COGNITION AND INSTRUCTION, *18*(4), 495–523

The Development of Cognitive Skills To Support Inquiry Learning

Deanna Kuhn, John Black, Alla Keselman, Danielle Kaplan
Teachers College
Columbia University

Establishing the value of inquiry learning as an educational method, it is argued, rests on thorough, detailed knowledge of the cognitive skills it is intended to promote. Mental models, as representations of the reality being investigated in inquiry learning, stand to influence strategies applied to the task. In the research described here, the hypothesis is investigated that students at the middle school level, and sometimes well beyond, may have an incorrect mental model of multivariable causality (one in which effects of individual features on an outcome are neither consistent nor additive) that impedes the causal analysis involved in most forms of inquiry learning. An extended intervention with 6th to 8th graders was targeted to promote (a) at the metalevel, a correct mental model based on additive effects of individual features (indicated by identification of effects of individual features as the task objective); (b) also at the metalevel, metastrategic understanding of the need to control the influences of other features; and (c) at the performance level, consistent use of the controlled comparison strategy. Both metalevel advancements were observed, in addition to transfer to a new task at the performance level, among many (though not all) students. Findings support the claim that a developmental hierarchy of skills and understanding underlies, and should be identified as an objective of, inquiry learning.

The argument for inquiry learning as an educational tool is being heard increasingly, especially as the technology and materials to support this kind of educational experience have expanded and become widely available. Amidst the widespread enthusiasm, the strongest criticism to be heard is that such methods are inefficient. Too little substantive knowledge is gained to justify the sizable expenditure of classroom time that such activities typically consume. But outweighing this criticism in a majority of educators' eyes are the potential benefits of the opportunities

Requests for reprints should be sent to Deanna Kuhn, Box 119, Teachers College, Columbia University, New York, NY 10027. E-mail: dk100@columbia.edu

afforded students to engage in genuine inquiry. Highly favored in a recent National Research Council report (Bransford, Brown, & Cocking, 1999) is a method in which students

> design studies, collect information, analyze data and construct evidence. … They then debate the conclusion that *they* derive from their evidence. In effect the students build and argue about theories. … Question posing, theorizing, and argumentation form the structure of the students' scientific activity. … The process as a whole provide[s] a richer, more scientifically grounded experience than the conventional focus on textbooks or laboratory demonstrations. (pp. 171–172)

In formulating questions, accessing and interpreting evidence, and coordinating it with theories, students are believed to develop the intellectual skills that will enable them to construct new knowledge (Chan, Burtis, & Bereiter, 1997). In addition, they ideally are also acquiring a set of intellectual values—values that deem activities of this sort to be worthwhile in general and personally useful. In the words of Resnick and Nelson-LeGall (1997), students who value intellectual inquiry

> believe they have the right (and the obligation) to understand things and make things work … believe that problems can be analyzed, that solutions often come from such analysis and that they are capable of that analysis … have a toolkit of problem-analysis tools and good intuitions about when to use them … know how to ask questions, seek help and get enough information to solve problems … have habits of mind that lead them to actively use the toolkit of analysis skills. (pp. 149–150)

In short, students come to understand that they are able to acquire knowledge they desire, in virtually any content domain, in ways that they can initiate, manage, and execute on their own, and that such knowledge is empowering. This outcome is believed to justify the time devoted to development of these skills and dispositions within the context of what is typically a circumscribed topic of investigation.

Is inquiry-based education capable of delivering on these promises? We argue here that the arguments supporting its merits rest on a critical assumption. The assumption is that students possess the cognitive skills that enable them to engage in these activities in a way that is profitable with respect to the objectives identified previously. If students lack the necessary skills, inquiry learning could in fact be counterproductive, leading students to frustration and to the conclusion that the world, in fact, is not analyzable and worth trying to understand—a conclusion that runs exactly opposite to the intellectual values that Resnick and Nelson-LeGall (1997) argued inquiry learning should promote.

At this point, it is necessary to become specific as to what we are referring to as inquiry learning because a wide range of educational practices have been described under this heading. Here, we define *inquiry learning* as an educational activity in which students individually or collectively investigate a set of

phenomena—virtual or real—and draw conclusions about it. Students direct their own investigatory activity, but they may be prompted to formulate questions, plan their activity, and draw and justify conclusions about what they have learned de Jong and van Joolingen (1998).

Inquiry activities targeted to young children may have simple goals that do not extend beyond description, classification, or measurement of familiar phenomena. More typically, however, inquiry activities are designed for older children or adolescents and have, as their goal, the identification of causes and effects. The context is typically a multivariable one, such that the goal becomes one of identifying which variable or variables are responsible for an outcome or how a change in the level of one variable causes a change in one or more other variables in the system. Equally important is the identification of noncausal variables, so that these can be eliminated as sources of influence in understanding how the system functions.

Are students of the elementary and middle school grades (in which inquiry activities are most commonly introduced) capable of inferring such relations based on investigations of a multivariable system? There exists little educational research on students engaged in inquiry learning that would answer this question directly. Evidence that is available, on the other hand, from the literature on scientific reasoning suggests significant strategic weaknesses that have implications for inquiry activity (Klahr, 2000; Klahr, Fay, & Dunbar, 1993; Kuhn, Amsel, & O'Loughlin, 1988; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Kuhn, Schauble, & Garcia-Mila, 1992; Schauble, 1990, 1996). Strategies, moreover, even though they have been the focus of attention in scientific reasoning research, may not be all, or even the most critical element, that is missing. In this article, we raise the possibility that students at the middle school level, and sometimes well beyond, have an incorrect mental model that underlies strategic weaknesses, and that impedes the multivariable analysis required in the most common forms of inquiry learning. Like many mental models, this model may be resistant to revision.

MENTAL MODELS UNDERLYING INQUIRY LEARNING

Numerous lines of cognitive and cognitively oriented educational research emphasize mental models as vehicles that students employ in coming to understand the workings of a system (Gentner & Stevens, 1983; Vosniadou & Brewer, 1992). Such models facilitate (or sometimes interfere with) understanding of how a system operates. We use the mental model terminology here, however, in a more generic sense. It is students' mental model of causality itself, we claim, that may be deficient, rather than a mental model of the workings of any particular causal system. This incorrect mental model can be contrasted to a normative analysis of variance (ANOVA) model of causality in a multivariable system—a model in which individual variables each manifest their individual effects on one or more dependent

variables. Such effects are normally additive, although one effect may in some cases influence (interact with) the effect of another variable.

If we expect students to understand the operation of a multivariable system, they must at least understand the concept of additive effects—effects that operate individually on a dependent variable but that are cumulative (additive) in their outcomes. A student who possesses this mental model of additive effects can understand much about a system and in many cases even predict outcomes fairly accurately without the more sophisticated concept of interaction effects as part of this model. The deficient mental model we describe here, in contrast, is one in which neither additive nor interactive effects are understood in a normative way.

The specific situation we refer to here in considering these mental models is one in which an outcome variable that can assume multiple levels on at least an ordinal scale (i.e., ordered from less to more of some quantity) is potentially affected by a set of independent variables, each of which can assume two different levels. For example, in the work described here, the variables of soil type (sand vs. clay), elevation (high vs. low), and water pollution (high vs. low) are among five potential features affecting the amount of flooding at building sites along a lake. This outcome variable can assume five different levels, from low flooding (1 ft) to high (5 ft). To investigate the system, a student has the opportunity to choose desired levels for each of the features and, once this is done, to observe the resulting outcome. The task presented to the student is to find out which features make a difference and which do not make a difference in determining the level of the outcome variable.

Students beginning to investigate such a system often focus exclusively on outcomes—achieving those deemed desirable and avoiding undesirable outcomes (Kuhn et al., 1995; Kuhn et al., 1992; Schauble, 1990; Schauble, Klopfer, & Raghavan, 1991). To make progress beyond an outcome focus, it is necessary to shift one's attention to what we can call an *analysis* focus—specifically, analysis in terms of the effects of individual features. Without the understanding that individual features will contribute their respective effects to the outcomes, the system cannot be analyzed and understood.

Consider now the mental model that might characterize the thinking of sixth-grader Matt (an actual case from the database of the research described here, although the student's name is changed). We label the five variable features of the system by number and the respective levels of each feature by the letters *a* or *b*.

Matt makes the following claims. Based on observation of the instance 1a2a3a4a5a in conjunction with a positive outcome (O1), Matt concludes that all of these contributed to the good outcome (the site is minimally flooded). In other words, the sandy soil, the lack of pollution, the high elevation, and so forth, "all make a difference, because it came out good." Next, Matt examines the instance 1b2b3b4a5a (i.e., the levels of three of the features are changed from what they were in the first instance and remain the same for the other two features) and observes a poor outcome (high flooding). This time, Matt says, "None of them made

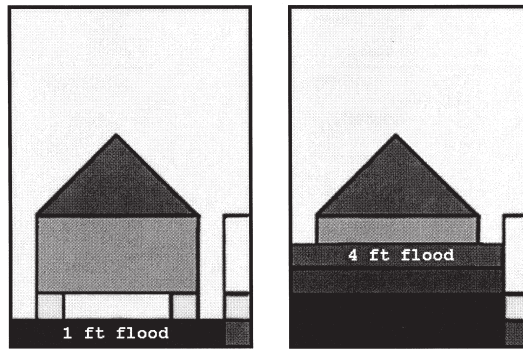


FIGURE 1 The co-occurrence mental model. Both features are implicated as causal in the outcome on the left and not implicated in the outcome on the right.

a difference—it came out bad." We can infer from these statements that Matt is not using the expression *make a difference* in the normative way dictated by the analysis model. Instead, *making a difference* appears to mean "helping to produce a good outcome."

Such a model of multivariable causality accommodates the seeming paradox of a variable *making a difference* on some occasions (when the outcome is good) and *not making a difference* on others (when the outcome is poor)—a state of affairs that we in fact have found to be common among, and not at all paradoxical, for many children of this age. In earlier work, for example, children of Matt's age who observed that sports balls with a certain type of surface produce a good serve half of the time and a poor serve half of the time, whereas balls with a different surface type produce the same results, often failed to make the normative inference that type of surface was noncausal with respect to this outcome variable. Instead, they concluded that the surface type "sometimes makes a difference" in the quality of the serve (Kuhn et al., 1988).

Formalizing this mental model, it can be described as stipulating the co-occurrence of a particular variable level and an outcome as a sufficient condition for implicating that variable as having played a role in the outcome (or, in the case of a negative outcome, excluding the variable as having played a role). We refer to this mental model as a *co-occurrence* model.

It is important to note that the variable level, not the variable itself, is implicated as causal in the co-occurrence model. In the depiction in Figure 1, for example, 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 on the left-hand side of the figure. In interpreting the unsuccessful outcome

on the right, the same water temperature is, this time, judged not to make a difference. Reflecting another form of inconsistency, rather than soil type making a difference, sand does (but clay does not) make a difference.

Causal attributions, then, fluctuate as functions of the particular constellation of feature levels that are present in a particular instance. Each constellation is a unique event (even though its components may be incompletely identified). Rather than representing a genuine interactive model, however, the co-occurrence model reflects failure to conceptualize even the main effects (of features as variables) on which statistical interaction effects are founded.

It should be noted finally that in addition to being inconsistent, effects of individual features are not additive in the co-occurrence model. Because co-occurrence of a particular feature level and an outcome is a sufficient condition for attributing causality, any co-occurring feature level may be implicated in what is regarded as a successful outcome. Implication of more co-occurring feature levels might be expected to produce an even more successful outcome—yet even one co-occurring feature is sufficient to explain even the most successful outcome.

MENTAL MODELS OF CAUSALITY AND INVESTIGATIVE STRATEGIES

Mental models, as noted previously, may be resistant to change, and it is not clear what the most effective way might be to effect a transition from a co-occurrence to a genuine analysis model of multivariable causality. In previous research (Kuhn et al., 1995; Kuhn et al., 1992), we focused on the investigatory strategies students use and the resulting validity of their inferences. To make a valid inference, it is necessary to make a controlled comparison between two instances that differ only with respect to a single feature that is the focus of analysis. In research on scientific reasoning, the lion's share of attention has gone to this controlled comparison, or "all other things equal" investigation strategy, as the hallmark of skilled scientific reasoning (DeLoache, Miller, & Pierroutsakos, 1998; Klahr, 2000; Kuhn et al., 1988; Zimmerman, 2000). The investigator needs to recognize that to conduct a sound test of the effect of one variable, all other variables must be held constant, so that the effects of these other variables do not influence the outcome.

In our research (Kuhn et al., 1995; Kuhn et al., 1992), we have found that use of a controlled comparison strategy and the valid inferences that result from it increase in frequency over a period of months among preadolescents when they are given the opportunity to engage in self-directed investigatory activity of a multivariable system. Some students, however, even after many weeks of investigation, remain stubbornly fixed at a level of confounded investigations and fallacious inferences. The mental model ideas proposed here suggest a possible reason for their lack of progress.

The analysis model of additive effects of individual variables is a logical prerequisite to the controlled comparison investigative strategy. This is so because the purpose of the latter is identification of the effect of a single variable. If one's mental model is not one of individual additive effects, neither attribute of the controlled comparison strategy is compelling. The "comparison" attribute is not compelling, given that it entails comparing the outcomes associated with two (or more) levels of a variable for the purpose of assessing the effect of that variable on outcome. Furthermore, the "controlled" attribute is even less compelling because it is the individual effects of other variables that need to be controlled. As we suggested previously, then, an incorrect mental model may underlie the strategic weaknesses that have been observed and impede the multivariable analysis central to inquiry learning.

As a procedure, the controlled comparison strategy is straightforward to teach ("Keep everything else the same and just change one thing"). By comparison, it is not easy to change mental models, and this would seem particularly so of the sort of generic model (of multivariable causality) that we discuss here. A number of studies over the years have undertaken teaching the use of the controlled comparison procedure in brief training sessions (Case, 1974; Chen & Klahr, 1999) with some degree of success, but such interventions are unlikely to effect change in underlying mental models of causality.

In our research (Kuhn & Angelev, 1976; Kuhn, & Ho, 1980; Kuhn, Ho, & Adams, 1979; Kuhn & Phelps, 1982; Kuhn et al., 1988; Kuhn et al., 1995; Kuhn et al., 1992), we have focused on longer term interventions (typically 8–10 weekly sessions), with an objective of promoting not just change in the strategies students use to acquire new knowledge about a causal system (referred to later as *knowing strategies*), but enhancement of their metastrategic understanding of why these are the strategies that must be used and why others will not suffice. Execution of the controlled comparison strategy, as just noted, is relatively easy to teach, but it is metastrategic understanding that determines whether the strategy will be selected when the student is engaged in self-directed activity (Kuhn, in press-c).

The argument we make here is that this metastrategic understanding requires a correct mental model of how a multivariable causal system (again, in the generic sense of any causal system) operates. A strategy that has the purpose of assessing the effect of an individual feature will not be understood and valued unless one's mental model of the operation of a multivariable system is based on the additive effects of individual features. Once this analysis mental model of individual additive effects is attained, the learner is in a position to proceed to a more complex analysis model in which these individual effects are interactive in their influence on outcomes. In the absence of this analysis mental model in which individual variables assert their respective effects on an outcome in an additive manner, the controlled comparison strategy for assessing these effects can be taught, but its logic will not be compelling—there will not be a deep level of understanding as to why it must be used.

METALEVEL FUNCTIONING

One way to formalize this deep level of understanding as a construct is to postulate a metalevel of operation that is distinct from the performance level (Figure 2). The knowing strategies depicted in Figure 2 are those we regard as central to inquiry activity. The metalevel is the level at which particular knowing strategies are selected for use and their application monitored and the results interpreted (left-hand side of Figure 2). Understanding why to use a strategy, then, occurs at the metalevel. Moreover, it is this metalevel understanding that should govern not only the use of a strategy but its generalization to a new context in which it is applicable (Crowley & Siegler, 1999).

Metalevel understanding, we can hypothesize, develops in parallel with strategic competence in a mutually facilitative relation. Exercise of strategies at the performance level feeds back and enhances the metalevel understanding that will guide subsequent strategy selection and, hence, performance. In other words, metalevel understanding both informs and is informed by strategic performance (Figure 2; see also Sophian, 1997).

Strategies exist only in relation to goals or objectives. Therefore, metalevel understanding of task objectives (metatask understanding) is as critical as metastrategic understanding of the strategies that are available to apply to the task (Kuhn & Pearsall, 1998; Siegler & Crowley, 1994). Both must be present and co-

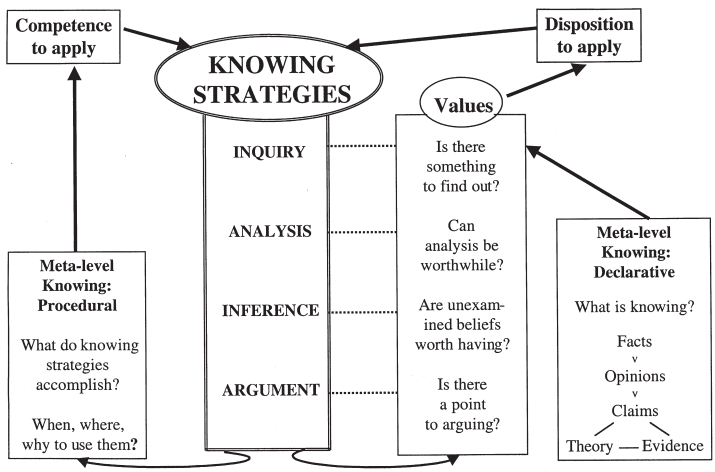


FIGURE 2 Phases of inquiry activity, with hypothesized bidirectional relations between the metalevel and the performance level.
Note. From "How Do People Know?" by D. Kuhn, in press, *Psychological Science.* Copyright 1999 by Blackwell. Reprinted with permission.

ordinated to guide performance successfully. The mental model of additive effects of individual variables, we have claimed, is essential for the controlled comparison investigative strategy. We can now elaborate that specifically it is necessary to metatask understanding of the task objective of identifying effects of individual variables. Without this understanding, the appropriate controlled comparison strategy will not be consistently selected.

In the research presented in this article, we examine the extent to which the mental model transition (from an incorrect to correct model of multivariable causality) that is discussed here is facilitated by metalevel exercise that occurs in addition to and in conjunction with performance-level exercise of strategies. In past work, we have undertaken to promote the development of metalevel understanding by externalizing it in collaborative discussion among peers, a method that works under certain conditions (Kuhn, in press-c). Another method is to engage students more directly in metalevel exercise by asking them to evaluate different potential strategies that could be applied to a problem. The contemplation of alternative strategies should promote not only attention to task objectives but also the essential task of coordinating task objectives with available strategies. This direct approach, we have found, also meets with some success (Pearsall, 1999).

It is this latter approach that is used in the work presented here, but we do so with a particular focus on the question of whether it will promote the transition to the more correct additive mental model of causality. As part of the metastrategic evaluation exercise, students are presented the situation of two individuals who disagree as to the effect of a particular feature with one individual, for example, claiming that soil type makes a difference and the other claiming that it does not. The students must then consider and evaluate the strategies that could be used to resolve the conflict. Note that the conflict is explicitly identified as one about the effect of a particular, individual feature. To what extent, we asked, would extended experience with the evaluation of such conflicts promote (a) at the metalevel, a mental model based on the effects of individual features, reflected in metatask understanding that the object of the activity is identification of effects of individual features; (b) metastrategic understanding of the need to control the influences of other features (the controlled comparison strategy); (c) at the performance level, successful use of the controlled comparison strategy; (d) resulting valid inferences regarding the status of causal and noncausal features in the system; and (e) superior acquisition of knowledge about the system, reflected in correct conclusions about its causal structure. Our past research indicated that performance-level exercise of investigative activity (with no feedback beyond that provided by the student's own activity) over a period of weeks is sufficient to induce some change on at least some of these dimensions among a majority of students. We, therefore, compare two conditions: one in which students engage only in this performance-level exercise and another in which students also engage in the metalevel exercise, described more fully subsequently.

METHOD

Participants

Participants were 42 middle school (6th, 7th, and 8th grade) students attending an urban public school. They came from two comparable intact science classes of mixed-grade (6th–8th) level. Each class participated over the same several-month period as part of their science curriculum. One class was arbitrarily chosen to serve as an experimental group, and the other class served as a control group. The former group consisted of 10 boys and 11 girls, and the latter had 12 boys and 9 girls. Students were of diverse ethnicity, with the majority being African American or Hispanic.

Task Environment

The main task, which students engaged repeatedly both individually and in dyads during the course of the study, is a multimedia research program, created with the Macromedia Director authoring tool. The program supports self-directed investigation of a multivariable environment consisting of a set of instances available for investigation, with instances defined by five variable features and an outcome—the degree of flooding of a building site.

Students are placed in the role of builders working for TC Construction Company, which builds cabins along the shore of a series of small lakes. The area is susceptible to flooding, and the cabins are, therefore, built on supports that raise them above the ground. It is the student's task to identify the optimum height of the supports for various buildings. It is explained in the introductory online presentation that the supports should be neither higher than necessary to avoid unnecessary building expense nor lower than necessary to avoid flooding and resulting damage to the building. Students are given a bank account at the beginning of their work, with money subtracted for incorrect predictions (of how much flooding will occur at that site and, therefore, how high the supports need to be built) and a bonus received for correct predictions.

The only way for students to generate correct predictions is to investigate effects of the five variable features on amount of flooding and draw appropriate inferences. Following an introductory session in which the program is introduced and students' initial beliefs assessed regarding the five variable features that may influence flooding, the student embarks on a series of investigatory sessions. The program includes the following sequence of activities: statement of investigatory intent (students indicate which features they intend to find out about), selection of feature levels in instances to be examined, prediction of outcomes, the opportunity to make inferences and justify them, and the option of making notes in an online notebook. During the second and subsequent sequences, the feature levels and out-

TABLE 1
Causal Structure of Flood Problem

Water pollution (high or low)	No effect
Water temperature (hot or cold)	Cold raises the flood level 1 ft
Soil depth (deep or shallow)	Shallow raises the flood level 2 ft
Soil type (clay or sand)	Sand reduces the flood level 1 ft for deep soil only
Elevation (high or low)	No effect

come of the immediately preceding instance remain visible to facilitate comparisons. The sequence is repeated five times during each session. At the end of a session, students are asked to draw conclusions about the causal and noncausal effects operating in the system. Students' activity within the program is tracked and recorded into word processing files by the program.

The causal structure of the task environment is shown in Table 1. Two of the five features are noncausal (i.e., have no effect on outcome). The other three features are causal, with an interactive effect between two features.

A second task was employed as a transfer task, to assess the generality of changes in students' strategies and understanding as a function of their work on the main task. The transfer task was identical to the main task in structure and computer interface. The content involved the effects of various features on job applicants' potential effectiveness as a teacher's aide in a classroom.

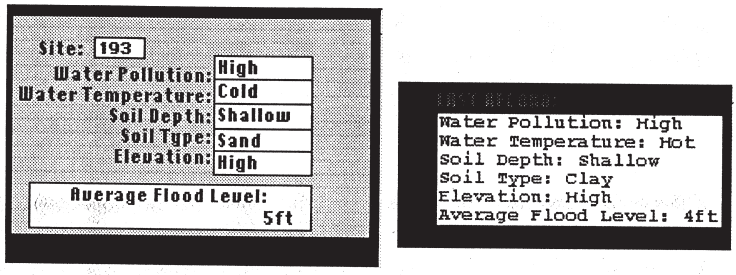
Procedure

Pretest assessment. Students from both classes participated in individually administered pretests. Following introduction of the program and assessment of initial beliefs, the initial investigatory session took place. During that session, the student repeated the investigatory cycle (selections of feature levels, prediction of outcome, inference, and justification) five times. Students worked one-on-one with a researcher during that session, so that any questions or misunderstandings could be addressed. An identical pretest assessment was administered for the transfer (teacher aide) task.

Performance-level exercise. A 2-week school vacation intervened between completion of pretest assessments and commencement of the main phase of the study. During that phase, participants worked in changing dyads in a series of 9 to 10 sessions that took place over a period of roughly 6 weeks, with an average of two sessions per week (and a range of 1–3, due to absences and scheduling constraints). Assignment to dyads was random except for the constraint of avoiding, as far as possible, pairing of the same two students for more than one session. At the beginning of the pair sessions, students were instructed to work collaboratively rather than in turns to discuss their views as to how to proceed or what to conclude,

TABLE 2
Sample Metalevel Exercise

This is Terry and Jamie's work:



They haven't finished because they can't agree.
Terry says soil type does make a difference.
Jamie says soil type does not make a difference.
What can settle the argument between them?

_____
_____

_____

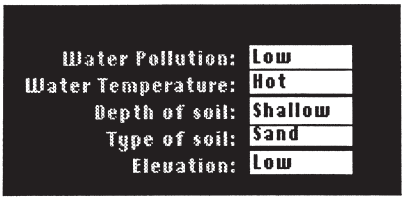
Do the records they looked at say anything about whether soil type does or does not make a difference? (circle one)
Yes No
What do the records suggest?
Soil type makes a difference
Soil type does not make a difference
Can't tell
What was different about this record and the last record they looked at?
Were they different on soil type? Same Different
Were they different on water pollution? Same Different
Were they different on water temperature? Same Different
Were they different on soil depth? Same Different
Were they different on elevation? Same Different
Did the two records have different amounts of flooding? (circle one or more)
Because of soil type
Because of water pollution
Because of water temperature
Because of soil depth
Because of elevation
Can't tell
Do the records they looked at say anything about whether soil type does or does not make a difference? (circle one)
Yes No
What do the records suggest?
Soil type makes a difference
Soil type does not make a difference
Can't tell

(*continued*)

TABLE 2 (Continued)

What grade would you give Terry and Jamie on their work? (circle one)
A B C D F
Why do they deserve this grade?

Suppose Terry and Jamie looked at this record:



And they wanted to find out FOR SURE if soil type makes a difference.
What record should they look at next, to be sure? (Circle your choices.)

If the second record comes out different from the first, what will the reason be?

and not to proceed until some agreement was reached. At each session, the pair worked collaboratively on the flood task, with an adult available for consultation if problems arose, but the adult otherwise did not intervene.

Metalevel exercise. In addition, students in the experimental condition engaged in a series of paper-and-pencil exercises related to the flood task, which they worked on in pairs within the classroom, twice each week for the duration of the period that they were working on the flood program. Pairing varied across occasions, and students were instructed to work together and agree on an answer before writing it down. Students completed one exercise per session. A sample exercise is shown in Table 2. In that example, the comparison is confounded (the record shown differs from the previous record with respect to two features) and the outcome varies. In other cases, the comparison was controlled and the outcomes either varied or remained constant.

Posttest assessment. The posttest assessment was conducted individually and duplicated the pretest assessment. Posttest assessments took place during the 2 weeks following completion of the intervention period.

Delayed posttest assessment of metalevel understanding. Approximately 1 week following the completion of posttest assessments, a paper-and-pencil measure was administered during class time by the classroom teacher in each of the classes. The researchers were not present during this administration. One student in the experimental condition and 5 students in the control condition were absent on the administration day and did not receive this assessment.

This measure was designed to assess metatask understanding of the task goal (identifying effects of individual features) and metastrategic understanding of the critical strategy (controlled comparison) that allowed this goal to be met. To serve as the most rigorous test of understanding, this measure was based on the content of the transfer (teacher aide) task rather than on the content of the main task (used in the intervention activities). The student was asked which of two records would be the better one to look at next: Pat's choice (which represented a controlled comparison relative to the initial record available) or Lee's choice (which represented a confounded comparison with respect to two features). The student was asked to justify why this was "a better plan for finding out." In addition, the student was asked what each person (Pat and Lee) will find out with the plan they have chosen.

RESULTS

Performance

Prediction error. A quantitative measure of performance is the degree of error in predicting outcomes. Average prediction error decreased from 1.23 errors at the pretest to 0.96 errors at the posttest (with one unit of error equaling a mismatch of 1 ft. between the predicted level of flooding and the actual level). This decline was significant, $F(1, 40) = 4.54$, $p = .039$, and did not differ by experimental condition.

Mean prediction error on the transfer task similarly decreased from 1.05 at the pretest to 0.74 at the posttest. This difference was also significant, $F(1, 40) = 7.87$, $p = .008$, and did not differ by experimental condition. Thus, students in both groups learned something about the causal system that was observable in their performance.

Valid inference. A more qualitative picture of performance is provided by analysis of the strategies students applied to the task. The key investigative strategy

of controlled comparison is not straightforward to assess because students did not always make the appropriate comparisons, even when they had selected for examination data that would allow them to make an informative comparison. Therefore, we were conservative in assessment of use of the controlled comparison strategy, judging it present only when students drew a justified inference, that is, drew a correct conclusion based on comparison of two instances that they had generated and that they referred to in justifying the conclusion.

The number of inferences justified by an appropriate controlled comparison of two instances (henceforth called *valid inferences*) was examined relative to number of possible inferences. This proportion of valid inferences was calculated for each student for the main and transfer tasks at pre- and posttest assessments. As seen in Table 3, patterns are similar for the two tasks. Students in both conditions show a low level of valid inference at the pretest, and both groups show improvement from pretest to posttest, with the experimental group showing somewhat greater improvement than the control group. The proportions summarized in Table 3 were subjected to arcsine transformation and analyzed by a repeated measures ANOVA with time of testing a within-subjects factor and experimental condition a between-subject factor. For the main task, time of testing was significant, $F(1, 41)$

TABLE 3
Proportion of Valid Inferences

Group	*Pretest*	*Posttest*
Main task		
Experimental group[a]		
M	.06	.45
SD	.11	.42
Control group[a]		
M	.12	.33
SD	.19	.42
Total group[b]		
M	.09	.39
SD	.15	.42
Transfer task		
Experimental group[a]		
M	.00	.43
SD	.00	.51
Control group[a]		
M	.10	.29
SD	.30	.46
Total group[b]		
M	.05	.36
SD	.26	.48

[a]$N = 21$; [b]$N = 42$.

TABLE 4
Mean Number of Inferences per Instance Examined

Group	*Pretest*	*Posttest*
Main task		
Experimental group[a]		
M	3.77	3.33
SD	1.32	1.66
Control group[a]		
M	3.98	4.02
SD	1.11	1.32
Total group[b]		
M	3.88	3.67
SD	1.21	1.51
Transfer task		
Experimental group[a]		
M	3.86	2.64
SD	1.57	1.92
Control group[a]		
M	3.64	3.50
SD	1.57	1.86
Total group[b]		
M	3.75	3.07
SD	1.55	1.92

[a]$N = 21$; [b]$N = 42$.

= 20.58, $p < .001$, but neither condition nor the interaction of time and condition reached significance. For the transfer task, time of testing was significant, $F(1, 41) = 19.21$, $p < .001$, and the Time × Condition interaction was marginally significant, $F(1, 41) = 2.84$, $p = .10$.

A decline in the number of inferences made also reflects improved performance. A student who declines to make an inference (choosing the "haven't found out" option) recognizes that the evidence he or she has generated does not allow for a definitive conclusion. The average number of inferences made per session was overall slightly below four (of a possible five). As seen in Table 4, this number declined noticeably only among the experimental group and more so on the transfer task than the main task. A repeated measures ANOVA yielded no significant effects for the main task. On the transfer task, however, the interaction effects of both time, $F(1, 41) = 7.01$, $p = .01$, and Time × Condition, $F(1, 41) = 4.37$, $p = .04$, were significant.

Some additional insight is gained by qualitative examination of patterns of change from pretest to posttest. These are summarized in Table 5, which shows the distribution of students showing no valid inference, a mixture of valid and invalid inference, and all valid inference at the two times for the main task. As seen in Ta-

ble 5, the majority of students show no valid inference at the pretest, and just less than half do not improve in this respect. Improvement, however, is more frequent in the experimental group. Mixture of valid and invalid inference is a common pattern at both times, consistent with previous research (Chen & Klahr, 1999; Crowley & Siegler, 1999; Kuhn et al., 1995). Results for the transfer task are similar, with slightly lower frequencies of valid inference usage at the posttest (9 students in the experimental group and 6 in the control group showing some or all valid inferences).

Understanding

Understanding inferred from performance. An indirect measure of students' understanding of the task objective is provided by their responses to the query regarding which features they intended to find out about, posed at the beginning of each investigative sequence. Did students understand the need to focus their investigative efforts on a single feature at a time? If so, this understanding should be reflected in answers to this question. A decline in the number of features for which a student expressed an intent (to investigate) in examining a single instance of evidence should reflect increased understanding of the need to focus on single features. Therefore, we compared mean number of intents (to investigate a feature) per instance at pretest and posttest assessments.

These means are shown in Table 6 for the two conditions and times of testing. As seen there, despite differences attributable to chance at pretest, number of intents declines over time, with the most sizable decline in the experimental group on the main task. An ANOVA yielded significant effects for the main task for both time, $F(1, 41) = 60.94$, $p < .001$, and the Time × Condition interaction, $F(1, 41) = 6.75$, $p = .013$. For the transfer task, only the effect for time was significant, $F(1, 41) = 5.92$, $p = .02$. Also relevant are the number of students for whom the mean number of intents declined to less than 2, indicating that at least some of the time this student had the intent of investigating a single feature. At the posttest, these

TABLE 5
Pre- and Posttest Distributions of Participants by Patterns of Valid Inferences (Main Task)

Group	*No Valid Inferences*	*Some Valid Inferences*	*All Valid Inferences*
Experimental group			
Pretest	16	5	0
Posttest	8	7	6
Control group			
Pretest	14	7	0
Postest	12	4	5

TABLE 6
Mean Number of Intents per Instance Examined

Group	*Pretest*	*Posttest*
Main task		
Experimental group[a]		
M	3.74	2.00
SD	0.87	0.72
Control group[a]		
M	3.05	2.24
SD	0.89	1.08
Total group[b]		
M	3.40	2.12
SD	0.94	0.91
Transfer task		
Experimental group[a]		
M	3.10	2.45
SD	1.5	1.41
Control group[a]		
M	2.57	2.02
SD	1.12	1.16
Total group[b]		
M	2.83	2.24
SD	1.33	1.29

[a]$N = 21$; [b]$N = 42$.

frequencies were 15 (71% of participants) for the experimental group and 12 (57%) for the control group on the main task. This difference was not maintained in the transfer task, however. Frequencies were 11 (52%) and 14 (67%) for experimental and control groups, respectively.

Relation of understanding to strategies. Qualitative analysis of patterns of performance indicated that focus on a single feature at a time as an investigatory intent at the posttest was associated with better strategies at the performance level. Of 10 participants (6 experimental and 4 control), who showed consistent single-feature investigatory intent at the posttest, all displayed valid inference at the posttest. Of the 6 experimental and 12 control participants who displayed no valid inference, conversely, none displayed single-feature investigatory intent. These students either intended to investigate multiple features at once, shifted their intent from one feature to another before the necessary evidence had been generated with respect to the first feature, or expressed no investigatory intent ("didn't know" what they were going to find out).

Direct assessment of understanding. The metalevel assessment measure was designed to provide a direct measure of what participants understood at the final assessment with respect to (a) the objective of the task and (b) why controlled comparison was the best strategy for achieving that objective. Both of these were assessed in a content domain other than the one in which students had had exercise.

Students who scored at the highest level (Level 3) chose Pat's plan (which allows unconfounded comparison) as the better one and were able to answer both questions about Pat's plan correctly—why it is better than Lee's plan (metastrategic understanding) and what Pat is intending to find out (metatask understanding). Typical of the correct answers to the first question were "because he only changed one thing" or "because everything is the same except age," although a few students showed very clear metastrategic understanding reflected in an answer such as "If you change only one and it makes a difference then you know what made the change." Typical of the correct answers to the second question were "if age makes a difference" or "if an older or younger teacher aide is better."

Students categorized as Level 2 chose Pat's plan as the better one but answered only one of the questions correctly, responding "I don't know" to the other or giving a vague answer (e.g., "She'll find out if her plan is better than Lee's").

Students categorized as Level 1 chose Pat's plan as the better one but offered no relevant justification (e.g., "Pat's plan is better because being a parent means she knows how to take care of her students").

Table 7 shows the number of students in each group categorized at each level. All students in the experimental group, it is seen, recognized Pat's plan as better, and all but 2 students in the control group did so. The number of students who were able to justify the superiority of Pat's plan in meeting the task objectives, however, is significantly higher among students in the experimental group—55% versus 38%, $\chi^2(1, N = 36) = 7.60$, $p < .01$. These results suggest that (a) overall, students' implicit understanding (reflected in the correct choice of Pat's plan) outstrips their explicit understanding (reflected in their justifications of the choice); and (b) the experimental condition facilitates the development of metalevel understanding.

Results also indicate that metastrategic understanding may remain incomplete even among students who show considerable understanding by correctly answering the two questions described. In response to the question "What will Lee find

TABLE 7
Number of Students at Each Level of Performance on the Metalevel Assessment Measure

Group	*Level 3*	*Level 2*	*Level 1*	*Level 0*
Experimental group	11	1	8	0
Control group	6	3	5	2

out with Lee's plan?" only 4 students in the experimental group and 5 students in the control group answered correctly, typically by identifying the limitation of the noncontrolled comparison strategy (e.g., "She won't find out anything because she won't know what caused the change"). The less common answer "He'll find out if anything matters" was also counted as correct. Others, when asked about Lee's plan, not only did not acknowledge its inferiority (e.g., "She will find out the same") but also indicated potentially productive outcomes of the plan (e.g., "She'll find out if the totally opposite person will make a difference"). The latter response, we would claim, invokes the faulty co-occurrence mental model of analysis via feature levels, rather than features.

Knowledge

Two measures of the posttest knowledge that students exhibited about the system following their investigations were examined. One was the total number of features they implicated as causal in interpreting outcomes. The other was the correctness of their conclusions as to which of the features were causal and which were noncausal.

At the pretest for the main task, students implicated a mean of 2.69 features as having causal status (compared to the correct number of 3). Following their investigations with the flood program, the mean number of features implicated declined to 2.22, a significant decrease, $F(1, 39) = 4.68$, $p = .037$. (This decrease did not differ significantly across conditions.) In this respect, then, students became less correct following investigation.

However, this conclusion must be tempered by the knowledge that students displayed as to which features were causal and which were noncausal. These findings are examined only for the main task. (Students' knowledge would not be expected to increase appreciably in the transfer task, given their limited exposure to it.) With respect to both noncausal features (water pollution and elevation), there was increase from pretest to posttest in the number of correct conclusions, indicating improved knowledge about the causal system. Many students, however, maintained their incorrect beliefs that these features had causal status. For water pollution, number of students exhibiting correct conclusions increased from 10 to 26 (of a total group of 42). For elevation, the number increased from 12 to 18. With respect to the causal feature water temperature, correct conclusions regarding the direction and nature of its causal status increased from 8 at the pretest to 18 at the posttest. (The remainder most commonly judged the feature noncausal, although a few students judged it causal but in the incorrect direction, or chose an "it depends" option.) Similarly, correct conclusions regarding the soil type feature increased from 9 at the pretest to 19 at the posttest, with most of the remaining students judging the feature noncausal, but 1 student correctly stipulated an interaction effect with soil depth. Soil type was initially (and correctly) judged causal by the largest number

of students—23. This number increased to 33 at the posttest, with a few students nonetheless retaining incorrect beliefs. Thus, students' interaction with the program over time enabled both groups to increase their knowledge of the causal system. The retention of incorrect beliefs, despite the substantial amount of evidence each participant generated, however, was common and did not differ significantly across conditions.

DISCUSSION

Increasingly, "authentic" scientific activity is being promoted as a model of good science education (Bransford et al., 1999; Cavalli-Sforza, Weiner, & Lesgold, 1994; Eisenhart, Finkel, & Marion, 1996; McGinn & Roth, 1999; Palincsar & Magnusson, in press). Such activity is contrasted to the allegedly more superficial observation, description, and laboratory exercises with well-known outcomes that long have been the staple of even the best science education. Students must engage in the genuine inquiry, it is argued—involving the formulation of questions, design of investigations, and coordination of theory and evidence with respect to multivariable systems—that is characteristic of real science.

The data presented here suggest that the skills required to engage effectively in typical forms of inquiry learning cannot be assumed to be in place by early adolescence. If students are to investigate, analyze, and accurately represent a multivariable system, they must be able to conceptualize multiple variables additively coacting on an outcome. Our results indicate that many young adolescents find a model of multivariable causality challenging. Correspondingly, the strategies they exhibit for accessing, examining, and interpreting evidence pertinent to such a model are far from optimal. We turn later to curriculum implications that we believe follow from these findings and consider first what the results suggest regarding the nature of these cognitive competencies and how they develop.

What Develops?

The performance skills (notably the controlled comparison strategy) that have been the focus of attention in research on scientific reasoning arguably are only one piece of a complex structure of related skills that undergoes development. This structure needs to be defined both horizontally (with respect to the components it includes) and vertically (with respect to first its emergent and ultimately its consolidated forms). An attempt to depict the horizontal structure appears in Figure 2, presented earlier. Key components of this model are (a) the full cycle of inquiry activity, beginning with the critical skill of identifying the questions to be asked and culminating in the advancement of claims in argumentive discourse; (b) the metalevel of *un-*

derstanding (of both strategies, depicted on the left side of Figure 2, and knowledge, depicted on the right side) that both directs and is influenced by performance, as discussed earlier; and (c) *values* associated with inquiry activity, highlighted by Resnick and Nelson-LeGall (1997) and discussed earlier. Related to values and also represented on the right side of Figure 2 is metalevel epistemological understanding of the nature of one's own and other's knowledge and knowing (Kuhn, Cheney, & Weinstock, in press). The broad implication to be drawn from Figure 2 is that there is more to effective knowing than the performance skills themselves (Kuhn, in press-b).

Vertical specification refers to the fact that a complex structure of this sort does not emerge fully formed but, more likely, undergoes a gradual evolution. Research with young elementary school children (Lehrer & Schauble, in press) has made it clear that even very basic forms of organizing and representing data (such as the frequencies of a set of possible outcomes) pose challenges to young children, and the relevant understandings and skills must be painstakingly constructed. In this sense, the finding highlighted in this work—that slightly older children have difficulty in representing relations between multiple antecedent variables and multiple outcomes—should not be surprising. At the other end of the vertical continuum, it is relevant to note that in earlier research (Kuhn et al., 1995), adult community college students who were readily able to use the controlled comparison strategy to identify effects of individual features nonetheless often had trouble explaining outcomes that were the additive product of two individual effects and fluctuated from one feature to the other in accounting for the outcome, seeing it as their task to explain which single feature had produced the outcome. Recognizing their simultaneous additive influence was a conceptual hurdle that rivaled in difficulty the conceptual hurdle posed by interaction effects. Unrepresented in the inquiry activity in which students engaged in this work is the further conceptual challenge that is posed when outcomes are not deterministic (as they were in our activity) but rather are a probabilistic distribution around some central tendency. Students of any age will not be successful in understanding interactive influences on probabilistic outcomes until they have mastered the more elementary model on which we focus here, involving multiple effects additively acting on an outcome.

Mental Models

Mental models of any sort remain essentially unobservable theoretical constructs. Performance indicators of various types serve as evidence that a particular mental model is in operation, but no empirical data can indicate with certainty the operation of a particular mental model. In inquiry activities, mental models are the individual's representation of the (virtual or actual) reality that is being investigated. For this reason, they are likely to influence the strategies that are brought to bear on

the task. Nonetheless, we cannot say with certainty that it was revision in mental models that brought about the changes observed over time in this work. Such revision could be an effect rather than a cause. Nor would we want to claim that the kind of intervention undertaken in this work represents the only sound approach to facilitating development of the cognitive competencies we have identified as involved in inquiry learning. However, this intervention was targeted at the metalevel of cognition depicted in Figure 2, and we do want to claim that this level of understanding about inquiry, in contrast to the "understanding how" emphasized in performance-focused interventions, plays an essential role in effecting change. Metalevel understanding can come about as a product of the exercise of performance skills, as well as by direct targeting, but it cannot be bypassed.

The importance of this metalevel of understanding about inquiry is also underscored by the fact that in most of the knowledge seeking that students may engage in outside of a formal school setting, they are unlikely to have the opportunity to devise and execute controlled experiments. Much more often, they will be in a position of interpreting evidence derived from partially controlled or natural experiment data (Kuhn & Brannock, 1977). It is all the more important, then, that their interpretations not be compromised by an inadequate mental representation of the multivariable causality that such data are likely to reflect. Equally critical is metalevel understanding of the strengths and weaknesses of inference strategies that may be effective, effective but inefficient, or ineffective and fallacious. Again, what to do (when controlled experimentation is possible) is only one piece of a larger knowledge structure that includes what not to do and why, as well as what to conclude when controlled experiment is not feasible—to know when we do not know, when we have a way to find out, and when we will never know (Kuhn, in press-b).

Patterns and Mechanisms of Change

The results presented here confirm earlier research (Kuhn et al., 1995; Kuhn et al., 1992; Schauble, 1990, 1996) indicating that exercise can be a sufficient condition to induce strategic change, both in increasing the frequency of effective strategies and decreasing the frequency of ineffective ones. This work extends these findings to metalevel understanding of task and strategies and the mental models of causality associated with them. In addition to performance, metalevel understanding (measured both directly and indirectly, the latter via investigatory intent) increases with exercise. This change at dual levels supports the kind of continuous feedback model depicted in Figure 2.

An additional finding of this work is that exercise directly at the metalevel (in the experimental condition) further enhances change. These benefits (indicated by significant effects of condition) were seen either specifically at the metalevel (in

both direct and indirect measures) or in the transfer to a new task at the performance level. (Condition differences, recall, did not reach significance at the performance level for the main task, though they were in the expected direction.) The social component of the exercise at both performance level and metalevel (in both cases, students worked in pairs), it should be noted, in itself provides a weak form of metalevel exercise for students in both conditions. If partners are suitably matched, students show higher levels of performance when working with a partner than they do when working alone on the same task (Kuhn, in press-c). The externalization of metalevel decisions in social dialogue presumably supports this normally covert level of processing. We did not make this comparison (between social and solitary conditions) in this study, however, because we wished to identify the effect of direct metalevel exercise.

More specific than this general model of dual-level change are the particular metalevel understandings and performance-level strategies that were the object of the present research. Although understanding of task objectives is critical to performance of most cognitive tasks (Kuhn & Pearsall, 1998; Schauble et al., 1991; Siegler & Crowley, 1994), in this case we have argued specifically that metalevel understanding of the task objective of identifying the effects of individual features (a) requires a correct mental model of multivariable causality and (b) is a prerequisite for consistent choice of the controlled comparison strategy. Logically, the value and power of the controlled comparison strategy cannot be appreciated in the absence of this mental model. Empirically, our data support this claim. Progress in understanding the task objective as one of the identifying effects of each of the individual features (which we took as an index of an accurate mental model of multivariable causality) showed significant effects of both time and experimental condition and, in analyses of individual patterns, was associated with good strategy usage. An implication for research on scientific reasoning is that investigatory intent is at least as important as the controlled comparison strategy as a topic of study.

In examining individual students' patterns of performance, we found mixture (of levels) and gradual change to be the rule rather than the exception, consistent with the findings of microgenetic research (Kuhn, 1995; Siegler & Crowley, 1991). Because students worked with a changing set of partners who produced a collaborative performance, these results do not allow microgenetic analysis of individual change patterns. Also, it is not obvious exactly what the parallel of strategy mixture might be in the domain of mental models. Students may display a confused or incoherent model in the course of transition from a less correct to a more correct model or, as they do in the case of strategies, they may rely on one approach (model) at one time and a different one at another. Our data do not allow us to choose definitively between these two alternatives, but they do suggest that a shift in mental models, like strategy shifts, is not abrupt and total, but more likely takes place slowly and in gradual steps.

The Process–Content Debate

The inquiry activity that students in this study engaged in was deliberately designed as "content-lean," in the sense that we were not undertaking to teach students any significant body of scientific knowledge. Instead, our approach was to examine in as simple a context as possible the strategies, metastrategic understanding, and attendant generic mental models required for productive inquiry regarding relations among variables. If faulty strategies and mental models are observed in this context, it is likely that they will be present as well in a more complex, content-rich environment (though they will be harder to identify and examine in that context).

A contrasting point of view is that a more content-rich context would have facilitated the reasoning observed in this study. In other domains of inferential reasoning—for example, Wason's (1983) four-card problem—performance has been shown to improve dramatically when the problem is situated in a familiar context. There is an important difference, however, between that reasoning paradigm and the one investigated here. In the former, the objective in providing a familiar context is to facilitate reasoners' recognition and, hence, application of a form of inference they already know well (e.g., permission and obligation).

This situation, in contrast, is a bit more complex because we are looking to do more than invoke a well-established reasoning scheme. The broad-level process skill in question, the coordination of theory with new evidence, can proceed in several different ways. If new evidence is entirely compatible with an existing theory, the evidence may readily be integrated into it and become part of its representation. However, this does not guarantee that this new evidence will be represented independently of the theory and brought to bear on it, which we identify as a hallmark of mature or skilled scientific thinking (Kuhn, in press-a; Kuhn & Pearsall, 2000). Instead, evidence may be integrated as an "illustration" of what is already accepted as true, or it may simply be assimilated without awareness.

The more interesting case, because it allows a clearer assessment of scientific thinking as a process, is one in which evidence conflicts with theory and, hence, is not readily assimilable, forcing the individual to ignore, dismiss, or distort it or, alternatively, to represent it accurately and evaluate its bearing on the theory. In the case in which the theoretical representation is richly elaborated and highly familiar, it is not clear that scientific thinking (again, as a process skill, in contrast to scientific understanding or knowledge) will be enhanced. Available evidence comparing scientific reasoning strategies across more and less familiar content suggests that contextually rich, highly elaborated, and highly familiar content, especially to the extent that it invokes entrenched beliefs, is motivating as a topic for contemplation but can resist the impingement of new evidence and, hence, work against proficient scientific thinking (Kuhn et al., 1995).

Implications for Science Education

An implication that should not be drawn from this research is that inquiry activity is inappropriate in the elementary or middle school science curriculum because students do not have the requisite skills to engage in it productively. The message we hope our work will convey is a different one, which is that supporting the design of inquiry curriculum for these critical years in science education should be identification of a sequence of well-delineated cognitive competencies that become the objective of this curriculum. In the absence of an explicit sequence of this nature, inquiry learning risks becoming a vacuous practice—one embraced without clear evidence of the cognitive processes or outcomes that it is likely to foster.

We believe this study makes a contribution in this respect, but such an effort is far from complete. The skills and understanding we have highlighted here lie somewhere in the middle of an extended developmental hierarchy. The kinds of elementary skills in posing questions and representing data that Lehrer and Schauble (in press) have studied form the initial levels of this hierarchy and are its essential foundation. At its upper levels are the skills and understanding needed to construct data-based models of causal systems that include multiple layers of causality and multiple variables (and variable levels) that interactively influence probabilistic outcomes. These are skills integral to the scientific inquiry that occurs in professional science.

The intervention aspect of this research similarly leaves much still to be learned. From an educational perspective, the major question is not exactly why the intervention was effective but why it was not more effective. At best, we can speculate as to what kinds of interventions might have been more effective for the sizable minority of students who showed little or no evident benefit from the experience we provided. Our work does point to (a) investigatory intent, (b) a mental model of multivariable causality, and (c) metalevel understanding as promising targets of future intervention efforts. However, more and different kinds of efforts certainly seem warranted, especially in view of the enormous current interest in inquiry as a teaching method.

A final comment has to do with the connection between scientific thinking and science education. A view emphasized in this work, and reflected in Figure 2, is that scientific thinking encompasses a good deal more than the controlled comparison strategy that has been the focus of attention in most developmental research on scientific thinking. A related view has been expressed in recent writing on science education that emphasizes the importance of formulating productive questions, representing observations in insight-generating ways, and advancing and debating claims in a framework of scientific argument (Lehrer, Carpenter, Schauble, & Putz, 2000). Mastering the coordination of questions, data representations, and argument, Lehrer et al. claimed, “puts students on the road to becoming authors of scientific knowledge” (p. 97).

Despite its comprehensiveness in encompassing all phases of scientific activity (from inquiry through argument), the kind of inquiry activity featured in this study is by itself far from a model of what a comprehensive science curriculum should be. Still, we do see such an activity as valuable as one strand interwoven into a rich middle school science curriculum. Its value as an educational tool, we believe, lies in its focusing attention on the forms of question asking and answering that are central to scientific thinking. By directing students' attention to the thinking they do in addressing scientific questions, we not only implicitly convey values and standards of science ("How do you know?"), but we also develop metalevel awareness and, ultimately, regulation of questions, of data representations, and of inferences that do—and especially that do not—follow from what is observed. Of course, we want students to acquire rich and deep understanding of the world around them as a goal of their science education, but awareness and understanding of their own and other's thinking about scientific questions seem important enough to warrant a prominent place in this curriculum.

REFERENCES

Bransford, J., Brown, A., & Cocking, R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school* (Report of the National Research Council). Washington DC: National Academy Press.

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

Cavalli-Sforza, V., Weiner, A., & Lesgold, A. (1994). Software support for students engaging in scientific activity and scientific controversy. *Science Education, 78,* 577–599.

Chan, C., Burtis, J., & Bereiter, C. (1997). Knowledge-building as a mediator of conflict in conceptual change. *Cognition and Instruction, 15,* 1–40.

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

Crowley, K., & Siegler, R. (1999). Explanation and generalization in young children's strategy learning. *Child Development, 70,* 304–316.

de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*(2), 179–201.

DeLoache, J., Miller, K., & Pierroutsakos, S. (1998). Reasoning and problem solving. In W. Damon (Series Ed.) & D. Kuhn & R. Siegler (Vol. Eds.), *Handbook of child psychology: Vol 2. Cognition, language. and perception* (5th ed., pp. 801–850). New York: Wiley.

Eisenhart, M., Finkel, E., & Marion, S. (1996). Creating the conditions for scientific literacy: A re-examination. *American Educational Research Journal, 33,* 261–295.

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

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

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

Kuhn, D. (1995). Microgenetic study of change: What has it told us? *Psychological Science, 6,* 133–139.

Kuhn, D. (in press-a). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Handbook of childhood cognitive development.* Oxford, England: Blackwell.

Kuhn, D. (in press-b). How do people know? *Psychological Science.*

Kuhn, D. (in press-c). Why development does (and doesn't) occur: Evidence from the domain of inductive reasoning. In R. Siegler & J. McClelland (Eds.), *Mechanisms of cognitive development: Neural and behavioral perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Kuhn, D., & Angelev, J. (1976). An experimental study of the development of formal operational thought. *Child Development, 47,* 697–706.

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., Cheney, R., & Weinstock, M. (in press). The development of epistemological understanding. *Cognitive Development.*

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

Kuhn, D., & Ho, V. (1980). Self-directed activity and cognitive development. *Journal of Applied Developmental Psychology, 1,* 119–133.

Kuhn, D., Ho, V., & Adams, C. (1979). Formal reasoning among pre- and late adolescents. *Child Development, 50,* 1128–1135.

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* (Vol. 17, pp. 1–44). New York: Academic.

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

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

Lehrer, R., & Schauble, L. (in press). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 5). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McGinn, M., & Roth, W. (1999). Preparing students for competent scientific practice: Implications of recent research in science and technology studies. *Educational Researcher, 28,* 14–24.

Palincsar, A., & Magnusson, S. (in press). The interplay of first-hand and second-hand investigations to model and support the development of scientific knowledge and reasoning. In S. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pearsall, S. (1999). *Effects of metacognitive exercise on the development of scientific reasoning.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Resnick, L., & Nelson-LeGall, S. (1997). Socializing intelligence. In L. Smith, J. Dockrell, & P. Tomlinson (Eds.), *Piaget, Vygotsky, and beyond* (pp. 145–158). Boston: Routledge & Kegan Paul.

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

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

Siegler, R., & Crowley, K. (1994). Constraints on learning in nonprivileged domains. *Cognitive Psychology, 27,* 194–226.

Sophian, C. (1997). Beyond competence: The significance of performance for conceptual development. *Cognitive Development, 12,* 281–303.

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

Wason, P. (1983). Realism and rationality in the selection task. In J. St. B. Evans (Ed.), *Thinking and reasoning: Psychological approaches* (pp. 44–75). London: Routledge & Kegan Paul.

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