990; Schauble, 1996; Zimmerman, Raghavan, & Sartoris, 2003). In particular, children and adults are more likely to begin the discovery process by attending to variables *believed to be causal* (e.g., Kanari & Millar, 2004; Klahr et al., 2007; Schauble, 1990, 1996) but over the course of experimentation, especially in microgenetic contexts, children start to consider hypotheses and make inferences about variables believed to be *non-causal* (e.g., Kuhn et al., 1995, 1992; Schauble, 1990, 1996). Standard, or expected hypotheses were proposed more frequently than hypotheses that predicted anomalous or *unexpected* results (Echevarria, 2003). Children's "favored" theories sometimes resulted in the selection of invalid experimentation and evidence evaluation heuristics (e.g., Dunbar & Klahr, 1989; Schauble, 1990). The choice of hypotheses to test as the session(s) progresses is a function of type of experiments conducted and the types of inferences generated (such choices will be addressed in subsequent sections).

Bridging the search for hypotheses and experiments: plausibility and predictions

As discussed in the evidence evaluation section, *plausibility* is a general constraint with respect to belief formation and revision (Holland et al., 1986) and has been identified as a domain-general heuristic (Klahr et al., 1993). That is, individuals may (or should) use the plausibility of a hypothesis to guide the choice of which experiments to pursue. Klahr et al.

provided third- and sixth-grade children and adults with hypotheses to test that were incorrect, but either plausible or implausible. For plausible hypotheses, children and adults tended to go about *demonstrating the correctness* of the hypothesis rather than setting up experiments to decide between rival hypotheses. When provided with implausible hypotheses to test, adults and some sixth-graders proposed a plausible *rival hypothesis*, and set up an experiment that would discriminate between the two. Third graders tended to propose a plausible hypothesis, but then ignore or forget the initial implausible hypothesis, getting sidetracked in an attempt to demonstrate that the plausible hypothesis was correct. One could argue that any hypothesis that is inconsistent with a prior belief could be considered "implausible." Therefore, because both adults and children tend to begin exploration of a causal system by focusing on variables consistent with prior beliefs, these are the variables that are considered to be plausibly related to the outcome (e.g., Kanari & Millar, 2004; Penner & Klahr, 1996a; Schauble, 1996).

The generation of predictions is a skill that overlaps the search for hypotheses and the design of experiments. Students' predicted outcomes could influence the choice of hypothesis to test, and the resultant selection of an experimental design. Once an experiment is set up, many researchers prompt individuals to express what they expect to happen, or the spontaneous utterance of predictive statement may be noted. Making predictions has been used to assess how well individuals understood a causal system, either immediately or after a delay (e.g., Kuhn et al., 2000; Kuhn & Dean, 2005; Reid et al., 2003; Zimmerman et al., 2003). Predictions have also been used as an assessment of how different types of errors (e.g., measurement, execution) are believed to influence the experimental outcome (Masnick & Klahr, 2003). Research by McNay and Melville (1993) showed that children in grades 1–6 are both aware of what predicting means, and are able to generate predictions for a number of science domains. Children are less likely than adults to generate predictions for experiments (e.g., Kuhn et al., 1995). As with hypotheses, students are more likely to make predictions about causal or covariation relations than they are about noncovariation outcomes (e.g., Kanari & Millar, 2004).

Searching experiment space: strategies for generating evidence

As discussed previously, there are a number of strategies for manipulating and isolating variables. Of these, the only one that results in an unconfounded design and is considered valid is the control of variables strategy (CVS; Chen & Klahr, 1999; also known as "vary one thing at a time" [VOTAT]; Tschirgi, 1980). The other strategies (hold one thing at a time [HOTAT] and Change All) are considered invalid strategies, as they produce confounded comparison resulting in ambiguous findings that cannot be unequivocally interpreted.[10] Only inferences of indeterminacy follow evidence generated by invalid strategies. Experimentation can be conducted for two purposes—to test a hypothesis (deductive step) or to generate a pattern of findings to generate a hypothesis (inductive step).

Of the general heuristics identified by Klahr et al. (1993), two focused on experimentation strategies: design experiments that produce informative and interpretable results, and

[10] For example, the HOTAT strategy is usually described as "inappropriate" and "invalid" but in some contexts, this strategy may be legitimate. For example, in real-world contexts, scientists and engineers cannot make changes one at a time because of time and cost considerations. Therefore, for theoretical reasons, only a few variables are held constant (Klahr, personal communication). In the tasks described here, HOTAT is interpreted as invalid because there are typically a countable number of variables to consider, each with only two or three levels.

attend to one feature at a time. Adults were more likely than third- and sixth-grade children to restrict the search of possible experiments to those that were informative (Klahr et al., 1993). Similarly, Schauble (1996) found that in an initial task domain, both children and adults started out by covering about 60% of the experiment space. When they began experimentation of a second task domain, only adults' search of experiment space increased (to almost 80%). Over six weeks, children and adults conducted approximately the same number of experiments. Therefore, children were more likely to conduct unintended duplicate or triplicate experiments, making their experimentation efforts less informative relative to the adults. Working alone, children explore less of the possible problem space, however, when children and parents worked collaboratively, they explored 75% of the possible experiment space (Gleason & Schauble, 2000). Children were more likely to devote multiple experimental trials to variables that were already well understood, whereas adults would move on to exploring variables they did not understand as well (Klahr et al., 1993; Schauble, 1996). This approach to experimentation, in addition to being less informative, illustrates the idea that children may view experimentation as a way of demonstrating the correctness of their current beliefs (Klahr et al., 1993).

With respect to the heuristic of attending to one feature at a time, children are less likely to use the control-of-variables (CVS) strategy than adults. For example, Schauble (1996) found that across two task domains, children used controlled comparisons about a third of the time. In contrast, adults improved from 50% CVS usage on the first task to 63% on the second task. Children usually begin by designing confounded experiments (often as a means to produce a desired outcome), but with repeated practice in microgenetic contexts, they began to use the CVS strategy (e.g., Kuhn et al., 1995, 1992; Schauble, 1990). However, both children and adults display intra-individual variability in strategy usage. That is, multiple strategy usage is not unique to childhood or periods of developmental transition (Kuhn et al., 1995). A robust finding in microgenetic studies is the coexistence of valid and invalid strategies (e.g., Garcia-Mila & Andersen, 2007; Gleason & Schauble, 2000; Kuhn et al., 1992; Schauble, 1990; Siegler & Crowley, 1991; Siegler & Shipley, 1995). Developmental transitions do not occur suddenly. Participants do not progress from an inefficient or invalid strategy to a valid strategy without ever returning to the former. An individual may begin with invalid strategies, but even when the usefulness of the CVS is discovered it is not immediately used exclusively. It is slowly incorporated into an individual's set of strategies, as participants become dissatisfied with the invalid strategies that produce ambiguous evidence. Experimentation and inference strategies often co-develop in microgenetic contexts, and because valid inferences require controlled designs, additional relevant findings will be discussed below.

Data management: recording designs and outcomes

In many SDE studies, participants are provided with some type of external memory system, such as a data notebook or record cards to keep track of plans and results, or access to computer files of previous trials. Tweney, Doherty, and Mynatt (1981) originally noted that many tasks used to study scientific thinking were somewhat artificial because real investigations involve *aided* cognition. Such memory aids ensure a level of authenticity and that the task remains centered on reasoning and problem solving and not memory.

Previous studies demonstrate that children are not often aware of their memory limitations (e.g., Siegler & Liebert, 1975). Recent studies corroborate the importance of an awareness of one's own memory limitations while engaged in scientific inquiry tasks, regardless of age. Carey et al. (1989) reported that prior to instruction, seventh graders did not spontaneously keep records when trying to determine which substance was responsible for producing a bubbling reaction in a mixture of yeast, flour, sugar, salt and warm water. Dunbar and Klahr (1989) also noted that children (grades 3–6) were unlikely to check if a current hypothesis was or was not consistent with previous experimental results. In a study by Trafton and Trickett (2001), undergraduates solving scientific reasoning problems in a computer environment were more likely to achieve correct performance when using the notebook function (78%) than were nonusers (49%), showing this issue is not unique to childhood.

Garcia-Mila and Andersen (2007) examined fourth graders' and adults' use of notetaking during a 10-week investigation of a number of multivariable systems. Notetaking was not required, so the focus was on participants' *spontaneous use* of notebooks provided. All but one of the adults took notes, whereas only half of the children took notes. On average adults made three times more notebook entries than children did. Adults' notetaking remained stable across ten weeks, but children's frequency of use decreased over time, dropping to about half of their initial usage. The researchers suggest that children may not have been aware of the utility of notetaking during investigations, or they may have underestimated the task demands (i.e., there were 48 possible combinations of variables). Children rarely reviewed their notes, which typically consisted of conclusions, but not the variables used or the outcomes of the experiments (i.e., the evidence for the conclusion was not recorded).

Gleason and Schauble (2000) found that in parent-child dyads, it was the parent who was responsible for both recording and consulting data while engaged in collaborative inquiry. Children may differentially record the results of experiments, depending on familiarity or strength of prior theories. For example, 10- to 14-year-olds recorded more data points when experimenting with factors affecting the force produced by the weight and surface area of boxes than when they were experimenting with pendulums (Kanari & Millar, 2004). Overall, it is a fairly robust finding that children are less likely than adults to record experimental designs and outcomes, or to review notes they do keep, despite task demands that clearly necessitate a reliance on external memory aids.

Given the increasing attention to the importance of metacognition for proficient performance on such tasks (e.g., Kuhn & Pearsall, 1998, 2000), it is important to determine at what point children and early adolescents recognize their own memory limitations as they navigate through complex tasks. Metamemory develops between the ages of 5 and 10, but with development continuing through adolescence (Siegler & Alibali, 2005) and so there may not be a particular age or grade level that memory and metamemory limitations are no longer a consideration. As such, metamemory may represent an important moderating variable in understanding the development of scientific thinking (Kuhn, 2001). If the findings of laboratory studies are to be informative to educators, children's metacognitive and metastrategic limitations must be recognized as inquiry tasks become incorporated into science curricula (e.g., Kolodner et al., 2003; White & Frederiksen, 1998). Record keeping is an important component of scientific investigation because consulting *cumulative* records is often a necessary part of the evidence evaluation phase. Children and early adolescents may require prompts and scaffolds to remind them of the importance of record keeping for scientific discovery.

Evaluating evidence: interpretation and inference

Inferences made based on self-generated experimental evidence are typically classified as causal (or inclusion), non-causal (or exclusion), indeterminate, or false inclusion. The first three types can be further classified as valid (i.e., supported by evidence, or in the case of inferences of indeterminacy, correctly "supported by" evidence that is ambiguous or results from a confounded experiment) or invalid. False inclusion, by definition, is an invalid inference but is of interest because in SDE contexts, both children and adults often incorrectly (and based on prior beliefs) infer that a variable is causal, when in reality it is not. Valid inferences are defined as inferences of inclusion (i.e., causal) or exclusion (i.e., not causal) that are based on controlled experiments that include observations for both levels of the target variable. Even after discovering how to make valid inferences, participants often do not give up less-advanced strategies such as making inferences that are (a) consistent with prior beliefs, (b) based on a single instance of covariation (or noncovariation), or (c) based on one level of the causal factor and one level of the outcome factor (e.g., Klahr et al., 1993; Kuhn et al., 1995, 1992; Schauble, 1990, 1996).

Children tend to focus on *causal inferences* during their initial explorations of a causal system. Schauble (1990) found that fifth- and sixth-graders began by producing confounded experiments and to rely on prior knowledge or expectations, and therefore were more likely to make incorrect causal inferences (i.e., false inclusions) during early efforts to discover the causal structure of a computerized microworld. In direct comparison, adults and children both focused on making causal inferences (about 75% of inferences), but adults made more valid inferences because they used a valid experimentation strategy. Children's inferences improved over the course of six sessions, starting at 25% but improving to almost 60% valid inferences (Schauble, 1996).

Adults were more likely to make exclusion inferences and inferences of indeterminacy than children (80% and 30%, respectively) (Schauble, 1996). Kanari and Millar (2004) reported that 10- to 14-year-olds struggled with exclusion inferences. Students explored the factors that influence the period of a pendulum or the force needed to pull a box along a level surface during one session of self-directed experimentation. Only half of the students were able draw correct conclusions about factors that do not covary with outcome, and in these cases, students were more likely to either selectively record data, selectively attend to data, distort or "reinterpret" the data, or state that non-covariation experimental trials were "inconclusive." Such tendencies are reminiscent of Kuhn et al.'s (1988) finding that some individuals selectively attended to or distorted researcher-selected data in order to preserve a prior theory or belief. Three of 14 students distorted or "reinterpreted" self-generated evidence to determine which factors influenced the tilt of a balance-of-forces apparatus (Zimmerman et al., 2003). Most students held prior beliefs the vertical height of a weight should make a difference (see also Aoki, 1991), but some were unable to reconcile this expectation with the data they collected during one session with the apparatus. The remaining students were able to reconcile the discrepancy between expectation and evidence by updating their understanding of the balance system and concluding that vertical distance was non-causal.

Kanari and Millar suggested that non-causal or exclusion inferences may be difficult for students because in the science classroom, it is typical to focus on variables that "make a difference" and therefore students struggle when testing variables that do not covary with the outcome (e.g., the weight of a pendulum does not affect the time of swing or the vertical height of a weight does not affect balance). In addition to extra exposure in the science classroom, Schauble's (1996) finding that three-quarters of inferences were causal means that both children and adults got much more practice and experience with inclusion inferences relative to exclusion inferences. Furthermore, it has been suggested that valid exclusion and indeterminacy inferences are conceptually more complex, because they require one to consider a pattern of evidence produced from several experimental trials (Kuhn et al., 1995; Schauble, 1996), which may require one to review cumulative records of previous outcomes. As has been shown previously, children do not often have the metamory skills to either record information, record sufficient information, or consult such information when it has been recorded.

After several weeks with a task in microgenetic studies, however, fifth- and sixth-grade children will start making more exclusion inferences (that factors are not causal) and indeterminacy inferences (i.e., that one cannot make a conclusive judgment about a confounded comparison) and not focus solely on causal inferences (e.g., Keselman, 2003; Schauble, 1996). They also begin to distinguish between an informative and an uninformative experiment by attending to or controlling other factors, which leads to an improved ability to make valid inferences. Through repeated exposure, invalid inferences, such as false inclusions, drop in frequency. The tendency to begin to make inferences of indeterminacy indicates that students may be developing an awareness of the adequacy or inadequacy of their experimentation strategies for generating sufficient and interpretable evidence.

Children and adults also differ in generating *sufficient evidence* to support inferences. In contexts where it is possible, children often terminate their search early, believing that they have determined a solution to the problem (e.g., Dunbar & Klahr, 1989). In microgenetic contexts where children must continue their investigation, this is less likely because of the task requirements. Children are also more likely to refer to evidence that was salient, or most recently generated. Whereas children would jump to a conclusion after a single experiment, adults typically needed to see the results of several experiments (e.g., Gleason & Schauble, 2000).

Kanari and Millar (2004) have suggested that evidence evaluation studies (e.g., Koslowski, 1996; Kuhn et al., 1988) were actually assessing “logical reasoning” rather than scientific reasoning because actual data were not presented. Rather, such studies present only “findings” or “results” for participants to evaluate, whereas real science involves reasoning from *data*. In developmental studies, children typically evaluate *categorical* evidence that is either self-generated or researcher-selected. That is, the outcome measures may be presented as simple differences (e.g., car A is faster than car B) or lack of difference (e.g., object A had the same sinking time as object B). A few studies have used quantitative measures rather than categorical outcomes. For example, children and adults generated quantitative data when exploring tasks involving hydodynamics and hydrostatics (Schauble, 1996). When repeating experimental trials, variation in resulting data occurred. Some children were confused by the different outcome on a duplicate trial. As children were more likely to conduct duplicate experiments, they were therefore faced with deciding which differences were “real” and which differences represented data variability. Prior expectation was often used to interpret whether numerical differences indicated that a variable had an effect (or not). That is, when in doubt, differences were interpreted as consistent with an effect if that effect was expected, but interpreted as measurement error if an effect was not expected. Therefore, the interpretation of evidence in the form of variable data was often done in such a way as to maintain consistency of belief.

Kanari and Millar (2004) reported that children were more likely to repeat measurements when exploring non-causal variables. As discussed previously, there were general difficulties with variables that “did not make a difference” and such measurement variability served to compound the difficulty of reconciling prior belief with the variable data that were generated. Such data were found to be puzzling to the 10- to 14-year-olds, also contributing to their tendency to distort or reinterpret the data. Based on interview comments, Kanari and Millar concluded that only a minority of students had any awareness of the idea of measurement error. Given the absence of statistical analysis, differences in measurements were thus interpreted by some students as indicating an effect consistent with expectation rather than as error variability.

Error is a part of all empirical investigations, from simple experiments conducted by science students to research conducted by scientists. Masnick and Klahr (2003) identified five stages during experimentation when errors could occur: (a) during the design phase, in which one selects variables to test and control, (b) during the set up of any physical apparatus or measurement device, (c) during the execution of the experiment, (d) during the measurement stage, or (e) during the analysis of the data. Each of these stages can be associated with some subset of four different types of error: design, measurement, execution, or interpretation error.

Masnick and Klahr (2003) examined young children's understanding of experimental error. During one phase of experimentation with features of ramps, students were asked to record the times that it took for balls to roll down two different ramps that varied on only one dimension. Unbeknownst to the children, the experimenter provided one data point that could be considered a noticeable "outlier." Second- and fourth-graders differed in the number and type of reasons they gave for the findings. Children in both grades were likely to mention execution errors, but fourth-graders were more sensitive to the idea of measurement error in the experimental context. An important component of scientific thinking involves an understanding of causes that produce systematic differences in patterns of data/evidence, and the "noise" and variation that is expected when making repeated measurements. Reasoning at the intersection of science and statistics is an important issue that has begun to be explored (see footnote 6). In order to evaluate a pattern of data and make a judgment that it is (a) random error, (b) an unexpected or surprising finding, or (c) a true difference requires one to draw on a knowledge base of concepts about the domain and about strategies for generating and evaluating evidence (Masnick & Klahr, 2003). Therefore, the full cycle of scientific investigation includes evaluating evidence in the light of current knowledge, which requires a coordination of existing knowledge with newly generated evidence that bears on the correctness of one's knowledge or expectations. The results of this coordination may or may not result in knowledge change.

Knowledge change: bridging evidence evaluation and hypothesis space

A key goal in scientific investigation is the discovery of new knowledge. SDE studies imitate the scientific discovery process in that participants start with an initial set of beliefs or knowledge, but these beliefs or knowledge are changed as a function of the evidence generated via observation and investigation. Numerous findings speak to the change in knowledge that occurs as a result of investigation and inferences drawn from different types of evidence (e.g., anomalous or surprising, physical versus social domains) that are evaluated and interpreted in the context of existing knowledge and beliefs.

For children and adults, it is more difficult to integrate evidence that disconfirms a prior causal theory than evidence that disconfirms a prior non-causal theory. The former case involves restructuring a belief system, while the latter involves incorporating a newly discovered causal relation (Holland et al., 1986; Koslowski, 1996). For example, students hold robust ideas that the weight of an object makes a difference in the period of a pendulum, and that heavy objects fall (and sink) faster than light objects. When confronted with evidence that disconfirms those beliefs, students may struggle with how to reconcile belief with newly generated evidence. In contrast, many children do not believe that string length is causal in the case of pendulums, or that wheel size is causal in the case of car speed. When experimental evidence shows that these variables do make a difference, they are more likely to accept the evidence as veridical—they are less likely to distort or misinterpret evidence in such cases. Such tendencies may be related to Kanari and Millar's (2004) speculation that school science biases students to be focus on factors that "make a difference." As mentioned previously, valid exclusion inferences require one to consider *patterns* of evidence (Kuhn et al., 1995; Schauble, 1996), whereas a single trial showing a difference (expected or not) may be sufficient to change one's belief from non-causal to causal. Belief change for both children and adults is far more likely for the variables for which individuals had no expectations (e.g., Schauble, 1996).

As suggested by the review of evidence evaluation studies, and further supported by SDE studies, some individuals cannot or do not disregard prior theories or expectations when they evaluate evidence. Children and adults differentially attend to variables that they already believe to be causal. More experiments are conducted and more inferences are made about factors that are selected based on prior belief or expectation. Concern for theoretical relationships is also evident by references to causal mechanisms. For example, in Schauble's (1996) study using the domains of hydrostatics and hydrodynamics, references were made to unobservable forces such as "currents," "resistance," "drag," and "aerodynamics" to help explain and make sense of the evidence.

One of the features of the causal systems used in SDE research is that they may be deliberately chosen to exploit known misconceptions (e.g., Schauble et al., 1991). Any time a finding is unexpected, it could by definition be considered an *anomaly*. "Surprising results" are an impetus for conceptual change in real science (Klahr et al., 1993; Klayman & Ha, 1987) and are discussed by Kuhn (1962) in the history and philosophy of science. A specific task variant that has been explored has been to examine the effect of "anomalous" evidence on students' reasoning and knowledge acquisition. For example, Penner and Klahr (1996b) used a task for which there are rich prior beliefs—most children believe heavy objects sink faster than light objects. For steel objects, sink times for heavy and light objects are very similar. Only 8 of 30 participants selected that particular set of objects to test, and all noted that the similar sinking time was unexpected. The process of knowledge change was not straightforward. For example, some students suggested that the size of the smaller steel ball offset the fact that it weighed less because it was able move through the water as fast as the larger, heavier steel ball. Other students tried to update their knowledge by concluding that both weight and shape make a difference. That is, there was an attempt to reconcile the evidence with prior knowledge and expectation by appealing to causal mechanisms, alternate causes or enabling conditions.

It is important to note that the children in the Penner and Klahr study did in fact *notice* the surprising finding. For the finding to be "surprising" it had to be noticed, and therefore these participants did not ignore or misrepresent the data. They tried to make sense of the surprising finding by acting as a theorist who conjectures about the causal mechanisms or boundary conditions (e.g., shape) to account for the results of the experiment. In Chinn and Malhotra's (2002a) study of students' evaluation of observed evidence (e.g., watching two objects fall simultaneously), the process of observation (or "noticing") was found to be important for conceptual change (see also Kloos & Van Ordern, 2005).

Echevarria (2003) examined seventh-graders' reactions to anomalous data in the domain of genetics to see if they served as a "catalyst" for knowledge construction. In general, the number of hypotheses generated, the number of tests conducted, and the number of explanations generated were a function of students' ability to encounter, notice, and take seriously an anomalous finding. The majority of students (80%) developed some explanation for the pattern of anomalous data. For those who were unable to generate an explanation, it was suggested that initial knowledge was insufficient and therefore could not undergo change as a result of encountering "anomalous" evidence. Analogous to case studies in the history of science (e.g., Simon, 2001) these students' ability to notice and explore anomalies was related to their level of domain-specific knowledge (as suggested by Pasteur's oft quoted "serendipity favors the prepared mind"). Surprising findings were associated with an increase in hypotheses and experiments to test these potential explanations, but without the domain knowledge to "notice," anomalies could not be exploited.

Research on SDE has included tasks in both physical and social science domains. Kuhn et al. (1995) found differential performance for physical domains (e.g., speed of cars) and social domains (e.g., determining the factors that affect students' school achievement). Performance in the social domains was inferior for both fourth-graders and adults (community college students). Percentage of valid inferences was lower than in the physical domains, participants made very few exclusion inferences (i.e., the focus was on causal inferences) and causal theories were difficult to relinquish, whether they were previously held or formed on the basis of (often insufficient) experimental evidence. Kuhn and Pearsall (1998) found that when fifth-graders investigated these same physical and social domains, that greater metastrategic understanding and strategic performance (e.g., valid inferences) were evident when working in the physical domains. Kuhn et al. (1995) suggested that adults and fourth-graders had a richer and varied array of existing theories in the social domains and that participants may have had some affective investment in their theories about school achievement and TV enjoyment, but not for their theories about the causal factors involved in the speed of boats or cars.

Although the influence of different types of domain knowledge on scientific thinking has not been systematically explored in SDE studies, this is an area that warrants further attention, especially if such findings are to be relevant for classroom science or relevant for students' long-term scientific literacy. Students learn science from many domains, and will go on to read and evaluate scientific findings from natural, physical, and social domains. For example, Zimmerman, Bisanz, and Bisanz (1998) found that undergraduates rated the credibility of physical science reports to be more credible than social science reports. The written justifications for these credibility ratings could be coded as appealing to elements of scientific research such as methods, data, and theory. An additional "belief" category was created because of the number of statements of belief or disbelief in the reported conclusion. Such belief justifications were much more common for social science research. For example, rather than critically evaluate a report on the benefits of meditation for senior citizens, one quarter of the sample of 128 students found it credible because of prior belief. Science education K-12 focuses largely on the natural and physical sciences, but much of the research students will be exposed to after graduation will be from the social and medical sciences.

Rozenblit and Keil (2002) presented a set of 12 studies that show that the "illusion of explanatory depth" varies as a function of domain. Although confidence in one's knowledge may not seem relevant to reasoning, the fact that prior knowledge has been shown to have a significant influence on scientific thinking tasks makes it an important factor to consider. Rozenblit and Keil found that the "degree of causal transparency for a system" (p. 554) (i.e., having visible parts rather than hidden causal mechanisms) was related to individuals' overconfidence about their understanding, controlling for familiarity and complexity. They suggested that people are more likely to think they understand quite well phenomena that are easy to visualize or to "mentally animate." This finding has implications in that the specific domain of prior knowledge (e.g., social versus physical) may be a factor in more or less proficient reasoning and conceptual development.

Individual approaches to self-directed experimentation

Experimentation has been characterized as a goal-directed problem solving activity (e.g., Klahr, 2000; Simon, 2001). The question then becomes, *Which goal?* Characteristic ways of approaching SDE tasks have been found that are related to an individual's perceived *goal.* As has been discussed, the selection of hypotheses, variable, designs and inferences may be a function of prior knowledge, but that prior knowledge also includes assumptions about what the ultimate objective of the investigation is.

Theorists versus experimenters

Simon (1986) noted that individual scientists have different strengths and specializations, but the "most obvious" is the difference between experimentalists and theorists (p.163). Klahr and Dunbar (1988) first observed strategy differences between *theorists* and *experimenters* in adults. Individuals who take a theory-driven approach tend to generate hypotheses and then design experiments to test the predictions of the hypotheses (i.e., using deduction). Experimenters tend to make data-driven discoveries, by generating data and finding the hypothesis that best summarizes or explains that data (i.e., using induction).

Dunbar and Klahr (1989) and Schauble (1990) also found that children conformed to the description of either theorists or experimenters. Penner and Klahr (1996a) had 10- to 14-year-olds conduct experiments to determine how the shape, size, material and weight of an object influence sinking times. Students' approaches to the task could be classified as either "prediction orientation" (i.e., a theorist; e.g., "I believe that weight makes a difference) or a "hypothesis orientation" (i.e., an experimenter; e.g., "I wonder if . . ."). Ten-year-olds were more likely to take a prediction (or demonstration) approach, whereas 14-year-olds were more likely to explicitly test a hypothesis about an attribute without a strong belief or need to demonstrate that belief. The age of 12 was suggested as the age at which students may begin to transition from using experiments to demonstrate a belief to using experiments for inquiry or investigation.

Zimmerman et al. (2003) could classify sixth-graders as either theorists (theory-modifying or theory-preserving) or experimenters (or "theory generating") in their approach to experimenting with three variables that did or did not influence a balance apparatus. The task was selected specifically because it was curriculum-neutral (none of the students were in classes that covered concepts of balance or torque). Students classified as theorists approached the task by explicitly stating and testing their theories about how the apparatus worked, using a combination of controlled tests and free-form exploration of the apparatus. Theory-modifying students evaluated evidence and, when based on controlled comparisons, were able to revise their theories based on the evidence they generated. In contrast, theory-preserving students would distort or interpret evidence as consistent with theory. Experimenters did not state theories in advance of evidence. Rather, they conducted controlled comparisons, determining the effects of each variable, and derived a quantitative rule (i.e., they *generated* the theory based on evidence).

Students from a curriculum that emphasized model-based reasoning and provided multiple opportunities to create and revise theories were successful at generating a quantitative rule for balance, regardless of whether they approached the task as a theorist or an experimenter. Students in a typical inquiry-based class (in which students engaged in only a single extended inquiry activity with plants) were only successful at discovering the quantitative rule when they were classified as experimenters. Because they experimented with the apparatus during a single session, students from the regular classroom only made progress if they did not have strong theoretical beliefs that they set out to demonstrate (i.e., the "prediction orientation"; Penner & Klahr, 1996a). Given more time on task, students from the regular class who took a "theorist" approach may have eventually discovered the causal status of all of the variables. Zimmerman et al. suggested that one possible reason for the success of the theorists from the model-based reasoning curriculum was due to their repeated exposure to and practice with activities that emphasized the generation, confirmation and revision of theories.

Across these studies, the general characterization of some participants as "theorists"—and that a theory-driven approach can lead to success in some discovery contexts—lends support to the idea that inadequate accounts of the development of scientific thinking will result from studying experimentation or evidence evaluation in the absence of any domain knowledge or under instructions to disregard prior knowledge. Although these patterns may characterize individuals' approaches to any given task, it has yet to be determined if such styles are idiosyncratic to the individual and would remain stable across different tasks, or if the task demands or domain changed if a different style would emerge.

Perceived goal of inquiry: scientists versus engineers

Research by Tschirgi (1980) initially suggested the possibility that the participant's goal could affect the choice of experimentation strategy. For positive outcome, participants selected the less valid HOTAT strategy. For negative outcomes, the more valid VOTAT (or CVS) strategy was used. This general pattern has been found by a number of researchers in different contexts. Schauble (1990) noted that fifth- and sixth-grade children often behaved as though their goal was to produce the fastest car in the Daytona microworld rather than to determine the causal status of each of the variables. Kuhn and Phelps (1982) noted that several children approached the colorless-fluids task as though they trying to produce the red colour rather than identifying which chemicals produced the reaction. Prior to instruction, students in the Carey et al. (1989) study behaved as though their goal was to reproduce the bubbling effect produced by mixing yeast, sugar, salt, flour and warm water—they did not distinguish between "understanding a phenomenon and producing the phenomenon" (p. 516). In several studies, Kuhn (1991, 1993a, 2001) and colleagues also reported that early in investigations, students tend to focus on desirable versus undesirable outcomes.

Schauble, Klopfer, and Raghavan (1991) addressed the issue of goals by providing fifth- and sixth-grade children with an "engineering context" and a "science context." Children worked on the canal task and the springs task. When the children worked as scientists, their goal was to determine which factors made a difference and which ones did not. As engineers, their goal was optimization, that is, to produce a desired effect (i.e., the fastest boat or the longest spring length). In the science context, children worked more systematically, by establishing the effect of each variable, alone and in combination. There was an effort to make both causal and non-causal inferences. In the engineering context, children selected highly contrastive combinations, and focused on factors believed to be causal while overlooking factors believed or demonstrated to be noncausal. Typically, children took a "try-and-see" approach to experimentation while acting as engineers, but took a theory-driven approach when acting as scientists. These findings support the idea that researchers and teachers need to be aware of what the student perceives the goal of experimentation to be: optimization or understanding. It is also a question for further research if these different approaches characterize an individual, or if they are invoked by task demand or implicit assumptions. It might be that developmentally, an engineering approach makes most sense as inquiry skills are developing. Schauble et al. (1991) found that children who received the engineering instructions first, followed by the scientist instructions, made the greatest improvements.

Summary of developmental differences and individual approaches to SDE tasks

As was found with experimentation, children and adults display intra-individual variability in strategy usage with respect to inference types. Likewise, the existence of multiple inference strategies is not unique to childhood (Kuhn et al., 1995). In general, individuals tend to focus on causal inferences early in an investigation (somewhat similar to a "confirm early, disconfirm late" heuristic), but a mix of valid and invalid inference strategies co-occur during the course of exploring a causal system. As with experimentation, the addition of a valid inference strategy to an individual's repertoire does not mean that they immediately give up the others. Early in investigations, there is a focus on causal hypotheses and inferences, whether they are warranted or not. Only with additional exposure (as with microgenetic contexts) do children start to make inferences of non-causality and indeterminacy. Knowledge change—gaining a better understanding of the causal system via experimentation—was associated with the use of valid experimentation and inference strategies. Knowledge change as a result of newly discovered evidence was also a function of one's ability to notice "surprising" or "anomalous" findings, and to use prior knowledge to reason about whether a pattern of data represented a real change or some type of random or systematic error.

The increasing sophistication of scientific thinking, whether in children or adults, involves both strategy changes and the development of knowledge. There is a dynamic interaction between the two, that is, the changes in knowledge and strategy "bootstrap" each other: "appropriate knowledge supports the selection of appropriate experimentation strategies, and the systematic and valid experimentation strategies support the development of more accurate and complete knowledge" (Schauble, 1996, p. 118).[11]

Children's performance was characterized by a number of tendencies: to generate uninformative experiments, to make judgments based on inconclusive or insufficient evidence,

[11] Detailed cases studies of individual students can be found in many SDE studies, including Schauble (1996) and Kuhn et al. (1992, 1995).

to vacillate in their judgments, to ignore inconsistent data, to disregard surprising results, to focus on causal factors and ignore noncausal factors, to be influenced by prior belief, to have difficulty disconfirming prior beliefs, and to be unsystematic in recording plans, data, and outcomes (Dunbar & Klahr, 1989; Gleason & Schauble, 2000; Keselman, 2003; Klahr et al., 1993; Kuhn et al., 2000, 1995, 1992; Penner & Klahr, 1996a; Schauble, 1990, 1996; Schauble & Glaser, 1990; Schauble et al., 1991; Zimmerman et al., 2003). In microgenetic studies, though, children in the fifth-grade or higher typically improve in the percentage of valid judgments, valid comparisons, and evidence-based justifications with repeated exposure to the problem-solving environment (Keselman, 2003; Kuhn et al., 2000, 1995, 1992; Schauble, 1990, 1996; Schauble et al., 1991).

A number of studies that followed students through repeated cycles of inquiry and all phases of the investigation showed the *co-development* of reasoning strategies and domain knowledge. Either acquisition alone will not account for the development of scientific thinking (e.g., Echevarria, 2003; Klahr et al., 1993; Kuhn et al., 1992; Metz, 2004; Penner & Klahr, 1996b; Schauble, 1996; Tytler & Peterson, 2004). The development of experimentation and inference strategies followed the same general course in children and adults (but that adults outperformed children) and that there were no developmental constraints on “the time of emergence or consolidation of the skills” involved in scientific thinking (Kuhn et al., 1995, p. 102).

Instructional and practice interventions

Metz (2004) observed that “cognitive developmental research . . . aspires to model the emergence of the children’s competence apart from any instructional intervention” (p. 222). Early studies examining children and adults’ self-directed instruction (SDE) were conducted largely in the absence of any specific instructional intervention. Children’s scientific thinking can be studied for what it informs us about the development of inductive, deductive and causal reasoning, problem solving, knowledge acquisition and change, and metacognitive and metastrategic competence. However, such studies can and should be informative with respect to the kinds of practice and instruction that may facilitate the development of knowledge and skills and the ages at which such interventions are likely to be most effective. In more recent SDE studies, there has been a shift to include instructional components to address such concerns. In this section, I will review studies that include instructional or practice interventions. Note that there exists a substantial body of research on classroom-based interventions in science education (and entire journals devoted to such research) but such studies are outside the scope of this review.

Recall that only a few studies focusing on experimentation and evidence evaluation skills incorporated instruction or practice. These studies, interestingly, foreshadowed the issues currently being investigated. Siegler and Liebert (1975) found that in the absence of instruction, students were largely unsuccessful at manipulating and isolating variables. Kuhn and Phelps (1982) reported that extended practice over several weeks was sufficient for the development and modification of experimentation and inference strategies, showing that “exercise of existing strategies in some cases will be sufficient to lead [students] to modify these strategies” (p. 3). Later SDE studies replicated the finding that frequent engagement with the inquiry environment can lead to the development and modification of cognitive strategies (e.g., Kuhn et al., 1995, 1992; Schauble et al., 1991).

Prompts as scaffolds? Kuhn and Phelps (1982) reported a variation of their procedure (Lewis, 1981, cited in Kuhn & Phelps) in which over the course of weeks, one group of children received only a simple prompt (i.e., "What to you think makes a difference?") with another group receiving the additional prompts as used by Kuhn and Phelps (e.g., "How do you think it will turn out?" and "What have you found out?", p. 33). No difference in the strategies used by these two groups was found. The presence of even the simple prompt and repeated practice led to strategic improvements.

Such prompts are used by researchers in SDE contexts in order to generate the verbal data that will serve as evidence of, for example, use of plans, intentions, rationale for the selection of variables, the use of evidence-based versus theory-based justifications for inferences, and so on. Later microgenetic studies examined children's performance in the absence of specific instructional interventions, but similar kinds of prompts were used as individuals explored a multivariable system. Klahr and Carver (1995) questioned whether the use of prompts and systematic probes do not in fact serve as a subtle form of instructional scaffolding that alerts participants to the underlying goal structure. An alternative interpretation exists for the finding of studies that report improvement in children's experimentation and inference strategies solely as a function of practice or exercise. Such prompts may cue the strategic requirements of the task or they may promote explanation or the type of reflection that could induce a metacognitive or metastrategic awareness of task demands. Unfortunately, because prompts are necessary to generate data in SDE studies, it may be very difficult to tease apart the relative contributions of practice and exercise from the scaffolding provided by researcher prompts. Gleason and Schauble (2000) used minimal intervention, but the parent-child collaborative discussion resulted in the verbal data needed to characterize dyads' performance.

Although conducted with undergraduate students, Wilhelm and Beishuizen (2004) reported no differences in learning *outcomes* for students who were and were not asked standardized questions during the process of experimenting with a computerized multivariable system. Differences in learning *processes* were found. For example, students who were not asked questions during exploration were more likely to repeat experiments. Repetition in experiments, as has been shown, may be indicative of experimentation done in the absence of plans, a less thorough search of the experiment space, and the generation of a smaller set of evidence.

In the absence of instruction or prompts, students may not routinely ask questions of themselves such as "What are you going to do next?"; "What outcome do you predict?"; "What did you learn?" and "How do you know?" Research on *self-explanations* supports this idea (e.g., Chi, de Leeuw, Chiu, & Lavancher, 1994) and moreover, that the process of self-explaining promotes understanding. Self-explanation is thought to be effective because it promotes the integration of newly learned material with existing knowledge (Chi et al., 1994). Analogously, the prompts used by researchers may serve to promote such integration. Recall that Chinn and Malhotra (2002a) incorporated different kinds of interventions aimed at promoting conceptual change in response to anomalous evidence. Only the explanation-based interventions were successful at promoting conceptual change, retention and generalization. The prompts used in microgenetic SDE studies very likely serve the same function as the prompts used by Chi et al. (1994). Incorporating such prompts in classroom-based inquiry activities could serve as a powerful teaching tool, given that the use of self-explanation in tutoring systems (human and computer interface) has shown to be quite effective (e.g., Chi, 1996; Hausmann & Chi, 2002).

Instructional interventions

Studies that compare the effects of different kinds of instruction and practice opportunities have been conducted in the laboratory, with some translation to the classroom. For example, Chen and Klahr (1999) examined the effects of direct and indirect instruction of the control-of-variables (CVS) strategy on students' (grades 2–4) experimentation and knowledge acquisition. Direct instruction involved didactic teaching of the CVS strategy along with examples and probes. Indirect (or implicit) training involved the use of systematic probes during experimentation. A control group did not receive instruction or probes. No group received instruction on domain knowledge for any task used (springs, ramps, sinking objects). CVS usage increased from 34% prior to instruction to 65% after, with 61–64% usage maintained on transfer tasks that followed after one day and again after seven months, respectively. No such gains were evident for the implicit training or control groups.

Students' mastery of CVS due to direct instruction allowed them to design unconfounded experiments, which facilitated valid causal and non-causal inferences, resulting in a change in knowledge about how various multivariable causal systems worked. Significant gains in domain knowledge were only evident for the direct instruction group. Fourth graders showed better skill retention at long-term assessment relative to second and third graders. In microgenetic contexts, extended practice has been shown to be sufficient for improving students' usage of valid experimentation and inference strategies. However, the younger children (grades 2 and 3) did not benefit as much from such exercise as older students here (grade 4), or as much as has been reported with students at the fifth-grade level or higher. It is perhaps prior to the fourth grade that self-directed experimentation in educational contexts requires the use of targeted and frequent scaffolding to ensure learning and strategic gains. The specific type of instruction and scaffolding most beneficial for younger students engaged in SDE is an empirical question awaiting further study.

Toth, Klahr, and Chen (2000; Klahr, Chen, & Toth, 2001) translated the direct instruction of CVS in the lab to a classroom environment. A classroom instructor taught the CVS strategy to fourth graders in a classroom setting, with students being assessed individually pre- and post-instruction. Toth et al. examined pre- to post-instruction gains in CVS usage, *robust use* of CVS (requiring correct justification of use), domain knowledge, and the evaluation of research designed by other children. Significant post-instruction increases were found for mean CVS usage (30–97%) and mean robust usage (6–78%). Although domain knowledge started high (79%), significant improvement (to 100%) was found. The percentage of students who were able to correctly evaluate others' research increased from 28% to 76%. Therefore, the effectiveness of a lab-based intervention could be "scaled up" to a classroom context. (See Klahr & Li, 2005, for a summary of research that alternates between the lab and the classroom.)

Klahr and Nigam (2004) explored the longer-term impact of learning CVS under two different conditions. Third- and fourth-graders engaged in self-directed experimentation to discover the factors that influence the distance a ball travels down various ramps. One group received direct instruction in CVS prior to SDE, and the control group explored the multivariable system without such training. Students in the direct instruction condition were more likely to master CVS, which resulted in better performance with respect to designing unconfounded experiments and thus making valid inferences. A minority of students (23%) in the control condition were able to master CVS. All students who attained mastery, *regardless of condition*, scored better on a transfer task that involved the evaluation of science projects completed by other students. Although the direct instruction group performed better, overall, on the immediate and transfer assessments, a quarter of the students did master CVS though exploration (which is not unexpected based on previous SDE studies, especially within the microgenetic context). Klahr and Nigam suggested that the next set of issues to address include determining the kinds of individual difference characteristics that account for some students benefiting from the discovery context, but not others. That is, which learner traits are associated with the success of different learning experiences? Answers to such questions would facilitate a match between types of students and types of pedagogy for a "balanced portfolio of instructional approaches to early science instruction" (p. 666).

Reid et al. (2003) also examined the influence of two different types of learning support on students' self-directed exploration of buoyancy. *Interpretive support* was designed to help students to access domain knowledge to generate appropriate and testable hypotheses and then develop coherent understanding of the domain. The virtual environment included a "reference book" that contained information about, for example, weight, upthrust, and motion. *Experimental support* was designed to scaffold the design of experiments, making predictions and drawing conclusions based on observations (e.g., how to vary one thing at a time) and included prompts to compare predictions to the outcome of the experiment.

The factorial combination of the presence or absences of the two support types resulted in four groups of students (i.e., no support, experimental only, interpretive only, or both supports). Reid et al. found that experimental support improved sixth-graders' performance on assessments of principled knowledge and intuitive understanding from pre- to post-test. At post-test, interpretive support was shown to improve intuitive understanding (predictions of upthrust for pairs of objects) and knowledge integration, relative to experimental support. Students were given scores for their use of experimentation strategies (e.g., use of CVS, percentage of experiment-space used). Although there was no difference for those who did or did not receive experimental support, students who had better experimentation strategies were more successful in discovering all of the causal and non-causal factors. All students benefited from participating in the scientific discovery activity, but different types of learning support facilitated performance on different outcome measures.

Ford (2005) used two different instructional methods based on different conceptualizations of experimentation. The *CVS instruction* focused on the logical aspects of experimentation, and was based on that developed by Toth et al. (2000) and Klahr et al. (2001). *Measurement instruction* focused on quantitative reasoning, including ideas such as measuring, measurement error, recording and visually displaying results. Sixth-graders were assessed with a CVS pre- and post-test, and a performance assessment along with interviews that allowed insight into students' reasoning. The multi-week units focused on determining how steepness affects the speed of a rolling ball. Students in the CVS instruction showed greater improvements on the CVS test, and their reasoning was consistent with a CVS strategy of identifying the focal variable and making sure the levels are different, while all other variables levels are the same.

The measurement group also had a specific pattern of reasoning that included a concern for the logic of experimentation, but it was also consistent with the idea that individuals are concerned with alternate causes and plausible sources of variation in the variable that is measured. In a performance assessment that involved an empirical investigation (i.e., "how does the release height of a ball affect bounce?"), other key differences were found for the two forms of instruction. Students in the measurement group they were more likely to

(a) quantify the outcome variable, (b) suggest a standardized release height, (c) plan multiple trials with the goal of verification, (d) use records to represent the results, and (e) produce conclusions based on summary of systematic and repeatable procedures. The identification of additional "fundamental aspects of practice" provides resources for guiding and improving instruction (Ford, 2005).

Practice interventions

Kuhn and her colleagues have been exploring a number of interventions aimed at increasing students' metacognitive and metastrategic competence. For example, Kuhn et al. (2000) incorporated performance-level practice and metastrategic-level practice in tasks explored by sixth- to eighth-grade students. Performance-level exercise consisted of standard exploration of the task environment (typical of SDE studies). Metalevel practice consisted of paper-and-pencil scenarios in which two individuals disagree about the effect of a particular feature in a multivariable situation. Students then evaluate different strategies that could be used to resolve the disagreement. Such scenarios were provided twice a week during the course of ten weeks. Although no differences between the two types of practice were found in the number of valid inferences (i.e., performance), there were more sizeable differences in measures of understanding of task objectives and strategies (i.e., metastrategic understanding). Keselman (2003) compared performance-level exercise (control) with two practice conditions: one that included direct instruction and practice at making predictions and one with prediction practice only. Sixth graders experimented with a multivariable system (an earthquake forecaster). Students in the two practice conditions showed better performance on a metaleval assessment relative to the control group. Only the direct instruction group showed an increase in use of evidence from multiple records and the ability to make correct inferences about non-causal variables.

Kuhn and Dean (2005) incorporated a very simple but effective instructional intervention. Over the course of twelve sessions, students interacted with a virtual environment to understand the effect of five binary variables on earthquake risk. Sixth graders were either given self-directed practice (control) or they were provided with the suggestion to focus on just one of the variables at a time. The suggestion to focus attention on just one variable was very effective: All students in the suggestion group were able to use CVS to design unconfounded experiments, compared to 11% in the control. CVS usage led to an increase in valid inferences for the intervention group, at immediate and delayed (3 months) assessment. Kuhn and Dean concluded that the manipulation influenced the question-formulation phase because it suggested to students *why* a strategy should be used. It is also possible that this simple instruction invoked strategic mastery, which as has been shown, bootstraps the ability to make valid inferences, such that knowledge of the domain accumulates which in turn facilitates advanced exploration and hypothesis generation. Extended engagement alone resulted in minimal progress, confirming that even minor prompts and suggestions represent potentially powerful scaffolds in the context of self-directed investigations.

Kuhn (2005b) has also been exploring the kinds of educational opportunities to engage urban, at-risk children in science. This is a particularly important question in the age of "no child left behind." Sixth-graders investigated a multivariable environment in a non-traditional domain that the students would find interesting. Students acted as advisors to a music club and investigated the features that had an influence on sales (e.g., cover illustration, color, format). The idea was to introduce students to inquiry in a topic that would engage them, while providing appropriate scaffolds. The scaffolds consisted of weekly conceptual goals, such as making explicit connections between claims and evidence, generalizing findings, and understanding additive multivariable causality. It was hypothesized that the development of inquiry skills would then transfer to a more traditional science domain. After 12 weeks of scaffolding on the music club software, students were then exposed to the earthquake forecaster software. Performance on this transfer task was compared to students who had unscaffolded practice and students who were not exposed to multivariable inquiry. The group of students that received scaffolding outperformed the students in the two control groups on a number of performance indicators, including intentions to focus on one variable, use of CVS, and making valid inferences. Three months later, there was a general decline in performance, but it was still generally superior to the control groups.

Start again here

Scaffolding in a classroom-based design experiment: an example

Metz (2004) conducted extensive analyses of children's interview data about their own investigations that they designed and executed. Second- and fourth/fifth-graders took part in a curriculum unit on animal behavior that emphasized domain knowledge, whole-class collaboration, scaffolded instruction, and discussions about the kinds of questions that can and cannot be answered by observational records. Pairs or triads of students then developed a research question, designed an experiment, collected and analyzed data, and presented their findings on a research poster.

As discussed previously, it has been suggested that developmental differences in scientific reasoning may be a function of task demands (e.g., Ruffman et al., 1993; Sodian et al., 1991). In contrast, Metz (2004) argues that the reason that researchers sometimes demonstrate that children fail to reason in a normative way on laboratory tasks my be due to the fact that they are *not demanding enough*: "This weak knowledge (including ignorance of the relevant variables and construct) has resulted in poorer reasoning and thus an underestimation of reasoning capacities ... [and] has resulted in unnecessarily watered-down curricula [which] have led to less opportunity to learn, and thus weaker domain-specific knowledge, again undermining children's scientific reasoning" (p. 284). Such classroom-based design experiments provide evidence that elementary school children can successfully participate in authentic inquiry, but that particular kinds of scaffolding are needed to support children's abilities to engage in independent investigations and to help students view science as a "way of knowing" (Metz, 2004).

The set of studies reviewed in this section were all conducted to ascertain the types of interventions that might promote the development of scientific thinking. The study by Metz (2004) serves as the "existence proof" of what even young children are capable of with appropriate classroom support, scaffolding, and collaborative inquiry. Lab-based studies have also explored interventions, and can be primarily categorized as one of two types. The first type involves a focus on promoting strategic skills, such as CVS (e.g., Klahr & Nigam, 2004), and the other type is intended to foster meta-strategic understanding – that is, the goal is to foster an awareness of the appropriateness a particular strategy.

Summary and conclusions

My goal in this review was to provide an overview of research on the development of scientific thinking, with a particular focus on studies that address children's investigation and inference skills. Although scientific thinking is multifaceted and a full account may need to take into account research on, for example, explanation, epistemology, argumentation, the nature of science, and conceptual understanding (and "misconceptions") in numerous domains of science, the focus of this review was the extensive literature on experimentation skills, evidence evaluation, and self-directed experimentation (SDE).

How do children learn science?

Recent approaches to the study of scientific thinking situate students in a simulated-discovery context, in which they investigate a multivariable causal system through active or guided experimentation. In these contexts, the development of both strategies and conceptual knowledge can be monitored. These two aspects of cognition *bootstrap* one another, such that experimentation and inference strategies are selected based on prior conceptual knowledge of the domain. These strategies, in turn, foster a deeper understanding of the system via more sophisticated causal or conceptual understanding, which (iteratively) foster more sophisticated strategy usage.

One of the continuing themes evident from studies on the development of scientific thinking is that children are far more competent than first suspected, and likewise, adults are less so. This characterization describes cognitive development in general and scientific thinking in particular. A robust finding is that during this long developmental trajectory, there is both inter- and intra-individual *variability* in performance, particularly with respect to inference and experimentation strategies. A number of generalizations can be extracted that address the issue of how children learn scientific inquiry skills. Children may have different assumptions and beliefs about the goals of experimentation and this claim is supported by their (a) evolving understanding of the nature of science and what experimentation is for (e.g., for demonstrating the correctness of current belief; producing an outcome versus understanding a phenomenon); (b) tendency to focus on outcomes by producing desired effects and reducing undesired effects; (c) tendency to ignore non-causal factors and focus on causal factors or what "makes a difference," and in doing so may, in some instances (d) tend to incorrectly encode, misinterpret, or distort evidence to focus on causes. Characteristics of prior knowledge such as (e) the type, strength, and relevance are potential determinants of how new evidence is evaluated and whether "anomalies" are noticed and knowledge change occurs as a result of the encounter. There are both (f) rational and irrational responses to evidence that disconfirms a prior belief. At the meta-level, children may not be aware of their own memory limitations and therefore may be unsystematic in (g) recording plans, designs and outcomes, and may fail to (h) consult such records. Likewise, there is a slow developmental course for the (i) metacognitive understanding of theory and evidence as distinct epistemological entities and the (j) metastrategic competence involved with understanding when and why to employ various strategies.

Scientific thinking involves a complex set of cognitive and metacognitive skills, and the development and consolidation of such skills require a considerable amount of exercise and practice. Given these generalizations about children's performance, researchers will be in a better position to explore the kinds of scaffolding, practice and instructional interventions that may be candidates to facilitate the development of increasingly proficient scientific thinking.

How can cognitive developmental research inform science teaching and learning?

Given the complexity of coordinating the cognitive and metacognitive skills involved in scientific thinking, and the potentially long developmental trajectory, it is necessary to consider the kinds of educational experiences that will foster and support the development of inquiry, experimentation, evidence evaluation, and inference skills. Basic research on children's developing scientific thinking can serve as a guide for targeting interventions. Given the numerous components and the iterative nature of investigation in SDE contexts, different researchers have targeted different phases of the inquiry cycle. Some of the research on instructional interventions reviewed here capitalized on basic findings and generalizations discussed in the previous section – findings related to the implementation of strategies, the role of prior knowledge, and the importance of meta-level understanding—and the general characterization of the bootstrapping of the conceptual and the strategic. Similarly, recent conceptual and empirical work points to the necessity for skilled scientific thinking to include flexible, *metacognitive* and *metastrategic* knowledge (Kuhn, 2002). Current research and curriculum development has been focused at exploring the types of scaffolding to support students' developing metacognitive abilities (e.g., Kolodner et al., 2003; Raghavan & Glaser, 1995; White & Frederiksen, 1998).

Cognitive developmental research has the potential to inform science teaching, as illustrated by some of the intervention research reviewed here. For example, beginning with the initial encoding of information, interventions can promote students' observational abilities and lead to appropriate encoding of information (e.g., Chinn & Malhotra, 2002a), which was shown to facilitate inferences, generalizations and retention. Domain knowledge and strategic performance have been shown to bootstrap one another, and as such interventions have targeted facilitating domain knowledge (e.g., Echevarria, 2003; Metz, 2004), strategy development (e.g., Chen & Klahr, 1999; Toth et al., 2000) or both (e.g., Reid et al., 2003). Metacognitive understanding and metastrategic competence have also been targeted as a means of promoting meta-level understanding (e.g., Keselman, 2003; Kuhn & Dean, 2005).

The summary of findings described above could be used to generate and target specific interventions. For example, given the tendency to initially focus on causal factors, students could be allowed to begin investigations of factors that make a difference, but then be guided into further investigations of how one determines that a factor is not causal and how to examine cumulative records to make such inferences. Given that children do not spontaneously record important information about their investigations, interventions could capitalize on initial investigations without records and then compare those to investigations in which thorough records are kept. Numerous additional examples could be suggested, but the point is to demonstrate that basic findings can be fertile source of research questions that can be explored and applied to teaching and learning situations. As research accumulates from laboratory studies on the conditions which support scientific thinking and conceptual change, continued research will need to explore the best ways to teach such skills. If science education is to be reformed on the basis of evidence-based research, specific questions about instruction will need to be tested empirically (e.g., How much support and structure are optimal? How much teacher control? What kinds of scaffolds and prompts are sufficient? Should domain knowledge and skills be taught concurrently or separately?).

Future directions for research

In the previous section, hypothetical examples of potential future research questions were briefly described to illustrate the potential for synergistic research (see also Klahr & Li, 2005). In this final section, I would like also to address some of the larger conceptual issues for future research to address. These include (a) methodological and conceptual issues in research on instructional interventions; and (b) the authenticity of scientific reasoning tasks used in schools and laboratories.

Which types of instruction are best?

This may be the key question that directs future research. Since scientific thinking involves a collection of intellectual skills, some of which do not "routinely develop" (Kuhn & Franklin, 2006), it is absolutely essential that basic research on understanding children's scientific thinking be used engineer better instructional interventions. Such studies may then be used as a source of further basic questions (Klahr & Li, 2005). A question about the relative efficacy of different interventions—whether they be prompts, scaffolds, didactic instruction or opportunities for particular types of practice—is a far trickier endeavor than would appear on the surface.

Recall that in the literature on evidence evaluation skills, numerous conceptual and methodological issues needed to be resolved to advance our understanding of the development of such skills (e.g., what is a rational or an irrational response to anomalous evidence? Does evidence always take precedence over prior knowledge?). Even current writings conflate the two connotations of theory-evidence coordination (i.e., one as inductive causal inference, one as epistemological categories). The issue of the best way to assess the effectiveness of instructional interventions will be the next issue in need of resolution, potentially in a joint effort by researchers and educators. A diversity of tasks are used in research on scientific thinking—tasks used to assess initial understanding, tasks used to exercise developing strategies and knowledge, tasks used to assess effectiveness of interventions, and tasks used to show the transfer and/or retention. Each of these tasks has the potential to be interpreted in multiple ways (e.g., as a valid measure of transfer, a valid measure of strategic competence).

Analogous to what was required in the evidence evaluation literature – discussions and negotiations about how to operationally define terms such as "direct instruction" or "transfer" will be needed in order to measure the success of various strategic, knowledge-based, or meta-level interventions. For example, Klahr and Li (2005) outlined different media reactions to the intervention studies conducted by Klahr and colleagues (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004; Triona & Klahr, 2003) based on multiple connotations of such terms (e.g., "discovery learning").

Similarly, the commentary between Klahr (2005b) and Kuhn (2005a) represents the types of debate and dialogue that will need to become an essential part of the next generation of research on evidence-based interventions (see also Dean & Kuhn, 2007). As mentioned previously, the intervention used by Kuhn and Dean (2005) included a simple prompt to students to "try to find out about just one feature to start." Whereas Kuhn and Dean concluded that this prompt invoked a metastrategic understanding, an alternate interpretation is that it simply invokes a strategic focus to vary one feature at a time (i.e., CVS). A similar discussion ensued in the commentary between Klahr (2005b) and Kuhn (2005a). Unfortunately, this is a definitional or interpretive issue rather than one that can be resolved by appealing to the data. A similar definitional issue at the core of this commentary is what counts as "transfer." Kuhn and Dean (2005) asserted that it was not possible to make comparisons with the work of Klahr and his colleagues because they did not provide specific data on transfer. Barnett and Ceci (2002), however, used the Chen and Klahr (1999) study in a detailed discussion of the nine relevant dimensions they proposed to classify transfer studies. Despite the proposed framework to make sense of the enormous transfer literature (spanning at least a century), an operational definition of "transfer" has not been widely accepted (Barnett & Ceci, 2002; Klahr, 2005b) and as Kuhn (2005a) notes, "evaluations of transfer depend on one's conception of the competency that is undergoing transfer, as well as the transfer data themselves" (p. 873).

As a final example, the use of the term "self-directed experimentation" as used widely within this review may be subject to interpretation. As noted earlier, the use of prompts and questions on such tasks to generate verbal data used to characterize performance may lead one to believe that there is nothing "self directed" about the endeavor. That is, one could argue that a lay (or social science) version of Heisenberg's Uncertainty Principle is at work, such that one cannot observe children's performance without changing what is being observed or measured.

In parallel, these same definitional issues will undoubtedly be (or are) of concern to researchers and educators who study educational assessment in science. In an educational climate that endorses increased standardized testing as one method of accountability, assessments of scientific thinking will be subject to the same discussions and disagreements about whether they are valid measures. The selection of terms, descriptors and operational definitions will become increasingly important. (Klahr & Li (2005) suggest that we follow the lead of physicists who invent novel terms like "lepton" or "quark" that cannot be subject to alternate interpretation.)

Authentic inquiry

There has been an increased interest in which features of authentic science should be incorporated into classroom and laboratory tasks (e.g., Chinn & Malhotra, 2001, 2002b; Kuhn, 2002; Kuhn & Pearsall, 2000; Metz, 2004; Tytler & Peterson, 2004; Zachos, Hick, Doane, & Sargent, 2000) and a call to incorporate more features of authentic science into educational contexts (see Chinn & Hmelo-Silver, 2002). Chinn and Malhotra (2001, 2002b) outlined the features of *authentic* scientific inquiry and compared these features to those used in classrooms and those used in cognitive developmental studies of scientific thinking. Although the tasks used by researchers (i.e., such as those reviewed here) were found to have more features of genuine research than tasks used in schools, Chinn and Malhotra argued that if schools do not focus on these "core attributes," then the cognitive processes developed will be very different from those used in real inquiry, and moreover, students may develop epistemological understanding that is not just different—but antithetical to that of authentic science.

Authentic scientific inquiry often requires the used of statistical procedures and statistical reasoning. Chinn and Malhotra (2001) mention "transforming observations," but certain sciences rely on statistics to support reasoning (e.g., Abelson, 1995). Kanari and Millar (2004) argued that tasks used in many previous studies should be classified as "logical reasoning" tasks because participants do not evaluate *numerical data*. In their view, authentic scientific thinking involves an evaluation of primary data sources. That is, "the effect size matters." As laboratory and classroom and lab tasks incorporate the evaluation of numerical data (e.g., Masnick & Klahr, 2003; Masnick & Morris, 2002; Schauble, 1996) issues that parallel the history of science and the need for statistics will emerge (see Salsburg, 2001, for an informal history). How can students know which differences matter without explicit instruction in statistics? Separating random error from true effects is not a skill that K-8 students spontaneously engage in without scaffolding (Metz, 1998) but emerging understanding is evident (Masnick & Klahr, 2003). Some investigations along this line have begun (e.g., Lajoie, 1998; Petrosino et al., 2003), but it is an area for continued investigation.

Lehrer, Schauble, and Petrosino (2001) recently initiated the question of how much emphasis should be placed on *experimentation* in science education. They suggested that experimentation can (or should) be thought of as a form of argument. That is, the experiment should be more closely aligned with the idea of *modeling* rather than the canonical method of investigation. As discussions turn to what is authentic for elementary- and middle-school science classes, this issue will need to be considered and revisited. Simon (2001) suggested that the best instruction in science will model the actual practice of science, but not the *stereotypes* or the standard prescriptive rules of what science is about. A current snapshot of the literature would support such claims about the primacy of experimentation as the prescribed method. Simon, in contrast, suggested that students need to learn science in contexts in which they are able to find patterns in the world, where curiosity and surprise are fostered—such contexts would be "authentic."

The identification of these authentic features can be used to guide the creation of classroom and laboratory tasks. The empirical issue that remains is whether, and to what extent, the inclusion of these various core attributes fosters more proficient scientific thinking, and whether they promote a more accurate understanding of the nature of science. An additional implication of the call to incorporate more authentic features is in that "there is no way to condense authentic scientific reasoning into a single 40- to 50-min science lesson" (Chinn & Malhotra, p. 213). Curricula will need to incorporate numerous composite skills, and further research will be needed to determine in what order such skills should be mastered, and which early acquisitions are most effective at supporting the development of subsequent acquisitions.

One goal of contemporary science education is to produce "scientifically literate" adults. Although all students do not pursue careers in science, the thinking skills used in scientific inquiry can be related to other formal and informal thinking skills (e.g., Kuhn, 1993a, 1993b, 2002). Recent efforts to reform and improve the way science is taught will ensure that even those who do not pursue a career in science will benefit from the skills that can be taught in the classroom (Ford & Forman, 2006; Metz, 2004; ONeill & Polman, 2004). By focusing on interventions that encourage the development and practice of investigation and inference skills—along with the metalevel understanding that such skills allow one to recognize the value of inquiry—science education will become increasingly relevant to the needs of all students.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299–352.

American Association for the Advancement of Science (1990). *Science for all Americans: Project 2061*. New York: Oxford University Press.

American Association for the Advancement of Science (1993). *Benchmarks for Science Literacy*. New York: Oxford University Press.

Amsel, E., & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development, 11*, 523–550.

Aoki, T. (1991). The relation between two kinds of U-shaped growth curves: Balance-scale and weight-addition tasks. *The Journal of General Psychology, 118*, 251–261.

Azmitia, M., & Crowley, K. (2001). The rhythms of scientific thinking: A study of collaboration in an earthquake microworld. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 51–81). Mahwah, NJ: Lawrence Erlbaum.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612–637.

Brewer, W. F., & Samarapungavan, A. (1991). Children's theories vs. scientific theories: Differences in reasoning or differences in knowledge. In R. R. Hoffman & D. S. Palermo (Eds.), *Cognition and the symbolic processes* (pp. 209–232). Hillsdale, NJ: Lawrence Erlbaum.

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: Cambridge University Press.

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

Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology, 21*, 13–19.

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.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405.

Chen, Z., & Klahr, D. (1999). All other things being equal: Children's acquisition of the control of variables strategy. *Child Development, 70*, 1098–1120.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10*, 33–49.

Chi, M. T. H., de Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.

Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching, 35*, 623–654.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction, 19*(3), 323–393.

Chinn, C. A., & Hmelo-Silver, C. E. (2002). Authentic inquiry: Introduction to the special section. *Science Education, 86*, 171–174.

Chinn, C. A., & Malhotra, B. A. (2001). Epistemologically authentic scientific reasoning. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 351–392). Mahwah, NJ: Lawrence Erlbaum.

Chinn, C. A., & Malhotra, B. A. (2002a). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology, 94*, 327–343.

Chinn, C. A., & Malhotra, B. A. (2002b). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*, 175–218.

Corrigan, R., & Denton, P. (1996). Causal understanding as a developmental primitive. *Developmental Review, 16*, 162–202.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition, 23*, 646–658.

Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education* doi:10.1002/sce.20194.

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

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science, 17*, 397–434.

Dunbar, K., & Klahr, D. (1989). Developmental differences in scientific discovery strategies. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 109–143). Hillsdale, NJ: Lawrence Erlbaum.

Echevarria, M. (2003). Anomalies as a catalyst for middle school students' knowledge construction and scientific reasoning during science inquiry. *Journal of Educational Psychology, 95*, 357–374.

Ford, M. J. (2005). The game, the pieces, and the players: Generative resources from alternative portrayals of experimentation. *The Journal of the Learning Sciences, 14*, 449–487.

Ford, M. J., & Forman, E. A. (2006). Redefining disciplinary learning in classroom contexts. In J. Green & A. Luke (Eds.), Review of research in education, Vol. 30.

Garcia-Mila, M., & Andersen, C. (2007). Developmental change in notetaking during scientific inquiry. *International Journal of Science Education.*

Gelman, S. A. (1996). Concepts and theories. In R. Gelman & T. Kit-Fong Au (Eds.), *Perceptual and cognitive development: Handbook of perception and cognition* (2nd ed., pp. 117–150). San Diego, CA: Academic Press.

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models.* Hillsdale, NJ: Lawrence Erlbaum.

Gleason, M. E., & Schauble, L. (2000). Parents' assistance of their children's scientific reasoning. *Cognition & Instruction, 17*(4), 343–378.

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*, 620–629.

Grotzer, T. (2003). Learning to understand the forms of causality implicit in scientifically accepted explanations. *Studies in Science Education, 39*, 1–74.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition, 61*, 233–259.

Hausmann, R. G., & Chi, M. T. H. (2002). Can a computer interface support self-explaining? *Cognitive Technology, 7*, 4–14.

Hirschfeld, L. A., & Gelman, S. A. (Eds.). (1994). *Mapping the mind: Domain specificity in cognition and culture.* New York: Cambridge University Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction.* Cambridge, MA: The MIT Press.

Hume, D. (1988/1758). *An enquiry concerning human understanding.* Buffalo, NY: Prometheus Books.

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

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching, 41*, 748–769.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge: MIT Press.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28*, 107–128.

Keselman, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *Journal of Research in Science Teaching, 40*, 898–921.

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development, 71*, 1347–1366.

Klaczynski, P. A., & Narasimham, G. (1998). Development of scientific reasoning biases: Cognitive versus ego-protective explanations. *Developmental Psychology, 34*, 175–187.

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

Klahr, D. (2005a). A framework for cognitive studies of science and technology. In M. Gorman, R. D. Tweney, D. C. Gooding, & A. P. Kincannon (Eds.), *Scientific and technological thinking* (pp. 81–95). Mawah, NJ: Lawrence Erlbaum.

Klahr, D. (2005b). Early science instruction: Addressing fundamental issues. *Psychological Science, 16*, 871–872.

Klahr, D., & Carver, S. M. (1995). Scientific thinking about scientific thinking. *Monographs of the Society for Research in Child Development, 60*, 137–151.

Klahr, D., Chen, Z., & Toth, E. E. (2001). From cognition to instruction to cognition: A case study in elementary school science instruction. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 209–250). Mahwah, NJ: Lawrence Erlbaum.

Klahr, D., & Dunbar, K. (1988). Dual search space during scientific reasoning. *Cognitive Science, 12*, 1–48.

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

Klahr, D., & Li, J. (2005). Cognitive research and elementary science instruction: From the laboratory, to the classroom, and back. *Journal of Science Education and Technology, 4*, 217–238.

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

Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical vs. virtual materials in an engineering design project by middle school students. *Journal of Research in Science Teaching, 44*, 183–203.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis-testing. *Psychological Review, 94*, 211–228.

Kloos, H., & Van Ordern, G. C. (2005). Can a preschooler's mistaken belief benefit learning? *Swiss Journal of Psychology, 64*, 195–205.

Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology, 64*, 141–152.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., et al. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting Learning by Design into practice. *Journal of the Learning Sciences, 12*, 495–547.

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

Koslowski, B., & Masnick, A. (2002). The development of causal reasoning. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 257–281). Oxford: Blackwell Publishing.

Koslowski, B., & Okagaki, L. (1986). Non-Humean indices of causation in problem-solving situations: Causal mechanisms, analogous effects, and the status of rival alternative accounts. *Child Development, 57*, 1100–1108.

Koslowski, B., Okagaki, L., Lorenz, C., & Umbach, D. (1989). When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development, 60*, 1316–1327.

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

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

Kuhn, D. (1991). *The skills of argument*. New York: Cambridge University Press.

Kuhn, D. (1993a). Science as argument: Implications for teaching and learning scientific thinking. *Science Education, 77*, 319–337.

Kuhn, D. (1993b). Connecting scientific and informal reasoning. *Merrill-Palmer Quarterly, 39*, 74–103.

Kuhn, D. (2001). How do people know? *Psychological Science, 12*, 1–8.

Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 371–393). Oxford: Blackwell Publishing.

Kuhn, D. (2005a). What needs to be mastered in mastery of scientific method? *Psychological Science, 16*, 873–874.

Kuhn, D. (2005b). *Education for Thinking*. Cambridge, MA: Harvard University Press.

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

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

Kuhn, D., & Dean, D. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition & Development, 5*, 261–288.

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16*, 866–870.

Kuhn, D., Franklin, S. (2006). The second decade: What develops (and how). In: W. Damon, R.M. Lerner, (Series Eds), D. Kuhn & R. S. Siegler (Vol. Eds), *Handbook of child psychology: Vol. 2. Cognition, perception and language* (6th ed.) (pp. 953-993). Hoboken, NJ: John Wiley & Sons.

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, 60*, 1–128.

Kuhn, D., & Ho, V. (1980). Self-directed activity and cognitive development. *Journal of Applied Developmental Psychology, 1*, 119–130.

Kuhn, D., & Pearsall, S. (1998). Relations between metastrategic knowledge and strategic performance. *Cognitive Development, 13*, 227–247.

Kuhn, D., & Pearsall, S. (2000). Developmental origins of scientific thinking. *Journal of Cognition and Development, 1*, 113–129.

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

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

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480–498.

Lajoie, S. (Ed.). (1998). . 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: Lawrence Erlbaum.

Li, J., Klahr, D., & Siler, S. (2006). What lies beneath the science achievement gap? The challenges of aligning science instruction with standards and tests. *Science Educator, 15*, 1–12.

Lien, Y., & Cheng, P. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology, 40*, 87–137.

MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology, 49*, 259–287.

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*, 67–98.

Masnick, A. M., & Morris, B. J. (2002). Reasoning from data: The effect of sample size and variability on children's and adults' conclusions. In *Proceedings of the 24th annual conference of the Cognitive Science Society* (pp. 643–648).

McNay, M., & Melville, K. W. (1993). Children's skill in making predictions and their understanding of what predicting means: A developmental study. *Journal of Research in Science Teaching, 30*, 561–577.

Metz, K. E. (1998). Emergent understanding of attribution and randomness: Comparative analysis of the reasoning of primary grade children and undergraduates. *Cognition and Instruction, 16*, 285–365.

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

Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

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

National Research Council (2000). *Inquiry and the national science standards*. Washington, DC: National Academy Press.

Okada, T., & Simon, H. A. (1997). Collaborative discovery in a scientific domain. *Cognitive Science, 21*, 109–146.

ONeill, D. K., & Polman, J. L. (2004). Why educate "little scientists?" Examining the potential of practice-based scientific literacy. *Journal of Research in Science Teaching, 41*, 234–266.

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

Penner, D. E., & Klahr, D. (1996b). When to trust the data: Further investigations of system error in a scientific reasoning task. *Memory & Cognition, 24*, 655–668.

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

Pfundt, H., & Duit, R. (1988). *Bibliography: Students' alternative frameworks and science education* (2nd ed.). Kiel: Institute for Science Education.

Raghavan, K., & Glaser, R. (1995). Model-based analysis and reasoning in science: The MARS curriculum. *Science Education, 79*, 37–61.

Reid, D. J., Zhang, J., & Chen, Q. (2003). Supporting scientific discovery learning in a simulation environment. *Journal of Computer Assisted Learning, 19*, 9–20.

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*, 521–562.

Ruffman, T., Perner, J., Olson, D. R., & Doherty, M. (1993). Reflecting on scientific thinking: Children's understanding of the hypothesis-evidence relation. *Child Development, 64*, 1617–1636.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: W.H. Freeman.

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

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

Schauble, L., & Glaser, R. (1990). Scientific thinking in children and adults. *Contributions to Human Development, 21*, 9–27.

Schauble, L., Glaser, R., Duschl, R. A., & Schulze, S. (1995). Students' understanding of the objectives and procedures of experimentation in the science classroom. *Journal of the Learning Sciences, 4*, 131–166.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *Journal of the Learning Sciences, 1*, 201–238.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1992). The integration of knowledge and experimentation strategies in understanding a physical system. *Applied Cognitive Psychology, 6*, 321–343.

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28*, 859–882.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*, 162–176.

Shaklee, H., Holt, P., Elek, S., & Hall, L. (1988). Covariation judgment: Improving rule use among children, adolescents, and adults. *Child Development, 59*, 755–768.

Shaklee, H., & Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Development, 52*, 317–325.

Shaklee, H., & Paszek, D. (1985). Covariation judgment: Systematic rule use in middle childhood. *Child Development, 56*, 1229–1240.

Shultz, T. R., Fisher, G. W., Pratt, C. C., & Rulf, S. (1986). Selection of causal rules. *Child Development, 57*, 143–152.

Shultz, T. R., & Mendelson, R. (1975). The use of covariation as a principle of causal analysis. *Child Development, 46*, 394–399.

Siegler, R. S., & Alibali, M. W. (2005). *Children's Thinking* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist, 46*, 606–620.

Siegler, R. S., & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Developmental Psychology, 11*, 401–402.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 31–76). Hillsdale, NJ: Lawrence Erlbaum.

Simon, H. A. (1957). *Models of Man*. New York: Wiley.

Simon, H. A. (1986). Understanding the processes of science: The psychology of scientific discovery. In T. Gamelius (Ed.), *Progress in science and its social conditions* (pp. 159–170). Oxford: Pergamon Press.

Simon, H. A. (1989). The scientist as problem solver. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 375–398). Hillsdale, NJ: Lawrence Erlbaum.

Simon, H. A. (2001). Seek and ye shall find. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 5–20). Mahwah, NJ: Lawrence Erlbaum.

Smith, C. L., Maclin, D., Houghton, C., & Hennessey, M. G. (2000). Sixth-grade students' epistemologies of science: The impact of school 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.

Sophian, C., & Huber, A. (1984). Early developments in children's causal judgments. *Child Development, 55*, 512–526.

Sperber, D., Premack, D., & Premack, A. J. (Eds.). (1995). *Causal cognition: A multidisciplinary debate*. Oxford: Clarendon Press.

Stanovich, K. E. (1998). *How to the think straight about psychology* (5th ed.). New York: Longman.

Swaak, J., & de Jong, T. (2001). Discovery simulations and the assessement of intuitive knowledge. *Journal of Computer Assisted Learning, 17*, 284–294.

Thagard, P. (1998a). Ulcers and bacteria I: Discovery and acceptance. *Studies in History and Philosophy of Science. Part C: Studies in History and Philosophy of Biology and Biomedical Sciences, 29*, 107–136.

Thagard, P. (1998b). Explaining disease: Correlations, causes and mechanisms. *Minds and Machines, 8*, 61–78.

Toth, E. E., Klahr, D., & Chen, Z. (2000). Bridging research and practice: A cognitively based classroom intervention for teaching experimentation skills to elementary school children. *Cognition and Instruction, 18*, 423–459.

Trafton, J. G., & Trickett, S. B. (2001). Note-taking for self-explanation and problem solving. *Human–Computer Interaction, 16*, 1–38.

Triona, L. M., & Klahr, D. (2003). Point and click or grab and heft: Comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition & Instruction, 21*, 149–173.

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

Tweney, R. D. (2001). Scientific thinking: A cognitive-historical approach. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 141–173). Mahwah, NJ: Lawrence Erlbaum.

Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (Eds.). (1981). *On scientific thinking*. New York: Columbia University Press.

Tytler, R., & Peterson, T. S. (2004). From "try it and see" to strategic exploration: Characterizing young children's scientific reasoning. *Journal of Research in Science Teaching, 41*, 94–118.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories in core domains. *Annual Review of Psychology, 43*, 337–375.

Wellman, H. M., & Gelman, S. A. (1998). Knowledge acquisition in foundational domains. In *Handbook of child psychology* (pp. 523–573). New York: Wiley.

White, P. A. (1988). Causal processing: Origins and development. *Psychological Bulletin, 104*, 36–52.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*, 3–118.

Wilhelm, P., & Beishuizen, J. J. (2004). Asking questions during self-directed inductive learning: Effects on learning outcome and learning processes. *Interactive Learning Environments, 12*, 251–264.

Wilkening, F., & Sodian, B. (2005). Scientific reasoning in young children: Introduction. *Swiss Journal of Psychology, 64*, 137–139.

Wolpert, L. (1993). *The unnatural nature of science*. London: Faber and Faber.

Zachos, P., Hick, T. L., Doane, W. E. J., & Sargent, C. (2000). Setting theoretical and empirical foundations for assessing scientific inquiry and discovery in educational programs. *Journal of Research in Science Teaching, 37*, 938–962.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*, 99–149.

Zimmerman, C., Bisanz, G. L., & Bisanz, J. (1998). Everyday scientific literacy: Do students use information about the social context and methods of research to evaluate news briefs about science? *Alberta Journal of Educational Research, 44*, 188–207.

Zimmerman, C., and Glaser, R. (2001). *Testing positive versus negative claims: A preliminary investigation of the role of cover story in the assessment of experimental design skills* (Tech. Rep. No. 554). Los Angeles, CA: UCLA National Center for Research on Evaluation, Standards, and Student Testing (CRESST) .

Zimmerman, C., Raghavan, K., & Sartoris, M. L. (2003). The impact of the MARS curriculum on students' ability to coordinate theory and evidence. *International Journal of Science Education, 25*, 1247–1271.