The Development of Scientific Reasoning Skills: What Psychologists Contribute to an Understanding of Elementary Science Learning

Final Draft of a Report to the National Research Council
Committee on Science Learning Kindergarten through Eighth Grade
August 2005

Corinne Zimmerman
Illinois State University

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. Scientific reasoning (SR), broadly defined, includes the thinking skills involved in inquiry, experimentation, evidence evaluation, inference and argumentation 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. Major empirical findings are discussed using the SDDS model (Klahr, 2000) as an organizing framework. Recent trends in SR research 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, explorations of different types of instructional and practice opportunities that are required for the development, consolidation and subsequent transfer of such skills. Rather than focusing on what children can or cannot do, researchers have been in a phase of research characterized by an "under what conditions" approach, in which the boundary conditions of individuals' performance is explored. Such an approach will be fruitful for the dual purposes of understanding cognitive development and the subsequent application of findings to formal and informal educational settings.

Acknowledgements

I thank Carol Smith, David Klahr, and Deanna Kuhn for helpful comments on initial drafts of this paper, and Amy Masnick and Andrew Shouse for helpful conversations during the process of writing. Address correspondence to: Corinne Zimmerman, Department of Psychology, Illinois State University, Campus Box 4620, Normal, IL 61790. Email: czimmer@ilstu.edu

Outline

OVERVIEW

APPROACHES AND FRAMEWORK

- The Domain-Specific Approach: Knowledge about Scientific Concepts
- The Domain-General Approach: Knowledge and Skills of Scientific Investigation
- Integration of Concepts and Strategies: A Framework for the Scientific Discovery Process
- Summary

THE DEVELOPMENT OF SCIENTIFIC THINKING

- *Research Focusing on Experimental Design Skills*
- *Research on Evidence Evaluation Skills*
  - The Evaluation of Covariation Matrices and Data Tables
  - Coordinating Theory with Covariation Evidence
  - Reactions to the Kuhn et al. (1988) Studies
  - Covariation Does Not Imply Causation: What's Normative?
    - The role of causal mechanism.
    - Considerations of plausibility.
  - Operationally Defining Performance on Evidence Evaluation Tasks
  - Causal versus Scientific Reasoning
  - Evaluating Anomalous Data: Instructional Interventions and Cognitive Processes
  - How Evidence is Evaluated: Chinn and Brewer's Models-of-Data Theory
  - Summary: The Development of Evidence Evaluation Skills
- *Integrated Approaches to Scientific Reasoning: Partially Guided and Self-Directed Experimentation*
  - General Features of Integrated Approaches
  - Developmental Differences
    - Searching Hypothesis Space: Prior Knowledge and the Selection of Hypotheses
    - Predictions and Plausibility: Bridging the Search for Hypotheses and experiment
    - Searching Experiment Space: Strategies for Generating Evidence
    - Data Management
    - Evaluating Evidence: Interpretation and Inferences.
      - Continuous outcome measures and the understanding of measurement error.
    - Knowledge Change: Bridging Evidence Evaluation and Hypothesis Space
      - Evaluating "anomalous" evidence.
      - Evaluating evidence in social versus physical domains
    - Bootstrapping Experimentation Strategies and Conceptual Change
    - Individual Approaches to Self-Directed Experimentation
      - Theorists versus experimenters.
      - Perceived goal: scientists versus engineers.
    - Summary Developmental Differences and Individual Approaches
  - *Instructional and Practice Interventions*
    - Prompts
    - Instruction and Practice
    - An Example of Classroom-Based Design Experiment

SUMMARY AND CONCLUSIONS

- How Do Children Learn Science?
- How Can Cognitive Developmental Research Inform Science Teaching and Learning?
- Future Directions for Research

The Development of Scientific Reasoning Skills: What Psychologists Contribute to an Understanding of Elementary Science Learning

Children's thinking has been of interest to both psychologists and educators for over a century. The past several decades have witnessed an increase in research on children's scientific reasoning. Developmental psychologists have been interested in scientific thinking because it is a fruitful area for studying conceptual formation and change, 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 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 of the development of scientific reasoning skills (Zimmerman, 2000), I outlined the main findings of studies that were focused on experimentation or evidence evaluation skills as well as studies that examined the co-development of reasoning skills and conceptual knowledge. That review 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, there is evidence that cognitive and developmental researchers have become aware of the objectives of science educators and the updated education standards which recommend a focus on investigation and inquiry in the science classroom at all educational levels (e.g., American Association for the Advancement of Science, 1990; 1993; National Research Council, 1996; 2000) and have, moreover, used such knowledge in guiding their research questions and studies completed 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).

The goal of the present paper is to reiterate that research by cognitive developmental and educational psychologists can inform efforts to reform science education and, potentially, teacher preparation. My primary objective is to summarize research findings on the development of scientific thinking skills, with a particular focus on studies that target elementary and middle

school students. Additionally, I will highlight conceptual and methodological changes, draw attention to the trends that have emerged in the scientific reasoning research literature in the past 5-7 years, and integrate the findings of new and previously reviewed research.

Scientific reasoning, by definition, involves both conceptual understanding and inquiry skills. Sufficient research has been compiled to corroborate the claim that in the context of investigation, domain-specific knowledge and domain-general strategies "bootstrap" one another, such that there is an interdependent relationship between these two types of knowledge. However, there is evidence that as is the case of intellectual skills in general, the development of the component skills of scientific reasoning "cannot be counted on to routinely develop" (Kuhn & Franklin, 2006, p. 47ms). That is, young children have many requisite skills needed to engage in scientific thinking, but there are also ways in which even adults do not show full proficiency in investigative and inference tasks. 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 (Kuhn, 1991; 1993a; 1993b; 2002; Kuhn & Pearsall, 2000).

Research efforts, therefore, have been focused on how such skills can be promoted by determining which types of educational interventions (e.g., amount of structure, amount of support, emphasis on strategic or metastrategic skills) will contribute most learning, retention and transfer, and which types of interventions are best suited to different students. There is a developing picture of what children are capable of with minimal support, and research is moving in the direction of ascertaining what children are capable of, and when, under conditions of practice, instruction and scaffolding so that it may one day be possible to tailor educational opportunities that neither under- or overestimate their ability to extract meaningful experiences from inquiry-based science classes.

Literature from the psychology of science and the history and philosophy of science has taught us much about the thinking of professional scientists (e.g., Dunbar, 1995; 2001; Feist & Gorman, 1998; Gholson, Shadish, Neimeyer, & Houts, 1989; Giere, 1991; Klahr & Simon, 1999; Thagard, 1998c; Tweney, 2001). Recently, 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). There has been a call to incorporate more features of authentic science into educational contexts (see Chinn & Hmelo-Silver, 2002) so that students

may develop reasoning processes and epistemological understanding that is truer to real scientific inquiry, and which will promote the skills and dispositions to help students to become "little scientists" and scientifically literate adults (Metz, 2004; O'Neill & Polman, 2004).

OVERVIEW

The plan of the article is as follows. In the first section, I will briefly describe the two main approaches to the study of scientific thinking: one focused on the development of conceptual knowledge in particular scientific domains, and a second focused on the reasoning and problem-solving strategies involved in diverse activities such as hypothesis generation, experimental design, evidence evaluation and drawing inferences. Both approaches will be introduced to distinguish two different connotations of "scientific reasoning" that exist in the literature, but it is the second line of research that is the primary focus of this review. Klahr's (2000) *Scientific Discovery as Dual Search* (SDDS) model is a descriptive framework of the cognitive processes involved in scientific discovery and integrates elements of the concept-formation approach with the reasoning and problem-solving approach into a single coherent model. The SDDS model will be described because it will serve as the framework around which the main empirical findings will be organized.

The second section will include a review of the literature. This review 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 of scientific reasoning, 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. The current review will focus on the reasoning or problem-solving solving skills involved in students' scientific inquiry. Many of these studies include SR tasks that are either knowledge lean, or are situated within a particular scientific domain (e.g., genetics). Studies that focus specifically on conceptual development in various scientific domains (e.g., physics, biology) will not be discussed here. Reviews and collections of work on domain-specific concepts can be found in Carey (1985), 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). Additional approaches to the study of scientific reasoning also exist, such as the study of explanation (e.g., Keil & Wilson, 2000;

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 important for a full understanding of scientific reasoning.

In the final section of the paper, I will provide a general summary around the points outlined above. I will 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.

APPROACHES AND FRAMEWORK

The most general goal of scientific investigation is to extend our knowledge of the world. "Science" is a term that has been used to describe both a body of knowledge and the activities that gave rise to that knowledge. Similarly, psychologists who study the development of scientific knowledge in children have distinguished between the *product* (i.e., individuals' knowledge about scientific concepts), and the *processes* or activities that foster that knowledge acquisition. Science also involves both the *discovery* of regularities, laws, or generalizations (in the form of hypotheses or theories) and the *confirmation* of those hypotheses (also referred to as justification or verification). That is, there has been interest in both the inductive and deductive processes used in the generation and testing of hypotheses.

Scientific investigation broadly defined includes numerous procedural and conceptual activities such as asking questions, hypothesizing, designing experiments, making predictions, using apparatus, observing, measuring, being concerned with accuracy, precision and error, recording and interpreting data, consulting data records, evaluating evidence, verification, reacting to contradictions or anomalous data, presenting and assessing arguments, constructing explanations (to self and others), coordinating theory and evidence, performing statistical calculations, making inferences, and formulating and revising theories or models (e.g., Carey, Evans, Honda, Jay, & Unger,1989; Chi, de Leeuw, Chiu, & Lavancher, 1994; Chinn and Malhotra, 2001; Keys, 1994; McNay & Melville, 1993; Schauble, Glaser, Duschl, Schulze, & John, 1995; Slowiaczek, Klayman, Sherman, & Skov, 1992; Zachos et al., 2000). Because of this complexity, researchers traditionally have limited the scope of their investigations by concentrating on either the conceptual or the procedural aspects of scientific reasoning. That is, the focus has been on the acquisition and development of two main types of knowledge, namely, domain-specific knowledge and domain-general strategies.

The Domain-Specific Approach: Knowledge about Scientific Concepts

One approach to studying the development of scientific reasoning has involved investigating the *concepts* that children and adults hold about phenomena in various content domains of science. In this approach, the focus is on conceptual development or conceptual change. Researchers are interested in what individuals understand about phenomena in domains such as biology (e.g., Carey, 1985; Hatano & Inagaki, 1994; Miller & Bartsch, 1997), evolution (e.g., Samarapungavan & Wiers, 1997), observational astronomy (e.g., Vosniadou & Brewer, 1992; Vosniadou, Skopeliti, & Ikospentaki, 2004), and physics (e.g., Baillargeon, Kotovsky, & Needham, 1995; Clement, 1983; diSessa, 1993; Hood, 1998; Hunt & Minstrell, 1994; Kaiser et al., 1986; Levin, Siegler, & Druyan, 1990; McCloskey, 1983; Minstrell, 2001; Pauen, 1996; Spelke, Phillips, & Woodward, 1995). This approach has historic origins in the pioneering work of Piaget, who was interested in the development of concepts such as life, consciousness, day-night cycles, weather, time, number, space, movement, and velocity (e.g., Flavell, 1963: Inhelder & Piaget, 1958; Piaget, 1970; 1972).

In the domain-specific approach, the primary goal is to determine the naïve mental models or domain-specific theories that children and adults hold about scientific phenomena and the progression of changes that these models undergo with experience or instruction. These naïve theories may or may not match currently-accepted, scientific explanations of those same phenomena (Murphy & Medin, 1985; Vosnaidou & Brewer, 1992; Wong, 1996). Of interest is the content and structure of these naive theories, possible misconceptions, conceptual change, and explanatory coherence (Samarapungavan & Wiers, 1997; Schauble, 1996; Thagard, 1989).

Although the primary focus of such research is what is referred to as conceptual development or change (e.g., Carey, 1985; 2000) or "scientific understanding" (Kuhn, 2002), such studies have been labelled as scientific *reasoning* because the tasks used to assess such knowledge often require participants to reason about situations, answer questions, or solve problems based on their current understanding (i.e., responses may not simply be retrieved from memory). In generating solutions to problems (e.g., predict the trajectory of a falling objects; McCloskey, 1983) or answers to questions (e.g., as "Is there an edge to the earth?"; Vosniadou & Brewer, 1992, p. 553) individuals do not conduct experiments, make observations or evaluate evidence to verify their solutions or answers. Inquiry and investigation skills typically are not included in studies focused exclusively on individuals' understanding of particular scientific phenomena.

The Domain-General Approach: Knowledge and Skills of Scientific Investigation

A second approach to understanding scientific reasoning has focused on the development of domain-general reasoning and problem-solving skills applied to the context of investigation. Such studies also evolved from the Piagetian tradition (e.g., Inhelder & Piaget, 1958), in which the tasks involved the manipulation and isolation of variables. Tasks such as the balance scale (e.g., Siegler, 1976), the pendulum task, flexibility of rods, projection of shadows (Case, 1974) and inclined planes were not used in the service of understanding the underlying science concepts (cf. Piaget, 1972) but with children's abilities to systematically explore these problems, formulate hypotheses, manipulate variables, and observe the consequences. The goal was to understand cognitive changes in problem solving and reasoning in a domain-general way (e.g., the transition to "formal operations"). Because many of these tasks have rich conceptual content, the interpretation of performance was not straightforward. Researchers started using knowledge-lean tasks in an attempt to isolate general skills involved in reasoning in both children (e.g., Siegler & Liebert, 1975) and adults (e.g., Mahoney & DeMonbreun, 1977; Mynatt, Doherty, and Tweney, 1978; Wason, 1960, 1968; Wason & Johnson-Laird, 1972). The goal was to reduce or eliminate the role of domain-specific knowledge in order to focus on domain-general strategies. These early efforts represented an attempt to determine the extent to which each type of acquisition (i.e., concepts or strategies) accounts for developmental differences in scientific reasoning (Klahr, Fay, & Dunbar, 1993). These lines of research were eventually abandoned because science can be characterized as both product and process, and as such, progress would be limited by research that focused on reasoning in the absence of content.

More recent and current research focuses on domain-general reasoning and problem-solving strategies that are involved in the discovery and modification of theories about categorical or causal relationships. These strategies include the general skills implicated in the cycle of scientific inquiry, such as experimental design and evidence evaluation. Although applied to meaningful content, the focus is on the cognitive skills and strategies that transcend the particular content domain to which they are being applied. Scientific thinking as defined in this approach involves the application of the methods or principles of scientific inquiry to reasoning or problem-solving situations (Koslowski, 1996). In contrast to the conceptual-development approach, participants are engaged some or all of the components of scientific inquiry, such as designing experiments, evaluating the findings from real or fictitious

experiments and making inferences in the service of forming and/or revising theories[1] about the phenomenon under investigation.

Integration of Concepts and Strategies: A Framework for the Scientific Discovery Process

As it became clear that the study of scientific thinking would not progress without recognition of the importance of both product (concept formation or knowledge acquisition) and process (experimental design and evidence evaluation skills), Klahr and Dunbar (1988; Klahr, 2000) developed an integrated model of the cognitive processes involved in scientific activity. The *Scientific Discovery as Dual Search* (SDDS) framework incorporates domain-general strategies with domain-specific knowledge. The SDDS model was influenced by the work and assumptions of Simon and his colleagues (e.g., Newell & Simon, 1972; Simon and Lea, 1974; Langley, Simon, Bradshaw, & Zytkow, 1987). Simon (1973; 1986; 1989) argued that scientific discovery is a problem-solving activity that uses the same information-processing mechanisms identified in other problem-solving contexts.

One of the most important findings about human problem solving is that people use *heuristic search* (e.g., Hunt, 1994; Newell & Simon, 1972; VanLehn, 1989). Klahr and Dunbar (1988; Dunbar & Klahr, 1989) conceived of scientific reasoning as problem solving that is characterized as a guided search and information-gathering task. The primary goal is to discover a hypothesis or theory that can account for some pattern of observations in a concise or general form (Klahr, 1994; 2000). Klahr and Dunbar argued that scientific discovery is accomplished by a *dual-search* process. The search takes place in two related problem spaces—the hypothesis space and the experiment space. The search process is guided by prior knowledge and previous experimental results. With respect to searching hypothesis space, one can use prior knowledge in order to constrain the search, but in other situations, one must make some observations (via experimentation) before constructing an initial hypothesis. One implication of this distinction is that the search through experiment space may or may not be constrained by a hypothesis. Initial search through the space of experiments may be done in the service of generating information

[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).

about the particular phenomenon under investigation. In order to test a hypothesis, in contrast, the search process involves finding an experiment that can discriminate among rival hypotheses. The search through these two spaces requires different representations of the problem, and may require different heuristics for moving about the problem spaces.

The first two cognitive processes of scientific discovery involve a coordinated, heuristic search. The third process of the SDDS model involves evidence evaluation. This process was initially described as the decision made on the basis of the cumulative evidence, that is, the decision to accept, reject, or modify the current hypothesis. Kuhn (1989, 2002) has argued that the heart of scientific thinking lies in the skills at differentiating and coordinating theory (or hypotheses) and evidence. As such, Klahr has elaborated upon the evidence evaluation process, indicating that it involves a comparison of results obtained through experimentation with the predictions derived from the current hypothesis (e.g., Klahr & Carver, 1995). Although original descriptions of SDDS highlighted a "dual-search" coordination, updated descriptions acknowledge the coordination and integration of all three components (Klahr, 1994; 2000; 2005a).

The SDDS framework captures the complexity and the cyclical nature of the process of scientific discovery (see Klahr, 2000, for a detailed discussion). In addition, most of the studies of these processes have focused either on very specific types of knowledge and processes or very general knowledge and processes. Thus one can use the top level categories of the model to organize the extensive literature on scientific reasoning by crossing the three major components of scientific discovery (columns) with two broadly defined knowledge types (rows) and then situating various studies in or across the cells shown in Table 1. In this review I will use Table 1 as a framework for the review of the empirical investigations of scientific reasoning, even though most of that work was conducted independently of the SDDS framework. Each of the cells in Table 1 will be described in the following section.

Studies in Cell A are most closely aligned with research that has been labeled as *conceptual development*, *conceptual change* (e.g., Carey, 1985) or *scientific understanding* (e.g., Kuhn, 2002) and have to do with the theories that individuals hold about particular phenomena. Such studies represent the domain-specific approach to studying scientific reasoning (as described above). Studies in Cell D are not common, but can be illustrated by Bruner, Goodenough and Austin's (1956) reception experiments. The focus of this review will be on the

remaining cells. Research on experimentation skills (cells B and E) and evidence evaluation (cell F, or research crossing cells C and F) will be discussed in separate sections. The process by which theories are developed, questioned, tested and/or revised has been referred to by such labels as scientific reasoning, scientific thinking, or scientific problem solving – and includes the coordination of all of the elements in this table. These integrated studies will then be reviewed.

Table 1

Klahr's (2000) Categorization of Types of Foci in Psychological Studies of Scientific Reasoning Processes and Representative Publications

	Type of Cognitive Processes		
Type of Knowledge	Hypothesis Space Search	Experiment Space Search	Evidence Evaluation
Domain-specific	A (Carey, 1985)	B (Tschirgi, 1980)	C (Chi & Koeske, 1983)
Domain-general	D (Bruner et al., 1956, Reception experiments)	E (Siegler & Liebert, 1975)	F (Shaklee & Paszek, 1985)

Summary

Scientific discovery is a complex activity that requires the coordination of many high-level cognitive skills, including heuristic search through problem spaces, inductive reasoning, and deductive logic. The main goal of scientific investigation is the acquisition of knowledge in the form of hypotheses or theories that can serve as generalizations or explanations. Psychologists have investigated the development of scientific concepts and the development of strategies involved in the discovery and verification of hypotheses. Klahr and Dunbar (1988; Klahr, 2000) proposed a framework for thinking about scientific reasoning in an integrated manner. The SDDS framework is a descriptive account of the processes involved in concept formation and strategy development in the service of scientific discovery. In the next section I will review

major empirical findings, beginning with early efforts to study scientific reasoning, in which only particular aspects of scientific discovery were of interest (as represented by the particular cells in Table 1) and ending with a description of more recent investigations that have focused on the integration of the processes and knowledge types represented by the SDDS framework as a whole.

THE DEVELOPMENT OF SCIENTIFIC THINKING

Initial attempts to study the development of scientific thinking skills began with investigations that followed a "divide-and-conquer" approach by focusing on particular cognitive components as represented by the cells in Table 1 (Klahr, 2000). The important findings to come out of this component-based approach will be described first – in particular, studies involving an exclusive focus on *experimentation skills* (cell E) and *evidence evaluation skills* (cell F). Investigations that use *partially guided* or *self-directed experimentation* tasks will then be reviewed. This recent line of research involves simulated-discovery tasks that allow researchers to investigate the dynamic interaction between domain-general strategies (i.e., experimentation and evidence evaluation skills) and conceptual knowledge in moderately complex domains. These tasks incorporate the three major processes of scientific discovery in the context of domain-specific knowledge (cells A through F). As mentioned previously, research focused *exclusively* on domain-specific hypotheses (cell A), exemplified by work on the development of conceptual knowledge in various domains such as biology or physics (e.g., Carey, 1985; McCloskey, 1983), has been reviewed elsewhere (e.g., Wellman & Gelman, 1992)

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 or an alternative (Simon, 1989). Although it is has been argued that there is no one "scientific method" (e.g., Bauer, 1992; Shamos, 1995; Wolpert, 1993), it can be argued that there are several characteristics common to experimentation across content domains. At a minimum, one must recognize that the process of experimentation involves generating observations that will serve as evidence that will be related to hypotheses. Klahr and Dunbar (1988) discussed the "multiple roles of experimentation" with respect to generating evidence. Experimentation can serve to generate observations 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).

Ideally, experimentation should produce evidence or observations that are *interpretable* in order to make the process of evidence evaluation uncomplicated. One aspect of experimentation skill is to isolate variables in such a way as to rule out competing hypotheses. An alternative hypothesis can take the form of a specific competing hypothesis or the complement of the hypothesis under consideration. In either case, the control of variables and the systematic combinations of variables are particular skills that have been investigated. The control of variables is a basic, domain-general strategy that allows valid inferences and is an important strategic acquisition because it constrains the search of possible experiments (Klahr, 2000). It is an important skill to attain because in addition to being essential for *investigation*, 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.

Early approaches to examining experimentation skills involved minimizing the role of prior knowledge in order to focus on the strategies that participants used. That is, the goal was to examine the domain-general strategies that apply regardless of the content that they are applied to (i.e., cell E in Table 1). 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. The problem involved determining how to make an electric train run. The train was connected to a set of four switches and the children needed to determine the particular on/off configuration required. The train was in reality controlled by a secret switch so that the discovery of the correct solution was postponed until all 16 combinations were generated. In this task, there was no principled reason why any one of the combinations would be more or less likely. That is, the task involved no domain-specific knowledge that would constrain the hypotheses about which configuration was most likely. Additionally, the children were provided with one specific goal and so a search of hypothesis-space was further constrained.

Siegler and Liebert (1975) used two instructional conditions. 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. Students in the control condition were only exposed to the train problem and 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 about factors, levels and copying tree diagrams were more successful in the case of eighth graders (50% produced all combinations). This intervention was not successful for the fifth graders (0%). In contrast, 70% of the fifth-graders and 100% of the eight graders in the *conceptual framework plus analogs* group were able to produce all the combinations. With 20-25 minutes of instruction and practice, the majority of fifth graders and all eighth graders were able to engage in the manipulation of variables necessary for success on this task.

An equally important finding from the Siegler and Liebert study 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 were presented to fourth- and fifth-grade children. 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 of the fluids or combinations of fluids was needed to reproduce the effect. Like the Siegler and Liebert study, search of hypothesis-space was constrained in that the specific goal or hypothesis to explore was provided to students and domain knowledge of the fluids (e.g., color or smell) could not be used to identify a likely hypothesis. Therefore, success on the task was dependent on the ability to isolate and control variables in the set of all possible fluid combinations in order to determine which one of the colorless fluids was causally related to the outcome (i.e., the one fluid that causes a mixture to turn cloudy or red).

Over the course of several weeks, different fluids were used so the problem space changed at each session. If an individual student mastered the problem, then a more advanced problem would follow (e.g., more than one fluid was causal). Neither specific instruction nor feedback was provided – the only feedback students received was the effects provided by the outcomes of

their experiments (i.e., a mixture changing or not changing color). Although an interviewer asked questions so that the researchers could interpret what the students were doing during the course of experimentation, reinforcement was not provided and solutions or strategies were not suggested.

Students' experimentation strategies could be classified as either one of three types of genuine (or valid) experimentation (e.g., conducted for the purposes of, and was capable of, testing a hypothesis because a variable was controlled/isolated) or 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 with the causal fluid isolated from the others) or invalid (e.g., based on intuition, uncontrolled tests, insufficient evidence, etc.).

Students' experimentation and inference strategies over the course of weeks were coded. 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. However, an abrupt change from invalid to valid strategies was not common. Rather, the more typical pattern was one in which there existed the presence of valid and invalid strategies both within sessions and across sessions, with a pattern of gradual attainment of stable valid strategies by some students (with stabilization point varying, but typically around weeks 5-7). Students who were ultimately successful showed a relatively frequent use of genuine experimentation strategies (60-100%) prior to stabilization, whereas genuine experimentation was used only 9-45% of the time by students whose performance was not deemed successful. Experimentation coded as genuine included the characteristic of "planfulness." That is, the experiment was conducted with a purpose in mind, which includes the possibility of alternative outcomes (i.e., producing or not producing the effect). The use of planful experimentation was one of the few similarities among the successful students, leading Kuhn and Phelps to speculate that students who slowly but eventually discarded invalid strategies were ones who 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 that when performing a test to produce evidence, the value of the outcome might be one condition that determines whether people will

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, making a paper airplane) and therefore involved some domain knowledge (i.e., cells B and E of Table 1). Tschirgi expected that when determining the cause of a negative event (e.g., a bad cake) in a multivariable situation, one is more likely to isolate the one variable thought to be causally responsible (e.g., change honey to sugar), keeping the others constant. In contrast, to determine the cause of a positive event, one's goal may be to reproduce that effect and therefore conduct a test in which the variable believed to be causally responsible (e.g., honey) is held constant, with a change to the other variables.

Adults and children in grades 2, 4, and 6 were asked to determine which levels of a variable to change and which ones to keep constant in order 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 baked a cake using 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 and were given three options to choose from: (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 flour (called the VOTAT strategy, for "Vary One Thing At a Time" and which is the only strategy that results in an unconfounded experiment[2]); or (c) changing all the ingredients (i.e., butter, sugar, and white flour) (Change All). Participants were presented with eight different multivariable problems (four good and four bad outcome) and told to pick the one *best* answer from the three choices provided. That is, participants did not manipulate the variables to produce a conclusive test, nor did they generate the hypothesis to be tested.

Tschirgi (1980) found that in familiar, everyday problem situations, 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. That is, for positive outcomes, the HOTAT strategy for manipulating variables was selected (choice *a* above) more frequently (54%) than VOTAT (33%) or CA (13%). All participants

[2] The VOTAT strategy is more recently referred to as the "control of variables" strategy or CVS.

selected disconfirmatory evidence when there was a "negative" outcome, picking the VOTAT strategy (choice *b* above) more frequently (55%) than HOTAT (21%) or CA (24%). This pattern suggests that when there is a negative outcome, there is a tendency to search for the one variable to change to eliminate the bad result (consistent with the elements of a controlled experiment). When there is a positive outcome, in contrast, there is a tendency to hold the presumed causal variable constant in order to maintain the good result (consistent with a confounded 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 the offending variable). Tschirgi suggested that the results supported 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 (i.e., Cells B and E). The task followed a curriculum unit in which groups of students designed and conducted experiments with plants. Students were provided with a hypothesis to test, but were not required to conduct the experiment or to evaluate evidence. All students who were asked to design an experiment to test the claim that "tap water is bad for plants" (i.e., a claim about a negative outcome with a familiar variable) 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 who were asked to test the claim that "coffee grounds are good for plants" (i.e., a claim about a positive outcome with an unfamiliar variable) designed experiments as though the goal was to test the generality of the claim. That is, rather than testing the veracity of the claim, they designed experiments to figure out which plant types coffee grounds are good for. About a quarter of the students designed experiments with a single manipulated variable, with a similar number selecting the correct variable to test (i.e., presence/absence of coffee grounds). Even with classroom experience in experimental design, variations in the form of the hypothesis to be tested (positive/negative; familiar/unfamiliar) affected students' search of the space of possible experiments. Although students were provided with a hypothesis to test, the design of the experiment was an open-ended task. Either one of these cover stories 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 (2000), the experimental designs that were chosen or suggested by participants may be defined as more or less valid with respect to a normative model of experimentation. Students' experimental skills appear to be influenced by situational factors, in these cases, whether the outcome can be interpreted as positive or negative. Under the conditions of a positive outcome, individuals seem to act as though they are certain about the causal status of a variable (e.g., that honey produces a good cake, or that coffee grounds are good for plants) and the task before them is to test the conditions under which that positive outcome holds. An alternate interpretation may be that the rationale behind such a strategy is to demonstrate that the claim holds under a variety of conditions (e.g., to show that honey produces a good cake regardless of the flour type or the shortening type, or to show that coffee grounds are good for a wide variety of plants). Normatively, the missing step is the initial confirmation of the claim in a controlled way -- showing that under some condition, honey is better than sugar, or that coffee grounds are good for some type of plant under some constant conditions. Variations in mental models of experimentation and/or mental models of causality may underlie these performance variations, 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).

Sodian, Zaitchik, and Carey (1991) investigated whether children in the early school years understand the difference between testing a hypothesis and reproducing an effect. Many of the tasks used to investigate children's experimentation skills previously involved producing an effect (e.g., making a train run, Siegler & Liebert; 1978; baking a good cake, Tschirgi, 1980). Although the participants were instructed to test a hypothesis, it is not possible to address the issue of whether they made the distinction because the specific hypotheses provided to the students required them to think about producing an effect. Moreover, researchers did not 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 in their home. One brother believes the mouse is small, the other believes it is large. 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 if the brothers wanted 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 asked 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 the goals of testing a hypothesis with an experiment versus generating an effect (i.e., feeding the mouse), then he or she should select different houses 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, a task was used in which story characters were trying to determine whether a pet aardvark had a good or a poor sense of smell. In the aardvark task, participants 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 the story characters should place some food very far away. If the aardvark has a good sense of smell, then it will find the food. The results support the general idea that children as young as 6 can distinguish between a conclusive and inconclusive experimental test of a simple hypothesis. It is important to point out, however, that the children were provided with the two mutually exclusive and exhaustive hypotheses, and in the case of the mouse-house task, were provided with two mutually exclusive and exhaustive experiments to select from (Klahr et al., 1993).

Summary of Experimentation Studies

In summary, a number of studies have been conducted which focused primarily on skills implicated in experimentation, in tasks that are either knowledge-lean or for which domain knowledge can be considered (i.e., Cells B and E of Table 1). 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 (or control) of variables on multivariable knowledge-lean tasks have been shown to emerge under conditions of practice or instruction.

Without instruction, few fifth- or eighth-graders were able to produce the full set of

possible combinations (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 able to produce the full combinatorial array. 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. Under conditions of repeated practice over the course of 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 moreover, were likely 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, 2000). Children and adults selected valid experimental tests when the hypothesized outcome was negative, but used a less valid strategy 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 be selected with the pragmatic goals of repeating positive effects and avoiding negative effects (and may perhaps, be loosely related to Herb Simon's concept of *satisficing*). A second explanation for such findings may be rooted in students' developing epistemologies and metacognitive understanding of the purposes of experimentation. For example, Carey et al. (1989) interviewed seventh graders about their understanding of the nature of science. Based on a coding of pre-instruction interview protocols, most students' beliefs were consistent with the ideas that "a scientist 'tries it to see if it works'" (p. 520); 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 do not differentiate between producing a particular phenomenon and understanding a phenomenon (Carey et al., 1989). It is not until a more advanced level of understanding that students differentiate ideas and experiments and believe that the goal is to use an experiment to test an idea and to construct explanations.

The research described in this section was limited to studies in which there was a particular focus on the experiment space search (cells B and E of Table 1). The specific set of skills included the control of variables (also called isolating variables and/or VOTAT), producing the

full set of factorial combinations in a multivariable task, selecting an appropriate design or a conclusive test, generating experimental designs or conclusive tests, and record keeping. Although limited with respect to the range of skills involved in scientific thinking, these studies provide a picture of the developing experimentation skills in students from first through eighth grade and the conditions under which more and less sophisticated use emerges.

All of the findings and conclusions from studies that focus on experimentation skills anticipate those to be reviewed in subsequent sections. In particular, findings that will be replicated with tasks that incorporate other skills (i.e., hypothesis generation and evidence evaluation) 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 the strategies selected, 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 scientific reasoning 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 has been of central interest in the work of Kuhn and her colleagues (e.g., 1989; 2002; Kuhn et al., 1988; Kuhn, Schauble, & Garcia-Mila, 1992; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Kuhn & Dean, 2004). Kuhn has argued that the defining feature of scientific thinking is the set of skills involved in differentiating and coordinating theory and evidence. Fully developed skills include the ability to consciously articulate a theory, to understand the type of evidence that could support or contradict that theory, and to justify the selection of one of competing theories that explain the same phenomenon. The ability to consider alternative hypotheses is an important skill, as evidence may relate to competing hypotheses. Kuhn has asserted that the skills in coordination of theory and evidence are the "most central, essential, and general skills that define scientific thinking" (Kuhn, 1989, p. 674). That is, these skills can be applied across a range of content areas. 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 connotations will be

discussed in more detail after the review of evidence evaluation studies.

A number of studies have examined the development of evidence evaluation skills in using knowledge-lean tasks (cell F of Table 1). With the addition of more sophisticated domain knowledge, later studies could be situated across cells A, C, D and F of Table 1. 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 (see Table 2). In the case of perfect covariation, one would only find cases in which both the antecedent and the outcome were present together (A) and cases in which they were both absent (D). Such instances confirm a relationship between two events. However, in a noisy and imperfect world, cases exist in which there is a violation of the sufficiency of causal antecedent (assumed cause present/outcome absent; B). Cases may exist that violate the necessity of the causal antecedent (assumed cause absent/outcome present; C). Whether evidence is presented in data tables, pictorially, or results from active experimentation (to be discussed in the next section), any data presented as a categorical (i.e., rather than quantitative) outcome is consistent with Table 2.

The correct rule for determining covariation between events in a 2 x 2 matrix is the *conditional probability strategy*, in which one compares $P(A|[A+C])$ with $P(B|[B+D])$. Mathematically, this simply requires a comparison of the frequency ratio in cells $A \div (A+C)$ with $B \div (B+D)$ (Shaklee & Paszek, 1985). If the ratios are the same, then there is no relationship between the antecedent (presumed cause) and the outcome (i.e., in statistical terms, the variables are independent). If there is a difference between these ratios, then the events covary (i.e., a relationship may exist). Researchers have only recently begun to address the issue of *how large* this difference must be to conclude that a relationship exists, given that scientists would use statistical techniques to analyze such data (in this case, by computing a χ^2 statistic and associated

probability). Data in the form of frequencies of the co-occurrence of events is the most common type of evidence used in such tasks, although some researchers have begun to explore how children and adults evaluate quantitative evidence (e.g., such as would correspond to computing a parametric *t*-test).[3]

Table 2

Cells in a 2 x 2 Contingency Table for Studies Using Covariation Evidence

		Outcome	
		Present	Absent
Antecedent	Present	A	B
	Absent	C	D

Note. The antecedent is the presumed causal factor. The cells represent the frequency of co-occurrences. See text for further explication

In the case of evidence evaluation tasks involving covariation of events, participants are provided with data corresponding to the frequencies in the cells of a 2 x 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 inductive or deductive inferences that can be made on the basis of the *pattern of evidence*. That is, for most tasks, participants were instructed to disregard prior domain knowledge while evaluating the evidence. Whether data are categorical or quantitative, or in numerical or pictorial form, another common feature of such studies is that the evidence that is

[3] A growing area of educational and psychological research that intersects with the scientific reasoning 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 (Ed.) (1998).

evaluated is *externally supplied*. The data are to be taken on the authority of the researcher – participants do not generate or produce the data. Therefore, experimental design skills are not of interest (e.g., isolating focal variables and controlling potentially confounding variables).

The Evaluation of Covariation Matrices and Data Tables

Early work on covariation detection was conducted by Shaklee and her colleagues (e.g., Shaklee & Mims, 1981; Shaklee & Paszek, 1985; Shaklee, Holt Elek, & Hall, 1988). Children in grades 2 through 8 and adults were presented with a series of 2 x 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 of evidence, which hypothesis was supported (i.e., if the events are related, and direction of the relationship if any). Shaklee and her colleagues found that the most sophisticated strategy that participants seemed to use, even as adults, was to compare the *sums-of-diagonals*. The *conditional probability* rule was only used by a minority of participants, even at the college level. Adults could readily learn this rule, if they were shown how to compare the relevant ratios. Children in grades 4 through 8, who were initially assessed as using a *cell-A* or an *A-versus-B* strategy 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 with 60-81% of 4th to 8th graders still using *sums-of-diagonals*. 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 they did consider the information from all four cells but in a less than ideal manner.

Using data tables in which only two conditions were compared, but for which the data were quantitative in nature, 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 with covariation evidence presented to them. Kuhn and her colleagues used simple, everyday contexts, 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, in Studies 1a and 1b, 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 ratings from the initial theory 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 the participant believed do not make a difference (e.g., type of potato, and type of condiment). This procedure allowed the evidence to be manipulated such that covariation evidence could be presented which *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. The specific manipulations, therefore, were tailored for each person in the study.

Kuhn et al.'s (1988) general method involved the presentation of covariation data sequentially and cumulatively. Participants were asked a series of questions about what the evidence shows for each of the four variables. Responses were coded as either *evidence-based* or *theory-based*. To be coded as evidence-based, a participant's response to the probe questions had to make reference to the patterns of covariation or instances of data presented (i.e., the findings of the scientists). 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 the participant's prior beliefs or theories about why the

scientists might have found that particular relationship. In the previous 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.

Kuhn et al. were also interested in both inclusion inferences (an inference that two variables are causally related) and exclusion inferences (an inference of no relationship between variables). Participants' inferences and justification types could be examined for covariation evidence versus noncovariation evidence and in situations where the prior theory was causal or noncausal. Other variations in the other studies included: (a) examining the effects of explicit instruction; (b) use of real objects for evidence (e.g., tennis balls with various features) versus pictorial representations of data; (c) task instructions to relate the evidence to multiple theories instead of a single theory; and (d) a reciprocal version of the task in which the participant generates the pattern of evidence that would support and refute a theory.

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, show 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 have a tendency to meld theory and evidence into a single representation of "the way things are." Second, participants have a variety of strategies for keeping theory and evidence in alignment with one another when they are in fact discrepant. One tendency is to ignore, distort, or selectively attend to evidence that is inconsistent with a favoured theory. 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 second 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 most often outside an individual's conscious awareness and control. Participants were often unaware of the fact that they were modifying their theory. When asked to recall their original beliefs, participants would often 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 each presentation of an instance of 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., 1988, p. 83). Kuhn argued that this tendency suggests that the subject'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 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.

Third, there were a variety of errors involved in understanding covariation evidence and its connection to causality. There were also problems in understanding noncovariation and its connection to an absence of causality. For example, when asked to generate a pattern of evidence that would show that a factor makes no difference in an outcome, participants often produced covariation evidence in the opposite direction of that predicted by their own causal theory.

Reactions to the Kuhn et al. (1988) Studies

Koslowski (1996) considers the work of Kuhn and her colleagues to be a significant contribution to research on scientific thinking, in part, because it raises as many questions as it answers. Various authors have criticized the Kuhn et al. (1988) research, however, on both methodological and conceptual grounds (e.g., Amsel & Brock, 1996; Koslowski, Okagaki, Lorenz, & Umbach, 1989; Ruffman, Perner, Olson, & Doherty, 1993; Sodian, Zaitchik, & Carey, 1991).

Methodological considerations. Sodian et al. (1991) first questioned Kuhn et al.'s interpretation that third- and sixth-grade children cannot distinguish between their beliefs (i.e., theories) and the evidence that would confirm or disconfirm those beliefs. Sodian et al. deliberately chose story problems in which children did not hold strong prior beliefs and they used a task that was less complex than those used by Kuhn et al. (1988). In order to demonstrate that children can differentiate beliefs and evidence, they selected a task that did not require a judgment about one causal factor while simultaneously ruling out the causal status of three other potential variables. As described in the previous section, Sodian et al.'s research 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. They also used less complex tasks with fewer factors to consider than Kuhn et al. (1988). 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. Ruffman et al. also ascertained that 6-year-olds were able to form a causal hypothesis on the basis of a pattern of covariation evidence (i.e., imperfect evidence).

In order 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 the hypothesis formed on the basis of covariation evidence. Children were asked to evaluate evidence and then form a hypothesis about which characteristics of tennis rackets were responsible for better serves (e.g., racket-size, head-shape). They were then asked which tennis racket they would buy and how good the next serve would be. The results were consistent with the idea that by age 7, children understood that the newly-formed hypothesis could be used to make predictions.

Ruffman et al. deliberately chose factors that were all equally plausible. Correct performance in the Kuhn et al. tasks was defined by considering covariation evidence as more important than the implausible hypothesis it was intended to support. For example, in Studies 3 and 4 of Kuhn et al., adults and third-, sixth-, and ninth-graders were to evaluate evidence to determine the features of tennis balls that resulted in good or poor serves (i.e., color, texture, ridges, and size). Most children and adults do not believe that color is causally related to the quality of a 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).

Amsel and Brock (1996) examined whether children and adults evaluated covariation evidence independently of prior beliefs or not. They also used a task that was less complex and cognitively demanding than Kuhn et al. (1988). Amsel and Brock argued that causal judgments should be assessed independently of the justification for that judgment and that these judgments about the causal status of variables should be assessed on a scale that reflects certainty, rather than a forced choice (i.e., the factor is causal, noncausal, or neither).

Unlike Ruffman et al.'s (1993) criticism about strong prior beliefs, the participants in Amsel and Brock's study were selected only if they did hold strong prior beliefs concerning the

variables. 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 were presented with four data sets to evaluate given by the factorial combination of prior belief (causal or non-causal) and type of contingency data (perfect positive correlation vs. zero correlation). Participants were asked whether giving the plants (sun/no sun) or (charm/no charm) was causally related to whether the plants were healthy or sick and to respond only based on the information given and not what they know about plants.

Standard covariation evidence served as the control (four instances in a 2 x 2 contingency table), while three conditions involved "missing data." Participants were told that the researcher forgot to record either the condition (e.g., if the plant got sun or no sun) and/or the outcome (i.e., if the plant was healthy or sick) for some of the experimental trials. Participants in the control group were presented with four data instances that represented covariation (or noncovariation) between the putative causal factor and outcome. Participants in the three missing data conditions were shown two additional instances in which (a) the antecedent was unknown, (b) the outcome was unknown or (c) both were unknown. Amsel and Brock reasoned that if participants were evaluating the evidence independently of their strongly-held prior beliefs, then the judgments in the control and missing data conditions should be the same. That is, participants would simply ignore the evidentially irrelevant missing data. If they were using prior beliefs, however, they might try to explain the missing data by judging the variables as consistent with their prior beliefs. If they were using newly-formed beliefs, then judgments would be consistent with the new belief and pattern of evidence (causal with covariation evidence; noncausal with noncovariation).

College adults were most like the "ideal reasoner" (i.e., defined as someone whose causal certainty scores were based solely on the four instances of contingency data). The pattern of mean causal certainty scores for both groups of children (2nd/3rd and 6th/7th grade) was such that they were making judgments consistent with prior beliefs, even when the evidence did not support those beliefs. For example, when presented with data showing covariation between the presence of a charm and 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 evidence was

disconfirming. The noncollege adults' judgments tended to be in between, leading the authors to suggest that there are differences associated with age and education in making causal judgments independently of prior beliefs. In the missing data conditions, participants did not try to "interpret" the missing data. Rather, the effect was to cause the children, but not the adults, to be less certain about the causal status of the variables. 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 their prior beliefs. When confronted with confirming evidence, however, both groups of children and adults made similar judgments.

The set of studies reviewed in this subsection addressed 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 what distinguishes more and less proficient scientific thinking is the ability to differentiate hypotheses from evidence as distinct epistemological categories, and that most children's and some adults' evaluation of evidence is done in a way that suggests they meld the two into one representation. The goal of the researchers was to show that with different methodological variants, children make a distinction between theory and evidence. When task demands are simplified such that a hypothesis can be induced from a pattern of evidence (e.g., Ruffman et al., 1993), children can detect those patterns and use the resultant hypothesis to make predictions. 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 really speak to the issue of the epistemological status of theory and evidence. Conceptual issues will be addressed next.

Covariation Does Not Imply Causation: What's Normative?

Koslowski (1996) criticized early research on the development of evidence evaluation skills based 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, chapter 5). 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 and considerations of plausibility. Koslowski argued that this reliance on tasks using covariation evidence has contributed to an incomplete or distorted picture of the reasoning abilities of children and adults. In some cases, tasks were so knowledge-lean that participants did not have the opportunity to use prior knowledge or explanation, thus contributing to an incomplete picture. When knowledge-rich tasks have been used, the operational definition of correct performance required participants to disregard prior knowledge. In this case, a distorted picture has resulted. As in Klahr and Dunbar's (1988) integrated model, Koslowski considers it legitimate to consider prior knowledge when gathering and evaluating evidence. Koslowski presented a series of 16 experiments to support her thesis that the principles of scientific inquiry are (and must be) used in conjunction with knowledge about the world (e.g., knowledge of plausibility, causal mechanism, and alternative causes).

The role of causal mechanism. Koslowski questioned the assumptions about the primacy of covariation evidence. One of the main concerns in scientific research is with the discovery of causes (Koslowski & Masnick, 2002). Likewise, previous researchers have used tasks requiring participants to reason about causal relationships. Psychologists who study scientific reasoning have been influenced by the philosophy of science, most notably the empiricist tradition which emphasizes the importance of observable events. Hume's strategy of identifying causes by determining the events that covary with an outcome has been very influential. In real scientific practice though, scientists are also concerned with *causal mechanism*, or the process by which a cause can bring about an effect. Koslowski noted that we live in a world full of correlations. 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 *e.coli* bacterium that we consider a causal relationship between hamburger consumption and illness or mortality. 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.[4]

[4] 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.

In the studies by Kuhn et al. (1988) and others (e.g., Amsel & Brock, 1996), correct performance entailed inferring causation from covariation evidence and lack of a causal relationship from noncovariation evidence. Evidence-based justifications are considered superior to theory-based justifications. In study 4, for example, a ninth grader was asked to generate evidence to show that the color of a tennis ball makes a difference in quality of serve, and responded by placing 8 light-colored tennis balls in the "bad serve" basket and 8 dark-colored balls in the "good serve" basket. When asked how this pattern of evidence proves that color makes a difference, the child responds in a way that is 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 by not trying to make sense of the evidence, participants would have to leave theory and evidence misaligned and therefore need to recognize them as distinct. Koslowski (1996), in contrast, would suggest this tendency demonstrates that participants' naive 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) do take mechanism into consideration when evaluating evidence in relation to a hypothesis about a causal relationship. In initial studies, Koslowski demonstrated that even children in sixth grade consider more than covariation when making causal judgments (Koslowski & Okagaki, 1986; Koslowski et al., 1989).

In subsequent studies, participants were given problem situations in which a story character is trying to determine if some target factor (e.g., a gasoline additive) is causally related to an effect (e.g., improved gas mileage). They were then shown either perfect covariation between the target factor and effect or partial covariation (4 of 6 instances). 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, the additive does not burn 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 also tried to determine if participants would spontaneously generate information about causal mechanisms when it was not cued by the task. Children and adults were presented with story problems in which a character is trying to answer a question about, for example, whether parents staying in hospital improves the recovery rate of their children. Participants were asked to describe whatever type of information might be useful for solving the problem. Half of the participants were told experimental intervention was not possible, while the other half were not restricted in this manner. Almost all participants showed some concern for causal mechanism, including expectations about how the target mechanism would operate. Although the sixth graders were less likely to generate a variety of alternative hypotheses, all age groups proposed appropriate contrastive tests.

In summary, Koslowski argues that sound scientific reasoning requires "bootstrapping," that is, using covariation information and mechanism information *interdependently*. Scientists, she argues, rely on theory or mechanism to decide which of the many covariations in the world are likely to be causal (or merit further study). To demonstrate that people are reasoning in a scientifically legitimate way, one needs to establish that they rely on both covariation and mechanism information and they do so in way that is judicious. As shown in the previous studies, participants did treat a covarying factor as causal when there was a possible mechanism that could account for how the factor might have brought about the effect, and were less likely to do so when mechanism information was absent. Moreover, participants at all age levels showed a concern for causal mechanism even when it was not cued by the task.

Considerations of plausibility. In another study, participants were asked to rate the likelihood of a possible mechanism to explain covariations that were either plausible or implausible. Participants were also asked to generate their own mechanisms to explain plausible and implausible covariations. When either generating or assessing mechanisms for plausible covariations, all age groups (sixth- and ninth-graders, adults) were comparable. When the covariation was implausible, sixth graders were more likely to generate dubious mechanisms to account for the correlation.

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. Koslowski provided an example from medical diagnosis, in which discovering the cause of Kawasaki's syndrome depended upon taking seriously the implausible correlation between the illness and having recently cleaned carpets. Similarly, Thagard (1998a; 1998b) 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 presented 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, 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 & Saramapungavan, 1991; Murphy & Medin, 1985) and causal reasoning literatures (e.g., Cummins, 1995; Shultz, Fisher, Pratt, & Rulf, 1986; Schulz & Gopnick, 2004; White, 1988) can be used to support the idea that both children and adults hold rich causal theories about "everyday" and scientific phenomena that include information about covariation and theoretically-relevant causal mechanisms (and possible alternative causes for the same effect). 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 be would overwhelming.

Operationally Defining Performance on Evidence Evaluation Tasks

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. This is where it is important to appeal 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, i.e., about the failure to recognize theory and evidence as distinct epistemological categories" (Kuhn & Franklin, 2006, p. 66ms).

Even though much of the research on evidence evaluation 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; 1998c; 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 the 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. An overarching theoretical goal for causal reasoning researchers has been to identify the universal inference rules that are used to make judgments of causality from covariation evidence (e.g., by appealing to a causal mechanism). Although researchers are interested in identifying the rules underlying all inductive causal inference (e.g., Cheng, 1997), most of the research has been conducted with college student populations. The few developmental studies conducted have been used to conclude that key features of causal inference may remain stable from childhood through adulthood (e.g., Harris, German, & Mills, 1996; cited in Kuhn & Dean, 2004). That is, the inference rules we use to make causal judgments

on the basis of evidence emerge in childhood and remain established well into adulthood, with key developmental differences being adults' superior in their abilities to differentiate between causes and enabling conditions, to consider a greater amount of information when making judgments, and be more accurate with respect to estimates of probability.

In contrast, the research literature 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. Evidence evaluation tasks are relatively open-ended (more current scientific reasoning tasks, to be reviewed next, are also participant-controlled with respect to strategies for investigating and generating evidence and the inferences made over cycles of inquiry).

Clearly, both groups of researchers are interested in the cognitive processes that allow one to make judgments of causality on the basis of evidence. Kuhn and Dean (2004) summarize the key differences in methodologies used in these two lines of research. In causal inference research, single session paper-and-pencil tasks with investigator-selected evidence to evaluate and for which judgments take the form of probabilities. Participants are most often college students. In the 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 some 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.

As discussed previously, theoretical accounts of causal reasoning (e.g., Cheng, 1997; Lien & Cheng, 2000) suggest that individuals possess a universal set of causal inference rules. Kuhn and Dean showed, however, 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. Half of the adults and three quarters of the children showed inconsistency across explicit judgments and implicit judgments (i.e., predictions about unique instances). Although adults almost always justified an inference 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 are not consistent with theories in the causal reasoning literature, suggesting that the causal theories that individuals hold do not translate into universal inference rules. Moreover, if such theories are primary or central when 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. The authors 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 that some evidence is inconsistent with those beliefs, that such evidence calls into question those beliefs, and in some cases, a belief will be altered or changed in response to the new (anomalous) evidence (Chinn & Brewer, 1998).

Chinn and Brewer propose that there are 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 about the relation of the data to the initial theory); (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 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.

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 (Cells A, C, F in Table 1). Children did not select the hypotheses or design the experiments. The goal was 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 experiment occurs quickly. An example of a topic that is counterintuitive but results in 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 the anomalous evidence produced by such experiments, 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 data that is ambiguous. At the level of interpretation, the resulting conclusion will be based on what was (or was not) observed (e.g., a child may or may not perceive the two objects landing simultaneously). Individuals may or may not align the observed evidence with theory, and may fail to do so in ways that vary in rationality (e.g., ignoring or distorting data may be less rational than discounting data because the data are considered flawed). 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 fourth-, fifth-, and sixth-grade students' evaluation of anomalous data. In the third study, one group of students was instructed to *predict* the outcomes of three experiments that produce counterintuitive but unambiguous data (e.g., reaction temperature). A second group answered questions that were 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. They were then tested for generalizations and a retention test followed 9-10 days later. There were main effects of age, with fifth- and sixth-graders' performance superior to fourth-graders (effect sizes 0.59 - 0.76), and a main effect of instructional condition, but no interaction. The explanation condition resulted in the best generalization and retention scores relative to the data-reflection and prediction conditions (effect sizes 1.39 – 1.60). Based on further analyses, 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 of results 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 that students observed. Studies of student-initiated experimentation will be described in the next section, but it is an interesting question if individuals evaluate evidence that is directly observable differently than evidence presented in the form of reports of evidence (e.g., Koslowski, 1996; Kuhn et al., 1988). Kuhn and Ho (1980) were interested in whether it was necessary for children to design their own experiments (using the colorless fluids task), or whether it was possible to make inferences from second-hand data already collected. Subjects were paired such that one child generated the evidence through experimentation, but the yoked-control child only made inferences from the evidence. Control children did make progress, but not to the same extent and speed as children who conducted the experiments. They 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 the two 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) has stressed the importance of the *interdependence* of theory and evidence, and that skilled individuals consider patterns of covariation evidence in conjunction with information about potential causal mechanisms, alternate causes, and the issue of plausibility. Similarly, Chinn and Brewer (2001) have proposed the *models-of-data* theory, in which they suggest that people 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 the evidence, procedural details, and the theoretical explanation of the observed findings (which may include unobservable mechanism such as molecules, electrons, enzymes or intentions and desires). The 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. This may 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 have used tasks or domains in which participants' beliefs are particularly "entrenched" or personally involving. For example, when faced with evidence that the type of condiment (ketchup vs. mustard) or the type of breakfast role 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). Thus, the *strength* of prior beliefs, and the *personal relevance* of those beliefs may influence the evaluation of the cognitive model. When individuals have reason to disbelieve evidence (e.g., because it is inconsistent with prior belief), they will search harder for flaws (Kunda, 1990). As such, individuals may not find the evidence 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. Additionally, in some cases, background knowledge can be used to identify methodological flaws (Chinn & Malhotra, 2002a). 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. Early research focused on the ways in which children and adults evaluated patterns of data in covariation matrices (cell F of Table 1). Later research focused on the conditions under which children and adults coordinate theory and evidence to make inferences (cells A, C, D and F of Table 1). 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 a more difficult time 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 is a complex, because as just one of the components of scientific thinking it also encompasses a large number of component skills. 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." Kuhn (e.g., 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 the 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 o f knowledge change in

response to evidence becomes increasingly within the student's control. A full account of developmental differences in scientific reasoning 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 reasoning 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 all of the cognitive skills implicated across the cells of Table 1 and require 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.

Integrated Approaches to Scientific Reasoning: Partially Guided and Self-Directed Experimentation

The "divide-and-conquer" strategy of focusing on particular components of scientific thinking 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, researchers are interested in examining 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 actively experiment with materials to determine and confirm the causal relations in multivariable systems. The student initiates all

phases of scientific discovery, 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 individual approaches, and then review research on the effects of instructional and practice differences. Many self-directed experimentation studies have been conducted only with undergraduates (e.g., Azmitia & Crowley, 2001; Dunbar, 1993; Okada & Simon, 1997; Schauble, Glaser, Raghavan, & Reiner, 1991; 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).

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).[5] In both, 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., Chen & Klahr, 1999; Masnick & Klahr, 2003) or the canal task (e.g., Gleason & Schauble, 2000; Kuhn et al., 1992). 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, such as the Daytona microworld (to discover the factors affecting the speed of race cars; Schauble, 1990). A variety of virtual environments have been created, in domains such as electric circuits (Schauble et al., 1992), genetics (Echevarria, 2003), earthquake risk and flooding risk (e.g., Keselman, 2003).

[5] 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-from 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.

Virtual environments to explore social science problems have also been created. For example, participants explore the factors that affect TV enjoyment (e.g., Kuhn et al., 1995) or CD catalog sales (e.g., Kuhn & Dean, 2005a; 2005b). The systems vary in conceptual complexity from fairly simple to moderately complex domains such as hydrostatics and genetics.

The virtual systems allow the experimenter to be in control of the "state of nature" (Klahr, 1994), but in a way that is not arbitrary or artificial like the task of determining the "physics" of a simulated universe (i.e., describing the motions of particles based on shape and brightness) used by Mynatt, Doherty, and Tweney (1978).[6] For any given system, some variables are consistent with participants' prior beliefs and some are inconsistent. For example, people hold strong beliefs that weight is a factor in how fast objects sink (Penner & Klahr, 1996a; Schauble, 1996) and that the color of a race car does not affect speed (Kuhn et al., 1992). The starting point is the participant's own theory, which presumably is the result of naïve conceptions and formal education. By starting with the participants' own theories, the course of theory revision can be tracked as participants evaluate self-generated experimental evidence that either confirms or disconfirms prior beliefs.

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. Table 3 includes common performance measures that may result for the different phases of experimentation. With respect to hypothesis search, participants' initial theories may be assessed, and if and when any changes occur during the course of experimentation and evidence evaluation (e.g., Schauble, 1996). Elements of initial and changing beliefs may also be noted (e.g., if only plausible hypotheses are mentioned, if causal mechanisms are posited for observed effects, or if there is a focus on a single hypothesis vs. competing hypotheses). A related measure involves the assessment of *comprehension* or knowledge gains. For example, seventh-graders knowledge of genetics was measured before and after three weeks of experimentation with genetics simulation software (Echevarria, 2003). An alternate measure of knowledge acquisition may be successfully determining the causal/non-causal status of all variables in the multivariable system (e.g., Penner & Klahr, 1996; Reid, Zhang, & Chen, 2003; Schauble, 1990).

---

[6] Most virtual systems are based in real science, but some may be simplified by making continuous variables categorical (e.g., temperature as "hot" or "cold" rather than a quantitative measure), or loosely based on real science (e.g., water temperature, water pollution, soil type, soil depth and elevation used to predict earthquake risk) (e.g., Kuhn & Dean, 2005) which are factors that may or may not map onto the real science of earthquake prediction.

Table 3

Common Measures used in Self-Directed and Partially Guided Experimentation Studies

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

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. The design that is selected 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 number of possible experiments is often larger than participants realize, therefore the use of data management (recording, consulting) is frequently noted (e.g., the percentage of experiments and outcomes recorded). Other common measures include participants' *intended plans* as well as their *predictions* before individual experiments are carried out.

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 or invalid (i.e., based on sufficient evidence and a controlled design), 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, Black, Keselman & Kaplan, 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.

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, allowing researchers to track progress on multiple performance indicators (e.g., change in knowledge and strategies). In other cases, participants work on the problem-solving task(s) for a single experimental session. Such studies may include a delayed post-test to assess retention on the same task or transfer to an isomorphic or related task. Based on the measures that can be collected in a SDE study (see Table 3), even a single session SDE can provide data on how participants negotiate all phases of hypotheses generation, experimentation, evidence evaluation and knowledge change. The goal is to track the co-development of strategies and knowledge (e.g., Schauble, 1996), and in some cases the co-development with metacognitive or metastrategic understanding (e.g., Kuhn & Peasall, 1998)

Now that I have described the features that are common or typical across studies, I will break the results into conceptual sections rather than talking about each study separately. I will begin with developmental differences, and address each of the performance indicators that map onto the cognitive components of the SDDS model by identifying. 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. Lastly, research addressing the effects of different instructional and practice interventions will be discussed. Although I will highlight patterns of findings across studies, Table 4 includes a summary of characteristics and main findings for each individual study.

Table 4
Self-Directed Experimentation Studies Focusing on K-8: Characteristics and Main Findings

Study	Participants	Task Domain(s)	Unique Feature	Hypotheses/Goals and Main Findings
Klahr & Dunbar (1988) Dunbar & Klahr (1989)	Expt 1 and 2: Undergraduates with (n = 20) & w/o programming experience (n = 10) Expt 3: Grades 3 – 6 (n = 22)	BigTrak Robot (discovery of a new function)	Development of SDDS model Focus on SDDS and developmental differences	*Goal*: replace the dichotomy of concept formation vs. problems solving into an integrated model Two main strategies of discovery: search hypoth. space (theorists) and search expt. space (experimenters); Scientific reasoning characterized as problem solving involving integrated search in 2 problem spaces Expt3: Given same data, children proposed diff. hypotheses than adults; Ss had difficulty abandoning current hypothesis (did not search H-space, or use expt. results); Ss did not check if hypoth. consistent w/ prior data; Expts designed to prove current hypoth. rather than discover correct one
Schauble (1990) Belief revision in children: The role of prior knowledge and strategies for generating evidence	Grade 5/6 (*n* = 22)	Daytona Microworld (cars)	Microgenetic study of evolving beliefs and reasoning strategies How reasoning is guided and constrained by knowledge	*Goal*: to describe changes in children's domain knowledge and reasoning strategies over an extended period and ID interactions b/w K and R Initial performance consistent with the goal or producing desirable outcomes (i.e., fastest car); Ss designed confounded expts and made invalid inferences, esp. incorrect causal inferences; invalid heuristics preserved favored theories Exploratory strategies improved with time; use of CVS and valid inferences; increase in inferences of non-causality and indeterminacy; Children using valid strategies gained better understanding of microworld structure
Schauble, Klopfer, & Raghavan (1991) Students' transition from an engineering model to a science model of experimentation	Grade 5/6 (*n* = 16) Task – WSs Context - BSs	Canal task (hydrodynamics); four variables with either 2 or 3 levels Spring task (hydrostatics); three variable with either 3 or 4 levels	Engineering context (goal = optimization of boat speed or spring length) Scientist context (goal = understanding the effects of variables on the boat speed or length of spring)	*Goal*: to investigate children's models of experimentation. *RQ*: Are children's goal orientations associated with different inquiry strategies? *Hyp*: Science context will be assoc,d w/broader search, incl. vars believed causal and non-causal; terminate when all vars/combos investigated; Eng context will be assoc.w/ a termination of search when some criterion is met; highly contrastive comparisons; focus on causal inferences only Ss belief about goal important: Science context resulted in broader exploration, more attention to all variables (incl. non-causal vars) & all possible combinations than in Eng. context; Greatest improvement when exploring Eng. problem first followed by Science problem

Table 4 (Continued) Study	Participants	Task Domain(s)	Unique Feature	Hypotheses/Goals and Main Findings
Kuhn, Schauble, & Garcia-Mila (1992) Cross-domain development of scientific reasoning	Expt1: Grade 4 (*n* = 12); boats & cars Expt2: Grade 5/6 (n = 20); cars & balls	Daytona Microworld (cars), Canal task (boats), Sports company research (balls)	Transfer paradigm (across domains)	*Goals*: (a) To ID if the same reasoning strategies are used across content domains; (b) to trace change in strategies over time; (c) to examine co-development of strategies across domains Developmental change in microgenetic context not specific to a single content domain; Co-existence of more & less advanced strategies; Domain knowledge alone does not account for development of scientific reasoning, co-development with reasoning strategies
Klahr, Fay, & Dunbar (1993)	Undergraduates, Commumity College students, Grades 3 & 6 (*n* = 64)	BigTrak Microworld (discovery of a new function)	Initial hypothesis provided (Incorrect & plausible or incorrect & implausible)	*RQ*: Do dev'l diffs in performance on SR tasks result from DG or DS acquisitions? *Hypothesis*: All Ss would use DS knowledge to assess relative plausibility of diff hypotheses, but DG heuristics (e.g., search of E-space, interpreting evidence) would vary with age/training Developmental differences in domain-general heuristics for search & coordination of searches in expt. and hypoth. spaces; Adults more likely to consider multiple hypotheses; Children focus on plausible hypotheses; Plausibility affects dual search – diff expt strategies used by children and adults as a function of whether hypothesis to test is plausible/implausible Principle heuristics ID'd: use plausibility of hypoth to select expt strategy; focus on one feature at a time; use knowledge of limits in own memory; design expts that give informative, interpretable results
Kuhn, Garcia-Mila, Zohar, & Andersen (1995)	Grade 4 (*n* = 15), Community College students (*n* = 17)	Daytona (cars), Canal task (boats), TV enjoyment task, School achievement task	Transfer paradigm (across Physical & Social domains)	*RQ*: In a microgenetic context, does substitution of new content affect the strategies that a Ss uses? Is variable strategy usage char. of dev'l transitions or a general char. of performance? Do Ss at diff dev'l levels show more rapid change in strategy usage given same starting point and amount of practice? Goals: How Ss generate evidence @ multi-var causal systems, form hypoth's @ relevant vars on the basis of evidence; DS knowledge & DG strategies; ID the mechanisms that affect theory change Strategic progress maintained by both groups when problem content changed midway; Social domain lagged behind Physical domain; Coexistence of valid & invalid strategies for children and adults (i.e., strategy variability not unique to periods of developmental transition)

Table 4 (Continued) Study	Participants	Task Domain(s)	Unique Feature	Hypotheses/Goals and Main Findings
Penner & Klahr (1996)	10-, 12-, & 14-year-olds (*n* = 30)	Sinking Objects task The effect of weight, size, shape (sphere, cube) material (steel, Teflon) on sinking time (min. 4 expts required)	Participants hold strong prior (incorrect) beliefs about role of weight Complex domain with real quantitative measures (and therefore somewhat ambiguous)	*Goal*: explore dev'l diffs in influence of DS knowledge on DG exptn strategies in a task w/real objects & forces; *RQs*: Are factors other than weight considered? Is task interpreted as assessing effect of each var on DV, or demonstrating correctness of belief? How does exptn affect belief revision? Prior belief affected initial goal (i.e., to demonstrate effect of weight); All Ss had pre-expt beliefs about weight, but only marginal diff w/younger children focusing on weight var; All children learned effects of other factors (e.g., material, shape) via experimentation; Older children more likely to view expts as testing hypotheses but younger children more likely to try to demonstrate a belief; 12- and 14-year olds conducted more unconfounded expts; Younger children experimented w/o explicit hypotheses
Schauble (1996) The development of scientific reasoning in knowledge-rich contexts	Grade 5/6 (*n* = 10), Non-college adults (*n* = 10)	Canal task (hydrodynamics) Spring task (hydrostatics)	Development of experimentation strategies in knowledge-rich contexts in which explanatory mechanisms between cause and effect may be invoked Continuous measures (varying effect sizes & measurement error)	*Goal*: to track changes in theories and reasoning strategies used by Ss who conduct/interpret cycles of expts to learn the causal structure of 2 phys systems; To address disagreement about the normative models of reasoning: logical validity vs. plausibility/explanatory coherence Success required both valid strategies and correct beliefs (bidirectional relation); Children and adults referred to causal mechanisms; variability in measurements limited progress in understanding, i.e., hard to distinguish error from small effect size--interpretation depends on theory Valid causal inferences acquired first, then inferences of non-causality and indeterminacy; Children more likely to conduct duplicate expts; Adults made greater gains on second task, conducting greater proportion of possible expts; Children spent time on trials w/vars they had correct beliefs about (demonstrating correctness of beliefs?); Both groups spent less time on trials of vars inconsistent with prior beliefs

Table 4 (Continued) Study	Participants	Task Domain(s)	Unique Feature	Main Findings
Kuhn & Pearsall (1998)	Grade 5 (10- & 11-year olds) (*n* = 47)	Physical tasks: Daytona (cars), Canal task (boats); Social tasks: TV enjoyment task, School achievement task	Microgenetic comparison of the co-development of strategic competence and two components of metastrategic knowledge (understanding the nature/requirements of the task & knowledge of strategies available and applicable) in multivariable tasks	Inferred hypothesis: in order for strategies to be used, there must be sufficient meta-strategic understanding that it applies to the goals/objectives of the task Metastrategic understanding was assessed by knowledge of task objectives and awareness of strategies needed to accomplish the task. Strategic performance was indicated by valid inferences (requiring unconfounded design, correct evidence interpretation and correct conclusion). Children working with physical task showed higher metastrategic understanding and strategic performance than those working on social task; however, 79% of all Ss showed increase in understanding, performance, or both over the course of 7 weeks; Metastrategic knowledge argued to be both distinct from, and related to, strategic performance; coherent pattern of connection found, with a "bootstrapping" or bidirectional relation between the two.
Chen & Klahr (1999) All other things being equal: Children's acquisition of the control of variables strategy	Grades 2, 3, & 4 (*n* = 87)	Springs task, Ramps task, Sinking objects task, Transfer: Paper & pencil tasks (natural & social science domains) given 7 months later	Explicit & implicit training of the control-of-variables (CVS) strategy; unprompted exploration (no direct instruction of domain knowledge)	*RQs*: Can early elem school children gain a genuine understanding of CVS and valid inferences? Can they transfer a learned strategy? Which type of training is most effective for learning and transfer? Are there dev'l diffs? Direct instruction, but not implicit probes, improved children's ability to design unconfounded experiments; CVS resulted in informative tests which facilitated conceptual change (domain knowledge); Ability to transfer the newly learned strategy increased with age – 4th graders showed skill retention 7 months later on transfer task w/diff experimenter, diff domain, diff test format
Toth, Klahr, & Chen (2000) Bridging research and practice: A cognitively based classroom intervention for teaching experimentation skills to elementary school children	Grade 4 (*n* = 77)	Ramps	Classroom verification of lab findings (Chen & Klahr, 1999); Explicit vs. implicit training of CVS strategy	*RQ*: can 4th graders learn and transfer CVS when taught in classroom setting? What relations exist b/w expt skills and acquisition of domain knowledge? Does direct instruction transfer to ability to evaluate the designs of others? CVS performance increased from 30% at pre-instruction to 96% with direct instruction; 78% were able to provide a rationale for the CVS strategy; Significant gains in domain knowledge from pre- to post-CVS instruction; CVS instruction resulted in gains in ability to evaluate the experiments of others

Table 4 (Continued) Study	Participants	Task Domain(s)	Unique Feature	Main Findings
Gleason & Schauble (2000) Parents' assistance of their children's scientific reasoning	9- to 12-year-olds Adults (parents) (*n* = 20 dyads)	Canal task (hydrodynamics)	Parent-child dyads in a museum context Given general superiority of adults on SR tasks, are they able to assist their children during collaborative discovery?	*RQ*: How do parents assist their children in a task when the parent does not know the solution? What evidence-generation and interpretation strategies do dyads use? What changes in knowledge occur? Dyads covered 75% of Expt space; Half were partially organized in approach, using logically related trials, other half used local chaining; Selecting features to test shared by parent & child; recording/consulting data done by parents; Inferences (causal/non-causal/indeterminate) were made on 48% of trials, of these 85% were deemed valid); Parents more likely to make inferences (67%); Parents' knowledge improved; children's knowledge did not change; children tended to adopt final beliefs endorsed by parents. Guidance of parent resulted in effective strategies & inferences; yet parents also missed opportunities to provide assistance, especially with respect to evidence interpretation
Kuhn, Black, Keselman, & Kaplan (2000) The development of cognitive skills to support inquiry learning	Grades 6, 7 & 8 (*n* = 42) *Aside*: Some evidence that "make a difference" means something different to a 6th grader. Protocol suggests it is used in a way to indicate producing a good or desirable outcome; but only one level of a var	Multivariable systems: Flooding problem task & Teacher's aide task (transfer) The assertion that inquiry learning benefit students rests on assumption that they have the cognitive skills to engage in activities in a way that will meet educational objectives	Compared two kinds of practice: Performance-level exercise (C) vs. Metastrategic-level exercise (E) Inquiry learning = educ'l activity in which Ss investigate phenom (real or virtual) and draw conclusions about it	In microgenetic context, some Ss use CVS & make valid inf's, but some Ss show strategic & inference weakness. *Hypothesis*: these Ss may have an incorrect mental model of causality and this explains lack of progress Prediction error & valid inferences: pre-post diffs but no diffs C vs. E Understanding: pre-post differences; Time x Cond intn: more sizeable change for the E group; At post-test, 71% of E and 57% of C understood need to investigate a single feature at a time; Implicit understanding better than explicit understanding of correct choice on transfer task, however, E group better at metalevel understanding. Overall, E group improved in strategic perf and in metalevel und of strategy, gains evident in transfer to a novel task Knowledge: both groups increased their knowledge of the causal system, no diffs E vs. C, both groups maintained incorrect beliefs despite much evidence generated over course of exploration of the system
Echevarria (2003) Anomalies as a catalyst for middle school students' knowledge construction and scientific reasoning during science inquiry	Grade 7 (n = 46)	Genetics simulator	Reactions to anomalous findings	Hypothesis or RQ: In context of scientific inquiry, do Ss choose to pursue/design tests that produce anomalous outcomes? Or do they favour tests that produce outcomes consistent with theories? How does either approach influence knowledge construction? "Ss generated hypotheses, ran tests, and constructed explanations in proportion to the extent to which they encountered anomalies. More anomalies meant more hypotheses, tests and explanations were generated."

Table 4 (Continued) Study	Participants	Task Domain(s)	Unique Feature	Main Findings
Masnick & Klahr (2003) Error matters: An initial exploration of elementary school children's understanding of experimental error	Grades 2 & 4 (8- and 10-year-olds) (*n* = 49)	Ramps task (4 binary variables: height, surface type, length, ball type)	Role of experimental error and how it affects the design, execution and interpretation of evidence	*RQs*: Can children differentiate the role of error in absolute versus relative measurements? Are children able to generate reasons for variations in repeated measures? Can children recognize potential sources of error? Students' reasons for data variability – 79% of 2nd graders and 100% of 4th graders mentioned execution errors; 34% of 2nd graders and 65% of 4th graders mentioned measurement errors. All students recognized hypothetical measurement and execution errors could affect outcome; 22% of 2nd graders and 50% of 4th graders offered correct predictions & correct reasons for error effects; Overall, 90% of children were able to name at least one source of experimental error (100% of 4th graders).
Keselman (2003) Supporting inquiry learning by promoting normative understanding of multivariable causality	Grade 6 (*n* = 74); 3 intact classrooms two-part intervention to improve Ss perf and metalevl sci competency by strengthening MM of causality	Multivariable systems: Earthquake forecaster Flooding problem task (modeled for direct instruction group) Avalanche hunter (transfer)	*Control* (performance-level exercise) Two forms of instructional support: *Practice condition* (performance-level plus prediction practice) *Instructional condition* (direct instruction plus performance-level and prediction practice)	*Hypothesis*: each part of the intervention would produce improved use of SR strategies (e.g., valid inferences + knowledge gains), and metalevel appreciation for strategies Low consistency between implicit and explicit causal theories; Pre- to post-test differences on number of valid inferences; trends toward differential improvement for intervention conditions not significant; Pre- to post-test differences on number of evidence-based responses with most improvement for interventions with practice and/or instruction; pattern held for transfer task; Only the instructional group showed increase in use of evidence from multiple records and correct beliefs about non-causal variables; Students in the practice and instructional conditions showed better performance on the metalevel assessment relative to the control.
Zimmerman, Raghavan & Sartoris (2003) The impact of the MARS curriculum on students' ability to coordinate theory and evidence	Grade 6 (*n* = 14)	Balance-scale task 2 causal variables (weight, horizontal distance from fulcrum) and one non-causal variable (vertical height on balance arm)	Comparison of effects of two science curricula: one inquiry-based, one inquiry-based with focus on quantitative models of data	*Hypothesis*: Repeated exposure to experiments which include a cycle of prediction, expt, etc., when compared to Ss exposed to a single curriculum unit on CVS, will perform better on a curriculum-neutral discovery task. *RQ*: Could Ss approaches to the task be characterized as consistent with the "experimenter" and "theorist" approaches? Students instructed in the model-based reasoning curriculum were more successful in focused exploration of the balance task, and all were able to induce a quantitative rule for predicting balance, regardless of whether they took an "experimenter" or a "theorist" approach to the task. As such, they were more successful at transferring this rule to a set of application questions. Only one student in the traditional inquiry-based curriculum discovered the quantitative rule (an "experimenter").

Reid, Zhang & Chen (2003) Supporting scientific discovery learning (SDL) in a simulation environment	12- and 13-year olds (approx. Grade 6) (*n* = 78) ES (y/n) x IS (y/n) design (BSs): ES&IS, ES, IS, no support	Simulated floating and sinking environment; Ss compare two trials IVs: shape, mass, volume (only mass affects DV); DV: upthrust of floating object; Online data-recording sheet Maximum 35 minutes on task, minimum of 5 experiments set for the exploration session	Types of learning support: *Interpretive support* (IS) – helps w/ knowledge access, hypoth. gen'n, etc. *Experimental support* (ES) – helps w/design, prediction, observation, conclusions 4 measures: Principled Knowledge (PK), Intuitive Understanding (IU): predictions of upthrust for pairs of objects), flexible application (Transfer prob: upthrust of boat on lake), and Knowledge Integration (KI). Subset of first 2 used at pre-test	Why does active inquiry not improve learning outcomes more consistently? Effectiveness of SDL *hypothesized* to be related to (a) meaningfulness of discovery processes (activ'n & mapping of prior knowledg to prob. represent'n and gen,n of hyp's);(b) systematic manip'n of vars/design and implementation of the expt; (c) reflective generalization–self-monitoring, integration of discovered rules/principles Diff types of support expected to enhance these activities components. Pre- Post: all groups showed improvement in PK/IU but groups with ES made greater knowledge gains. Post-test: No effect of cond. on PK; Groups with IS showed greater IU/prediction scores; marginal int'n w/ES (advantage of ES&IS); for KI there was an advantage for groups with IS (no int'n) Evaluating experiments (% controlled expts; avg. # vars varied; % of E-space used): No diffs b/w ES and no-ES; however, diffs on all 3 indices for students grouped by success (discovery of which IVs do/not affect DV) or no-success. Successful discovery associated with better experimentation strategies.
Triona & Klahr (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	Grades 4 and 5 (*n* = 92)	Ramps task Springs task	Compared instructional methods using either physical or virtual materials	*Hypotheses*: Based on previous views: (a) computer presentation will decrease learning b/c phys interaction is essential; OR (b) no difference b/w Phys & Virtual as long as instruction is similar Two types of materials were both effective in training students to design unconfounded experiments; Both groups were able to justify their design, make valid inferences, and transfer skills to a new domain; Confidence in conclusions was similar for both groups; Results call into question the pedagogical assumption about the necessity of "hands on" learning
Klahr & Nigam (2004) The Equivalence of Learning Paths in Early Science Instruction: Effects of Direct Instruction and Discovery Learning	Grades 3 and 4 (*n* = 112)	Acquisition task: ramps task Transfer task: evaluating science fair projects (about 1 week later with a different experimenter blind to condition)	Comparison of direct instruction vs. discovery learning Discovery = no teacher intervention, no guiding questions, no feedback Direct Instruction = teacher controlled goals, materials, examples, explanation and pace	*H1*: Direct instruction is more effective than discovery learning in teaching children CVS (replication); *H2*: Students who have mastered CVS will outperform those who do not on a transfer task involving the evaluation of science fair posters (e.g., critique of design, measures, conclusions, etc.); *H3*: What is learned is more important than how it is taught – i.e., the *Path-independent transfer hypothesis*: Mastery, regardless of method of attainment, will result in transfer of knowledge DI better than DL for designing unconfounded expts and appropriate inferences; more direct instruction Ss reached mastery of CVS strategy. Ss who attained mastery of CVS, regardless of learning path, scored better on a transfer task

Table 4 (Continued) Study	Participants	Task Domain(s)	Unique Feature	Main Findings
Kanari & Millar (2004) Reasoning from data: How students collect and interpret data in science investigations Very little SR literature cited; nothing past 1996 except Chinn& 1998 ref and the Masnick error stuff	10-, 12-, & 14-year-olds (*n* = 60)	Pendulums task (IVs: string length, mass: DV: time of swing) or Boxes (friction) task (IVs: weight, surface area; DV: force needed to pull the box) Both tasks involved one IV that covaried with the DV, and one that did not; Each IV had 5 possible levels Getting from "measured primary data" to a statement of "results" 3 hypothesis cards; Initial, Change, & Final e.g. increase in X = (increase, decrease, no change) in Y	Little "guidance" from experimenter during exploration; also watched videos of Ss investigating the 2nd task domain and then interviewed * assert that many previous SR studies are actually *logical reasoning* tasks (e.g., Kuhn et al. 1988, Koslowski 1996) because "data" was not presented – only "results" or "findings" were presented - the "effect size matters"	*AIMS*: to ID common approaches and patterns in Ss' reasoning as they collect data and draw conclusions; to explore how performance varies as a function of task type, age and school experience; to explore how Ss' deal with measurement uncertainty/error when collecting/interpreting data *Hypothesis*: Ss will experience greater difficulty, and will be less successful in reaching the correct conclusion, in investigations of a variable that does not affect an outcome than of a 'non-causal variable Tendency to explore covariation hypotheses (vs. non-covariation); 85-95% of Ss used CVS strategy; Ss recorded more data points for Boxes task compared to Pendulums task; In exploring levels of IV, strategies tended to be difference- or trend-focused; No age differences were found for CVS or for correct conclusions Marked differences when exploring IV that covaried vs. one that did not; all students drew correct conclusion when exploring an IV that covaried; only half drew correct conclusion when exploring IV that did not covary with the DV; repeat measurements more likely when exploring non-covar IV; selective recording or repeat measures taken to make non-covar data fit a trend to be consistent with (incorrect) conclusion (i.e., distort or "(f) reinterpret" data); puzzled by data and thought 'inconclusive'; selective attention to certain data; Students lacked an awareness of measurement error; general difficulties when IV does not covary with DV
Metz (2004) Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design	Grade 2 (*n* =21; 10 teams) Grade 4/5 (*n* = 31; 14 teams)	Structured interviews following completion of Student-regulated investigations of animal behavior (crickets) Analytical lens: focus on children's understanding of uncertainty in various components of their investigations	Children were asked about their research question; their findings and level of confidence; what could be done to be more confident; how the study could be improved, what question and method would be used for a new study; generalizability of their findings; how the project was similar to the work of scientists	Any of the following 5 categories could be a "sphere of uncertainty": (a) producing the desired outcome; (b) data; (c) identified trends; (d) generalizability of trend; (e) theory to account for the trend At both grade levels, majority of students (71–83%) conceptualized at least one sphere of uncertainty, with half of each age conceptualizing 2 or 3 spheres. Most common: trend identified and theory to account for trend. Trend as uncertain differed by age; common reasons: insufficient data (7 types, e.g., too few org's, too few trials, etc.), research design, expt procedure/apparatus, and weak/conflicting trend. Children understood that data collection is not a straightforward process (e.g., researcher or instrumentation error); understood generalizability related to research plan, sampling; Most students who were aware of uncertainty (80-97%) could propose a study modification that would address the source of the uncertainty

Tytler & Peterson (2004) From "try it and see" to strategic exploration: Characterizing young children's scientific reasoning	5-6 year olds (*n* = 15)	Classroom tasks: mealworm exploration, buoyancy (floating and sinking), whirlybird exploration (rather than "rarefied tasks that comprise cognitive research") Semi-structured interview with kids as they free-form explored different domains	Note: This paper talks about Piaget being "enormously influential"; it does not cite any recent SR literature (1997 at the latest); outlines 3 arguments against or objections to cognitive developmental work; No new findings given critical stance of SR literature and lack of awareness of new lit	*Goal*: "there is a need to tease apart the dimensions that describe SR" (nature of approach to exploration; levels of processing; response to anomalous data; dealing with competing knowledge claims; knowledge constraints on performance); *RQ*: How can we characterize SR? Can these dimensions be related to epistemological reasoning? Do Ss demonstrate a coherent approach to exploration? What links between SR and conceptual knowledge? Pattern of individual differences across kids; consistency within; some patterns across dimensions; performance differed for the 3 different domains: *Results consistent with other SDE studies, but with younger Ss and more qualitative methodology and with "classroom" tasks.* Suggest that knowledge influence evidence evaluation and vice versa. There is something somewhat vague about some of the dimensions listed; "The classroom teacher, Sally, did not regard herself as science-focused" (p. 100); They did not use "lab tasks" used by cognitive science but then failed to show how their "results" will actually make more sense to teachers
Garcia-Mila & Andersen (2005; unpublished manuscript under review)	Grade 4 (n = 15) and community college students (n = 16)	Boats, Cars, TV enjoyment task, School achievement task 48 combinations of variables per task	Focused on Ss notetaking during inquiry over the course of 10 weeks (20 sessions); Ss given a notebook to use in any way they chose	RQs: Do Ss choose to take notes {and when?}?; Which kinds of info are recorded by Ss (e.g., evidence, theory)? Are notes complete? Ss progress in notetaking hypothesized to be related to progress in scientific thinking. Adults more likely to take notes (15/16) than children (8/15); adults took more notes (*M*=65 entries) than children (*M*=21.5 entries); adults' usage remained constant across sessions; children's usage decreased over time; use of evidence-based notes increased over time for adults, but decreased for children; high within-subject variability in notetaking; children's notes were not complete enough to allow a replication or be useful in drawing a conclusion; about a third of adults' notes were complete. Making complete notes was correlated with making valid inferences.
Kuhn & Dean (2005) Is developing scientific thinking all about learning to control variables?	Grade 6 (n = 42) E=12; C1=18; C2=12 academically at risk students	Earthquake Forecaster inquiry software; 5 binary variables Control 1: Single individual session with E.F., otherwise in regular science classes Control 2: Practice with E.F. minus suggestions to focus on a single variable Transfer: Ocean Voyage; Delayed Transfer (3 mos): Earthquake Forecaster	Intervention was a simple suggestion to examine only one variable at a time. (I interpret the intervention as one that induces the CVS strategy; vs. a question-formulation intervention as interpreted by the authors; may invoke meta-level und~ but at core is instruction to "VOTAT" or use CVS)	*RQ*: The goal was to test an intervention involving suggesting to students that they try to find out about one thing at a time. Did Ss act on the instruction to focus inquiry on the effects of a single variable? All Ss in the E group (100%) used CVS vs. 11% in C groups; mean number of vars investigated was 3.1 (of 5 possible), so C Ss simultaneously manipulated many variables.Diffs b/w E and 2 C groups on mean # of valid inferences found at immediate and delayed (3 mos) assessments; marginally sig diff in mean valid inferences for the transfer task. Conclusion was that the initial question-formulation phase is as important as CVS. Operationalized as a decision about which one of a # of vars to be investigated. Suggest it is as impt to teach Ss *why* to use a strategy rather than how to execute the strategy

Table 4 (Continued) Study	Participants	Task Domain(s)	Unique Feature	Main Findings
Kuhn & Dean (under review) Scaffolded development of inquiry skills in academically-at-risk urban middle-school students	Grade 6 (n = 52) E=22; C1=18; C2=12 E = Scaffolding C1 = regular instruction C2 = Inquiry practice	Music Club (for scaffolding group); sales: illustration (CD cover, musician); format (booklet, foldout); color type (color, B&W) Earthquake Forecaster inquiry software; 5 binary variables	Scaffolding 12 sessions as Advisors to a music club (examining the effect of catalog changes on CD sales); Data presented for a variety of cities with different weekly conceptual goals Session 13 – SDE with Earthquake forecaster	Hypothesis: at-risk classrooms do not provide occasions to engage in sustained thinking; By allowing sustained, scaffolded [instruction] with a topic of interest there would be the facilitation of the development and transfer of inquiry skills; Challenge: to get students to realize in such tasks there is "something to find out" Immediate Assessment: Intent to focus on one var: E = C2 > C1 CVS: E > C1 = C2 Valid inferences: E > C1 = C2 Delayed assessment (3 mos later): E had no decline in perf; E = C2 > C1 CVS: decline for E; E = C2 > C1 Valid inf: decline for E; E > C1 = C2
Kuhn (under review) Developing mental models of multivariable causality	Grade 4 (n = 30) Ss worked in pairs	Investigation module Earthquake Forecaster (EF) inquiry software Ocean Voyage (OV) inquiry software; 5 binary variables (2 non-causal; 3 have additive causal effects) Prediction module to assess skill; based on a set of vars, task is to predict the outcome; no feedback provided	Exploration of how mental models of causality are related to performance on SDE task Will students who are successful at developing traditional SR skills be able to perform on the prediction task (aka have a correct mental model of causality)	*RQ*s: Do mental models incorporate a consistency principle?); Do MM allow for operation of multiple causal factors? *H1:* If immature MM of causality are an epiphenomenon of immature scientific method, then weaknesses in attributing causality should disappear if such skills are taught/dev'd; *H2*: weaknesses in MM of C~ are attributable to lack of familiarity in content domain Ss comparisons across two groups: (1) successful progress (n=19) and (2) those who did not; Successful Progress: focused intent to investigate a single var, choice of CVS; approp. evidence-based inferences; in both the practice domain (OV) and the new domain (EF) Successful group – does their mastery of SR skills and familiarity w/content show evidence of improving MM of causality? Success did not translate to ability to make predictions about a particular constellation of variables (op. def. of MM of C~); Ss often explained prediction by mentioning a single var rather than the actual 3 causal vars which were discovered through experimentation

Developmental Differences

In this section I will describe findings that characterize children's performance, and where possible, make comparisons among groups of children or make comparisons between children and adults. Findings that relate to changes that occur within or across sessions will also be noted. Studies included for review have been published since 1988, and include all or most of the characteristics of SDE studies described previously (i.e., include components from cells A through F from Table 1). There are 27 studies that include children as participants (23 published, 4 manuscripts in press/under review). Six studies included both children and adults as participants, whereas 21 studies reported data from children only (K-8, or age 5 through 14). Twelve of these studies follow only one group of children, and nine report findings from two or three age/grade levels. Studies reviewed in this section will primarily focus on SDE studies that do not include any specific interventions. Studies that incorporate instructional or practice manipulation will be discussed in a separate section.

Searching Hypothesis Space: Prior Knowledge and the Selection of Hypotheses

When asked to engage in an investigation or discovery task, both knowledge and problem-solving strategies are important. Individuals come to the task with either existing conceptual knowledge of the task domain, or hypotheses are developed about how a system operates during the course of investigation: "What defines the enterprise, and itself undergoes development, is the subject's effort to coordinate this existing understanding with new information." (Kuhn et al., 1992, p. 320). How individuals begin this enterprise is of interest

In multivariable systems, participants may 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, 1996). Success in SDE tasks is associated with a search for hypotheses to guide experimentation (Schauble & Glaser, 1990). Children tend to focus on plausible hypotheses and often get "stuck" focusing on a single hypothesis (e.g., Klahr et al., 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, 1996; Schauble, 1990; 1996; Zimmerman et al., 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; 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., 1992; 1995; 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) progress(ed) was a function of type of experiments conducted and the types of inferences generated (such choices will be addressed subsequent sections).

Predictions and Plausibility: Bridging the Search for Hypotheses and Experiments

Predictions. 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 the experiments that they run (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, 2003).

Plausibility. 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 as a guide for 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. For implausible hypotheses (provided to participants 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, the finding that both adults and children tend to begin exploration of a causal system by focusing on variables consistent with prior beliefs are thus variables considered plausible candidates to be causally related to the outcome (e.g., Kanari & Millar, 2003; Penner & Klahr, 1996; Schauble, 1996).

Searching Experiment Space: Strategies for Generating Evidence

As discussed in the review of studies focusing on experimentation strategies, 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 (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.[7] 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 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 an initial task domain (first three-week period), both children and

[7] 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.

adults started out by covering about 60% of the experiment space. When they began experimentation of a second task domain (second three weeks), only adults' search of experiment space increased (to almost 80%). Over the 6 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, who were selected a broader range of experiments. 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 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., 1992; 1995; 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., Kuhn et al., 1992; Garcia-Mila & Andersen, 2005; Gleason & Schauble, 2003; Schauble, 1990; Siegler & Crowley, 1991; Siegler & Shipley, 1995). Developmental transitions do not occur suddenly. That is, participants do not progress from an inefficient or invalid strategy to a valid strategy without ever returning to the former.[8] With respect to experimentation strategies, an individual may begin with invalid HOTAT or change-all strategies, but once the usefulness of the CVS is discovered it is not immediately used exclusively. The newly discovered effective strategy

[8] Multiple strategy use has been found in research on the development of other academic skills such as math (e.g., Bisanz & LeFevre, 1990; Siegler & Crowley, 1991), reading (e.g., Perfetti, 1992) and spelling (e.g., Varnhagen, 1995).

(CVS) is only slowly incorporated into an individual's set of strategies, potentially because repeated exposure to a problem results in participants' dissatisfaction with the strategies that do not result in progress or 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 et al. (1981) originally noted that many of the early tasks used to study scientific reasoning were somewhat artificial because real scientific investigations involve *aided* cognition. Such memory aids ensure a level of authenticity that the task remains centered on reasoning and problem solving and not memory.

Previous studies of experimentation demonstrate that children are not often aware of their own 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 and keep track of which substance was responsible for producing a bubbling reaction in a mixture of yeast, flour, sugar, salt and warm water. Dunbar and Klahr (1988) 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 (2005) examined fourth graders' and adults' use of notetaking during a 10-week investigation of a number of multivariable systems. Unlike some studies, notetaking was not required, and 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. Moreover, despite variability in the amount of notebook usage in both groups, on average adults made 3 times more notebook entries than children did. Adults' notetaking remained stable across the ten weeks, but children's frequency of use decreased over time, dropping to about half of their initial usage. The researchers suggest that the children may have been unaware 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 experimental tests (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 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 what 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 a complex task. Metamemory development has been found to develop 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 reasoning (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., White & Frederiksen, 1998; Kolodner et al., 2003). Record keeping is an important component of scientific investigation in general, and of SDE tasks because access to and consulting of *cumulative* records often is an important component 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

The inferences that are made based on self-generated experimental evidence are typically classified as either 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, such as that which 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., that a variable is causal) or exclusion (i.e., that a variable is not causally related to outcome) that are based on controlled experiments that include both levels of the causal and outcome variables (e.g., Kuhn et al., 1992; 1995; 2000; Schauble, 1990; 1996; Schauble, Klopfer, et al., 1991). Even after discovering how to make inferences under these conditions, participants often have difficulty giving up less-advanced inference strategies such as false inclusion and exclusion inferences that are consistent with prior beliefs, or are based on a single instance of covariation (or noncovariation) between antecedent and outcome, or are based on one level of the causal factor and one level of the outcome factor (Klahr et al., 1993; Kuhn et al., 1992; 1995; Schauble, 1990; 1996).

Children have a tendency to focus on making 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 (or "false inclusion" inferences) 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 their experimentation was more often done using CVS (Schauble, 1996). Children's inferences were valid 44% of the time, compared to 72% for adults. The fifth- and sixth-graders did improve 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 relative to 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 swing of a pendulum or the force needed to pull a box along a level surface, but their self-directed experimentation only lasted for one session. 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 student 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 micorgenetic contexts where children must continue their investigation (e.g., Schauble et al., 1991), 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).

Continuous outcome measures and the understanding of measurement error. Recently, Kanari and Millar (2004) 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 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; Schauble, 1990) or lack of difference (e.g., object A had the same sinking time as object B; Penner & Klahr, 1996) or a set of categorical outcomes (e.g., a particular variable has a low, medium-low, medium-high, or high risk of earthquake, avalanche or flooding; Keselman, 2003; Kuhn & Dean, 2005).

Schauble (1996) incorporated the use of quantitative measures as part of children and adults' exploration of tasks involving hydrodynamics and hydrostatics. 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, whether they be simple experiments conducted by a science student or research conducted by a scientist. 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-grader 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 3). 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

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 the belief with the 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." Alternatively, 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. Most of the belief changes for both children and adults were for the variables for which individuals had no expectations (Schauble, 1996).

As suggested by the review of evidence evaluation studies, some individuals cannot or do not disregard prior theories or expectations when they evaluate evidence. That children and adults' do in fact pay attention to theoretical concerns is evidenced by the fact that individuals 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. Children and adults' consideration of theoretical relationships is also evident by their 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.

Evaluating "anomalous" evidence. One of the features of the causal systems used in SDE research is that they may be deliberately chosen to exploit known misconceptions (Schauble et al., 1991), such as that heavy objects fall faster than light objects. In the case of computer simulations, a task may incorporate factors that do or do not conform to intuitions for unfamiliar domains (e.g., predicting earthquake or flooding risk; Keselman, 2003; Kuhn & Dean, 2005). Any time a finding is unexpected, it could by definition be considered an *anomaly*. However, a specific task variant that has been explored has been to examine the effect of "anomalous" evidence on students' reasoning and knowledge acquisition. Researchers have argued that "surprising results" are an impetus for conceptual change in real science (e.g., Klahr et al., 1993; Klayman & Ha, 1987) and which is consistent with the work of T. Kuhn (1962) in the history and philosophy of science.

Penner and Klahr (1996) used a task in which there are rich prior beliefs – most children believe that heavier objects sink in fluid 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.

What is also important to note about the children in the Penner and Klahr study is that they did in fact <u>notice</u> 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 an

important mediator of conceptual change.

Echevarria (2003) examined seventh-graders reactions to anomalous data in the domain of genetics and whether they served as a "catalyst" for knowledge construction during the course of self-directed experimentation. 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 the initial knowledge was insufficient and therefore could not undergo change as a result of the encounter with "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.

Evaluating evidence in physical versus social domains. Kuhn et al. (1995) found differential performance for physical domains (i.e., microworlds involving cars and boats) and social domains (i.e., determining the factors that make TV shows enjoyable or make a difference in students' school achievement) in many respects. 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 the experimental evidence (often insufficient or generated from uncontrolled comparisons). 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 reasoning 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 both natural, physical, and social domains. For example, Zimmerman, Bisanz, & 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 reasoning tasks makes it an important factor to consider. Rozenblit and Keil found that the "degree of causal transparency for a system" (p. 554) 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 vs. physical) may be a factor in more or less proficient reasoning and conceptual development.

Knowledge of how variables and variables levels causally "make a difference." Kuhn et al. (2000; Kuhn, under review) have recently suggested that performance on SDE tasks may also be the result of a faulty mental model of "multivariable causality" in general (as opposed to being context specific) and that many children (and adults) have an insufficient understanding that the effects of variables are additive and consistent. Kuhn (2005b) found that fourth graders who made substantial progress over the course of several weeks on typical outcome measures (e.g., focused intent, CVS use, evidence-based inferences) were unable to make predictions about

constellations of variables (i.e., the task used to assess mental model of causality) that were learned through SDE. That is, another piece of the puzzle may involve individuals' general understanding of the nature of causality itself (Grotzer, 2003).

Additionally, an interesting finding in the Kuhn et al. (2000) study points to the idea that some students may have an alternative interpretation of what "makes a difference" means. The set of variables in SDE studies are almost always pre-selected by the researcher, and moreover, the levels of those variables are pre-selected. For example, a student directed to find out if the evidence shows whether "type of condiment" makes a difference in catching a cold (Kuhn et al., 1988), the values of *ketchup* and *mustard* are provided. A student with an alternative understanding of this situation may instead interpret this as a question that requires a "ketchup versus no ketchup" design (or cognitive model; Chinn & Brewer, 2001) and may interpret evidence in light of that interpretation. Similarly, in the flooding prediction task used by Kuhn et al. (2000), variables and values such as water temperature (hot, cold) and soil type (sand, clay) are pre-selected. It is suggested that a "co-occurrence model" may influence students' understanding and interpretation of the task requirements: "it is the feature levels sandy soil and hot water (rather than soil type or water temperature, as features) that are implicated as causal in interpreting the successful outcome" (p. 499). That is, "[r]eflecting another form of inconsistency, rather than soil type making a difference, sand does (but clay does not) make a difference" (p. 500).

Bootstrapping Experimentation Strategies and Conceptual Change

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 reasoning, 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).[9]

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). Bauer (1992) also noted that despite the great differences among the various scientific disciplines, within each there are individuals who specialize as theorists or experimenters. Klahr and Carver (1995) observed that "in most of the natural sciences, the difference between experimental work and theoretical work is so great as to have individuals who claim to be experts in one but not the other aspect of their discipline" (p. 140).

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 test the predictions of the hypotheses, or as Simon (1986) described: "draw out the implications of the theory for experiments or observations, and gather and analyse data to

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

test the inferences" (p. 163). Experimenters tend to make data-driven discoveries, by generating data and finding the hypothesis that best summarizes or explains that data.

Dunbar and Klahr (1989) and Schauble (1990) also found that children conformed to the description of either theorists or experimenters. In a number of studies with adults, success was correlated with the ability to generate multiple hypotheses (e.g., Schauble, Glaser, et al., 1991) and, in particular domains, theorists were more successful than the experimenters (Schauble & Glaser, 1990). Penner and Klahr (1996) had 10- to 14-year-olds conducting 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 as inquiry or investigation.

Zimmerman et al. (2003) were able to 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 their reasoning profile (i.e., 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 governing balance when they were classified as experimenters. Because they only interacted with the apparatus over the course of one 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, 1996). Given more time on task, it is conceivable that the students from the regular class who took a "theorist" approach would 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 was due to their repeated exposure to and practice with curriculum 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 reasoning 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. When testing the factors that produced a positive outcome, participants selected the less valid HOTAT strategy was. 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. In Kuhn and Phelps' (1982) study of experimentation strategies on the colorless fluids tasks, several children approached that 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 in a corked flask – they did not distinguish between "understanding a phenomenon and producing the phenomenon" (p. 516). In several studies, Kuhn and her colleagues (1992; 1995; 2000) also reported that early in investigations, students tend to focus on desirable versus undesirable outcome.

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." They suggested that some aspects of children's and adults' performance on scientific reasoning tasks could be elucidated by a consideration of what the participant believed the goal of experimentation was. Children worked on the canal task (an investigation in hydrodynamics) and the spring task (an investigation of hydrostatics).

When the children were working as scientists, their goal was to determine which factors made a difference and which ones did not. When the children were working as engineers, their goal was optimization, that is, to produce a desired effect (i.e., the fastest boat in the canal task, and the longest spring length in the springs problem). When working in the science context, the children worked more systematically, by establishing the effect of each variable, alone and in combination. There was an effort to make inclusion inferences (i.e., an inference that a factor is causal) and exclusion inferences (i.e, an inference that a factor is not causal).

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 to experimentation 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

Children's performance was characterized by a number of tendencies: to generate uninformative experiments, to make judgments based on inconclusive or insufficient evidence, 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., 1992; Kuhn et al., 1995; Kuhn et al., 2000; 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., 1992; Kuhn et al., 1995; Kuhn et al., 2000; Schauble, 1990; 1996; Schauble, Klopfer, 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 reasoning (e.g., Echevarria, 2003; Klahr et al., 1993; Kuhn et al., 1992; Metz, 2004; Penner & Klahr, 1996; 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 reasoning (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) was conducted largely in the absence of any specific instructional intervention. Children's scientific reasoning 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 focus on empirical investigations of scientific reasoning that include some type of instructional or practice intervention. 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 handful of studies focusing on the development of experimentation and evidence evaluation skills explicitly addressed issues of instruction and practice. These studies, interestingly, foreshadowed the very issues that are being investigated and addressed today. Siegler and Liebert (1975) incorporated instructional manipulations aimed at teaching children about variables and variable levels with or without practice on analogous tasks. In the absence of these conditions, no fifth graders and a small minority of eighth graders were successful. In the absence of explicit instruction, Kuhn and Phelps (1982) reported that despite the intra- and inter-session variability in strategy usage, extended practice and exercise over several weeks was sufficient for the development and modification of experimentation and inference strategies. This early microgenetic study tested the assumption 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., 1992, 1995; 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 on SDE tasks in the absence of

specific instructional interventions, but similar kinds of prompts were used continually through the course of the individual's exploration of 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 serves to alert participants to the underlying goal structure. Therefore, 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 of their role in generating 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 that no differences in learning outcomes were found between 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 a experimentation done in the absence of plans, a less thorough search of the experiment space, and the generation of a smaller set of evidence.

Students may not, in the absence of instruction or prompts, 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 the use of *self-explanations* supports this idea (e.g., Chi et al., 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, questions such as 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 experimental evidence. Interventions included practice at making predictions, reflecting on data, and explanation. 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 the course of children's 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.

Direct instruction about CVS improved children's ability to design informative experiments, which in turn facilitated conceptual change in a number of domains. Students' mastery of CVS 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. Although in other microgenetic contexts, extend practice with or without probes has been shown to be sufficient for improving students' usage of valid experimentation and inference strategies. 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% to 97%) and mean robust usage (6% to 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 (9 of 10 designs) increased from 28% to 76%. Therefore, the effectiveness of a lab-based intervention could be "scaled up" to a classroom context. (See Klahr & Li [in press] for a summary of research that alternates between the lab and the classroom.)

Klahr and Nigam (2003) 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 speed of a ball on a ramp. One group of students 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, <u>regardless of condition</u>, 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, Zhang and Chen (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 in order 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. Also, questions were offered to activate prior knowledge about, for example, forces acting on a floating object. *Experimental support* was designed to scaffold the design of experiments, making predictions and drawing conclusions based on observations. For example, an introductory part of the simulation environment included explanations about designs (e.g., how to vary one thing at a time). Another example of an experimental scaffold was a prompt to compare their 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.

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 (as is 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 (i.e., an evaluation of the strategies of two individuals who proposed different designs) relative to the control group. Only the instructional 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 in a SDE study with low performing inner-city students. Over the course of twelve sessions, sixth graders interacted with a virtual environment to understand the effect of five binary variables on earthquake risk. Students were either given self-directed practice (control) or they were provided with the suggestion "to find out about just one feature to start." 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. The use of CVS led to an increase in valid inferences for the intervention group, at both immediate and delayed (3 months) assessment. Performance on a transfer task was marginally better than the control. Kuhn and Dean concluded that the manipulation influenced the question-formulation phase and that it suggested to students *why* a strategy should be used. While this may be the case, it is also possible that this simple instruction did in fact invoke a 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, suggesting that the inclusion of even minor prompts and suggestions may represent potentially powerful scaffolds in the context of self-directed investigations, especially for at-risk populations of students.

Kuhn and Dean (2005b) have 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 were provided with 12 weeks of exposure to a multivariable environment in a domain that the students would find interesting, but that is not considered a traditional science topic. 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 the transfer task was compared to students who had only extended practice with the music club software 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 for CVS and valid inferences (but not for focused intent) for the scaffolded group, but their performance was still generally superior to the control groups.

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 in the evidence evaluation section, it has been suggested that developmental differences in scientific reasoning may be a function of task demands. For example, Sodian et al. (1991) and Ruffman et al. (1993) altered task demands to show that under simpler conditions, children could distinguish a hypothesis from evidence. 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, 2003), 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 reasoning, 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 reasoning in particular:

> . . . the complex and multifaceted nature of the skills involved in solving these problems, and the variability in performance, even among the adults, suggest that the developmental trajectory of the strategies and processes associated with scientific reasoning is likely to be a very long one, *perhaps even lifelong*. Previous research has established the existence of both early precursors and competencies . . . and errors and biases that persist regardless of maturation, training, and expertise. (Schauble, 1996, p.118; emphasis added)

A robust finding is that during this long developmental trajectory, there is both inter- and intra-individual variability in scientific reasoning performance, particularly with respect to inference and experimentation strategies. There is also much variability in the tasks that have been used to study scientific reasoning but there are some generalizations that that can be made from an examination of this literature to 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 vs. 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 limited instances (d) 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 be unsystematic in (g) recording plans, designs and outcomes, and may fail to (h) consult such records. Likewise, there is a slow development course for the (h) metacognitive understanding of theory and evidence as distinct epistemological entities and the

(i) 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 have been 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 reasoning.

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 reasoning 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 reasoning 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 & Frederikson, 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., Ecchevarria, 2003; Metz, 2004), strategy development (e.g., Chen & Klahr, 1999; Toth, Klahr & Chen, 2001) 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. Support for this may be found in Schauble et al.'s (1991) finding that students made the most progress when beginning in the "engineering context" and then moving to the "science context." 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, in press). 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 reasoning

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 reasoning be used engineer better instructional interventions. Such studies may then be used as a source of further basic questions (Klahr & Li, in press). 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.

How Do We Define and Measure "Best" and "Instruction"?

This may seem like an odd question, but issues are already emerging about how particular interventions should be labeled, and how such interventions should be assessed. As a parallel, 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 scientific reasoning research – 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 transfer or retention. Each of these tasks has the potential to be interpreted in multiple ways (e.g., as a valid measure of transfer, as a valid measure of strategic competence).

Similarly, the interventions used by educators or researchers may open to multiple interpretations. For example, Klahr and Li (in press) outlined different media reactions to the intervention studies conducted by Klahr and colleagues (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2003) based on multiple connotations of "direct instruction" and "discovery learning." Although Klahr and Nigam used "direct instruction" as a condition label, the students engaged in self-directed, hands-on experimentation after a brief didactic presentation and demonstration. Although performance in this group was better, a quarter of the "discovery learning" students mastered the task and showed equivalent performance on a transfer task. This example illustrates two issues of interpretation. First, the labels used to describe the different conditions were

interpreted in unintended ways, that is, "direct instruction" was taken to mean "passive learning." Second, these findings were interpreted to mean that "direct instruction" was advocated as the most efficient and desirable way to teach science. 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. An example of such a debate, that represents both fundamental issues – which instruction is best, and how should these terms be used – can be found in the set of commentaries by Klahr (2005b-in press) and Kuhn (2005a-in press).

These commentaries represent the types of debate and dialogue that will need to become an essential part of the next generation of research on evidence-based interventions. Another issue tackled is how to interpret the nature of the intervention itself. As mentioned previously, the intervention used by Kuhn and Dean (2005a) included a simple suggestion to students: "Today, let's try to find out about just one feature to start." (p. 7ms). Whereas Kuhn and Dean concluded that this suggestion influenced the question-formulation phase, an alternate interpretation is that it elegantly 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 issue cannot be resolved simply by appealing to the data. A similar definitional issue at the core of this commentary involves the best way to demonstrate that "transfer" has occurred. For example, Kuhn and Dean (2005b) 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) proposed nine relevant dimensions 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. 2ms).

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 the 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) be 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 reasoning 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 and Li [in press] 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

Chinn and Malhotra (2001; 2002b) recently 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 reasoning. 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 many previous studies of scientific reasoning should be classified as "logical reasoning" tasks because participants do not evaluate *numerical data*. In their view, authentic scientific reasoning 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 & Morris, 2002; Masnick & Klahr, 2003; 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. Herb 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 reasoning, 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" (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. 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. 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.

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist, 48*, 35-44.

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

Baillargeon, R., Kotovsky, L., & Needham, A. (1995). The acquisition of physical knowledge in infancy. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 79-116). Oxford: Clarendon Press.

Bauer, H. H. (1992). *Scientific literacy and the myth of the scientific method*. Urbana: IL: University of Illinois Press.

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.

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

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.

Case, R. (1974). Structures and strictures: Some functional limitations on the course of cognitive growth. *Cognitive Psychology, 6*, 544-573.

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

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

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.

Chi, M. T. H., & Koeske, R. D. (1983). Network representations of children's dinosaur knowledge. *Developmental Psychology, 19*, 29-39.

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,* 327-43.

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

Clement, J. (1983). A conceptual model discussed by Galileo and used intuitively by physics students. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 325-339). Hillsdale, NJ: Lawrence Erlbaum.

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.

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. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 365-395). Cambridge, MA: MIT Press.

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.

Feist, G. J., & Gorman, M. E. (1998). The psychology of science: Review and integration of a nascent discipline. *Review of General Psychology, 2*, 3-47.

Flavell, J. H. (1963). *The developmental psychology of Jean Piaget*. Princeton, NJ: Von Nostrand.

Garcia-Mila, M., & Andersen, C. (2005). Developmental change in notetaking during scientific inquiry. Manuscript submitted for publication.

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.

Gerber, B. L., Cavello, A. M. L. & Merek, E. A. (2001). Relationships among informal learning environments, teaching procedures and scientific reasoning ability. *International Journal*

of Science Education, 23, 535-549.

German, T. P. (1999). Children's causal reasoning: Counterfactual thinking occurs for 'negative' outcomes only. *Developmental Science*, *2*, 442-447.

Gholson, B., Shadish, W. R., Neimeyer, R. A., & Houts, A. C. (Eds.). (1989). *Psychology of science: Contributions to metascience*. Cambridge, MA: Cambridge University Press.

Giere, R. N. (1979). *Understanding scientific reasoning*. New York: Holt, Rinehart and Winston.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39*, 93-104.

Gleason, M. E., & Schauble, L. (2000). Parents' assistance of their children's scientific reasoning. *Cognition & Instruction, Vol 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.

Greenhoot, A. F, Semb, G., Colombo, J., & Schreiber, T. (2004). Prior beliefs and methodological concepts in scientific reasoning. *Applied Cognitive Psychology, 18*, 203-221.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93*, 216-229.

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

Hatano, G., & Inagaki, K. (1994). Young children's naive theory of biology. *Cognition, 56*, 171-188.

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

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology, 37*, 668-683.

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.

Hood, B. M. (1998). Gravity does rule for falling events. *Developmental Science, 1*, 59-63.

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

Hunt, E. (1994). Problem solving. In R. J. Sternberg (Ed.), *Thinking and problem solving* (pp. 215-232). San Diego, CA: Academic Press.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. McGilly, Kate (Ed). (1994). Classroom lessons: Integrating cognitive theory and classroom practice. (pp. 51-74). Cambridge, MA: The MIT Press.

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

Kaiser, M., McCloskey, M., & Proffitt, D. (1986). Development of intuitive theories of motion. *Developmental Psychology, 22*, 67-71.

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.

Karmiloff-Smith, A., & Inhelder, B. (1974). If you want to get ahead, get a theory. *Cognition, 3*, 195-212.

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

Keys, C. W. (1994). The development of scientific reasoning skills in conjunction with collaborative writing assignments: An interpretive study of six ninth-grade students. *Journal of Research in Science Teaching, 31*, 1003-1022.

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

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.

Klahr, D. (1994). Searching for the cognition in cognitive models of science. Psycoloquy, 5(94), scientific-cognition.12.klahr.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge: MA: 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-in press). Early science instruction: Addressing fundamental issues. *Psychological Science*.

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., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, *125*, 524-543

Klahr, D., & Simon, H. A. (2001). What have psychologists (and others) discovered about the process of scientific discovery? *Current Directions in Psychological Science*, *10*, 75-79.

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

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., Puntambekar, S., & Ryan, N. (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: MA: MIT Press.

Koslowski, B., & Maqueda, M. (1993). What is confirmation bias and when do people actually have it? *Merrill-Palmer Quarterly, 39*, 104-130.

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, 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-in press). What needs to be mastered in mastery of scientific method? *Psychological Science*.

Kuhn, D. (2005b). Development mental models of multivariable causality. Manuscript under review.

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., & Brannock, J. (1977). Development of the isolation of variables scheme in experimental and "natural experiment" contexts. *Developmental Psychology, 13*, 9-14.

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

Kuhn, D., & Dean, D. (2005a-in press). Is developing scientific thinking all about learning to control variables? *Psychological Science*

Kuhn, D., & Dean, D. (2005b). Scaffolded development of inquiry skills in academically-at-risk urban middle-school students. Manuscript Under Review.

Kuhn, D., & Franklin, S. (2006-in press). The second decade: What develops (and how)? To appear in *Handbook of child psychology* (6th ed.).

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.

Kuh, 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., & Phelps, E. (1982). The development of problem-solving strategies. In H. Reese (Ed.), *Advances in child development and behavior, 17,* 1-44.

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

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

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

Lajoie, S. (Ed.) (1998). Reflections on statistics: learning, teaching and assessment in grades K-12. Mahwah, NJ: Erlbaum.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

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.

Levin, I., Siegler, R. S., & Druyan, S. (1990). Misconceptions about motion: Development and training effects. *Child Development, 61*, 1544-1557.

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

Mahoney, M. J., & DeMonbreun, B. G. (1977). Psychology of the scientist: An analysis of

problem-solving bias. *Cognitive Therapy and Research, 1*, 229-238.

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. *Proceedings of the 24th annual conference of the Cognitive Science Society,* 643-648.

Mayer, R. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist, 59*, 14-19.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299-324). Hillsdale, NJ: Lawrence Erlbaum.

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

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.

Miller, J. L., & Bartsch, K. (1997). The development of biological explanation: Are children vitalists? *Developmental Psychology, 33*, 156-164.

Miller, K. F. (2000). Representational tools and conceptual change: The young scientist's tool kit. *Journal of Applied Developmental Psychology, 21*, 21-25.

Miller, S. (2004). Public understanding of science at the crossroads. *Public Understanding of Science*, *10*, 115-120.

Miller, J. D. (2004). Public understanding of, and attitudes toward, scientific research: What we know and what we need to know. Public Understanding of Science, 13, 273-294.

Minstrell, J. (2001). The role of the teacher in making sense of classroom experiences and effecting better learning. In Carver, S M & Klahr, D. (Eds) , Cognition and instruction: Twenty-five years of progress. (pp. 121-149). Mahwah, NJ: Lawrence Erlbaum.

Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological*

Review, 92, 289-316.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of information and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology, 30*, 395-406.

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.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Norris, S. P., Phillips, L. M., & Korpan, K. A. (2003). University students' interpretation of media reports of science and its relation to background knowledge, interest, and reading ability. Public Understanding of Science, 12, 123-145.

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.

Pauen, S. (1996). Children's reasoning about the interaction of forces. *Child Development, 67*, 2728-2742.

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.

Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145-174). Hillsdale, NJ: Erlbaum.

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.

Piaget, J. (1970). *Genetic epistemology*. New York, NY: Columbia University Press.

Piaget, J. (1972). *The child's conception of the world*. Totowa, NJ: Littlefield, Adams & Co.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Journal of Research in Science Teaching, 66*, 211-227.

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.

Samarapungavan, A., & Wiers, R. W. (1997). Children's thoughts on the origin of species: A study of explanatory coherence. *Cognitive Science, 21*, 147-177.

Sandoval, W. A., & Reiser, B.J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*, 345-372.

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

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.

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., 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.

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

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.

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

Shamos, M. H. (1995). *The myth of scientific literacy*. New Brunswick, NJ: Rutgers University Press.

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. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*, 481-520.

Siegler, R. S. (1978). The origins of scientific reasoning. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 109-149). Hillsdale, NJ: Lawrence Erlbaum.

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.

Simon, H. A. (1973). Does scientific discovery have a logic? *Philosophy of Science, 40*, 471-

480.

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.

Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 105-128). Hillsdale, NJ: Lawrence Erlbaum.

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition, 20*, 392-405.

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.

Solomon, G. E. A., Johnson, S. C., Zaitchik, D., & Carey, S. (1996). Like father, like son: Young children's understanding of how and why offspring resemble their parents. *Child Development, 67*, 151-171.

Spelke, E. S., Phillips, A., & Woodward, A. L. (1995). Infants' knowledge of object motion and human action. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 44-78). Oxford: Clarendon Press.

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

Stanovich KE 1998 How to the think straight about psychology (5th ed.). New York: Longman.

Strauss, S. (1998). Cognitive development and science education: Toward a middle level model. In W. Damon (Series Ed.), I. E. Sigel, & K. A. Renninger (Vol. Eds.), *Handbook of child psychology: Volume 4. Child psychology in practice* (5th ed., pp. 357-399). New York: John Wiley & Sons.

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

Taconis, R., Ferguson-Hessler, M., G., M., & Broekkamp, H. (2001). Teaching science problem solving: An overview of experimental work. *Journal of Research in Science Teaching*, *38*, 442-468

Thagard, P. (1989). Explanatory coherence. *Behavioral & Brain Sciences, 12*, 435-502.

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). Ulcers and bacteria II: Instruments, experiments, and social interactions. *Studies in History and Philosophy of Science. Part C: Studies in History and Philosophy of Biology and Biomedical Sciences*.

Thagard, P. (1998c). 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.

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

Tweney, R. D. (1991). Informal reasoning in science. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 3-16). Hillsdale, NJ: Lawrence Erlbaum.

Tweney, R. D., (2001). Scientific thinking: A cognitive-historical approach. In Crowley, K., Schunn, C. D., & Okada, T. (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.

VanLehn, K. (1989). Problem solving and cognitive skill acquisition. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 527-579). Cambridge, MA: MIT Press.

Varnhagen, C. K. (1995). Children’s spelling strategies. In V. W. Berninger (Ed.), *The varieties of orthographic knowledge II: Relationships to phonology, reading, and writing* (pp. 251-290). Dordrecht, Netherlands: Kluwer Academic Publishers.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*, 535-585.

Vosniadou, S., Skopeliti, I., & Ikospentaki, K. (2004). Modes of knowing and ways of reasoning in elementary astronomy. *Cognitive Development, 19*, 203-222.

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

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 63-71.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. London: B. T. Batsford.

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

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.

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

Wong, D. E. (1996). Students’ scientific explanations and the contexts in which they occur. *The Elementary School Journal*, *96*, 495-509.

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., & 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.