Reasoning About Multiple Variables: Control of Variables Is Not the Only Challenge

DEANNA KUHN
Teachers College Columbia University, New York, NY 10027, USA

Received 25 October 2006; revised 12 February 2007; accepted 21 February 2007

DOI 10.1002/sce.20214
Published online 21 May 2007 in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: Thirty fourth-grade students participated in an extended intervention previously successful in fostering skills of scientific investigation and inference, notably control of variables (COV). The intervention was similarly successful for a majority of students in the present study, enabling them to isolate the three causal and two noncausal variables operating in a multivariable system. However, when asked to predict outcomes of various constellations of variable levels, they tended not to take into account the effects of all of the causal variables they had identified. Moreover, they did not adhere to a consistency principle, i.e., that a factor that produces an effect can be expected to produce the same effect in the future, given similar conditions. These findings suggest that COV is not the only challenge students experience in reasoning about multiple variables. Elementary-school students' mental models of multivariable causality appear to deviate from a normative, scientific model, even after they have mastered that aspect of scientific method having to do with the design of controlled experiments to isolate effects of individual variables. The challenges, beyond COV, that appear to be involved in making prediction judgments involving multiple variables warrant attention in the design of curricula to foster development of scientific thinking skills. © 2007 Wiley Periodicals, Inc. *Sci Ed* **91:**710–726, 2007

INTRODUCTION

What skills do students need to have developed to be regarded as competent with respect to scientific method? This question is of theoretical significance in the study of cognitive development but also of enormous practical significance to science educators. Mastery of scientific method now appears as a goal in virtually all U.S. state and national curriculum standards (National Research Council, 1996) and commonly in the science curricula of other countries as well (Abd-El-Khalick et al., 2004), despite a lack of consensus as to exactly what it entails (Abd-El-Khalick et al., 2004; Duschl & Grandy, 2005; Kuhn, 2005). Developmental psychologists show a similar lack of consensus regarding the age at which

Correspondence to: Deanna Kuhn; e-mail: dk100@columbia.edu children show competence in scientific method, with some researchers emphasizing early competence (Gopnik, Sobel, Schulz, & Glymour, 2001; Ruffman, Perner, Olson, & Doherty, 1993; Schulz & Gopnik, 2004; Sodian, Zaitchik, & Carey, 1991), and others later lack of competence (Klaczynski, 2004; Klahr, 2000; Klahr, Fay, & Dunbar, 1993; Koslowski, 1996; Kuhn, 1989, 2002; Kuhn, Amsel, & O'Loughlin, 1988; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Kuhn, Katz, & Dean, 2004; Schauble, 1990, 1996).

Progress in resolving such controversies necessitates achieving a clear understanding of exactly what aspects of scientific method students need to acquire competence in, a prerequisite to addressing the equally fraught question of how educators can best realize such goals (Dean & Kuhn, 2007; Klahr & Nigam, 2004). Research in developmental psychology on the *process* of scientific reasoning (as opposed to scientific concepts or understanding) has been devoted almost entirely to development of the control of variables (COV) strategy (i.e., holding constant all other variables than the one under investigation, to eliminate their influence on outcome). (For review see Klahr, 2000; Kuhn, 2002; Kuhn & Franklin, 2006; Zimmerman, 2000, in press.)

Here and elsewhere, I raise the question of whether COV is all that is critical for students to learn about scientific method. Argumentation, explanation and model building, as well as experimentation figure importantly in scientific method (Kuhn, 1993, 2002; Lehrer & Schauble, 2006). Even within the domain of experimentation, a recent study by Kuhn and Dean (2005) suggests that this focus may be overly narrow. In an intervention otherwise confined to exercise of investigation and inference strategies (in the absence of direct instruction) in a multivariable context, we introduced the simple suggestion to middle-school students that they identify a single variable to find out about. This minimal intervention had a pronounced effect on students' experimentation strategies, greatly enhancing the frequency of controlled comparison and valid inference, relative to a control group, in both the original and a new context. This initial phase of identifying a question plays such a powerful role in the subsequent conduct of investigation and inference, we suggested, because it gives meaning and direction to what follows. In the multivariable context of isolation and control of variables, the student may cease to vary other variables across two-instance comparisons because of a gradually increasing sense that they are not relevant to the question at hand.

The work reported here raises further question as to whether the implementation of COV to design and evaluate experiments is all that is important to teach students about scientific method. To anticipate my conclusion, I claim that scientific reasoning about multivariable phenomena poses significant challenges above and beyond COV and that these challenges warrant attention in their own right.

To begin, it is useful to situate COV in its broader framework of causal inference. Scientific reasoning and multivariable causal inference are in fact closely connected (Kuhn & Dean, 2004), although they have been examined in almost entirely separate literatures. An analysis of variance (ANOVA) framework is applicable to both.[1] Among the assumptions that are part of this framework is first the assumption that causes have consistent effects under the same conditions, and second that multiple effects may operate jointly on an outcome, in either additive or interactive fashion. In both scientific reasoning and multivariable causal inference, a number of potential causal variables may or may not be associated with an outcome. In empirical research on individuals' use of these kinds of reasoning, the key difference between them is that in studies of scientific reasoning, individuals typically choose instances to investigate as a basis for subsequent inferences, whereas in

[1] The analysis-of-variance framework is not invoked here in the sense of a claim that it describes students' reasoning, but only in the more limited sense that it provides a model incorporating criteria for systematic integration of multiple simultaneous effects on an outcome.

studies of multivariable causal inference (Cheng, 1997, Downing, Sternberg, & Ross, 1985; Glymour, 2001; Hilton & Slugoski, 1986; Schustack & Sternberg, 1981; Sloman & Lagnado, 2005; Spellman, 1996; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Waldmann & Hagmayer, 2005; Waldmann & Martignon, 1998), individuals typically draw inferences on the basis of instances presented to them. In both cases, however, the ANOVA assumptions apply—multiple effects may be in operation and effects are repeatable (consistent).

In the scientific reasoning context, the individual's task is to identify those effects that are in operation, i.e., those variables that affect a designated outcome variable, a task that requires use of COV. The bulk of studies of scientific reasoning have limited their focus to execution of the COV skill. They examine whether an individual can correctly identify each of the operative variables and exclude the inoperative ones. They have not asked whether, having done so, the individual is then able to reason appropriately about the simultaneous effects of multiple operative variables (the task required in multivariable causal inference tasks), once each of these effects has been identified.

The most straightforward way of assessing this ability is to ask the individual to predict and justify the outcomes for new multivariable combinations not previously examined. Kuhn and Dean (2004) found that preadolescents, as well as some adults, perform poorly on such a multivariable prediction (MVP) task. In addition to typically faulty predictions, individuals often make inconsistent causal attributions across consecutive predictions, for example implicating variable A as causal (when asked, following a prediction, to indicate which variables influenced their prediction) for one prediction and variable B as causal for the next prediction. Furthermore, they often fail to implicate as many variables as influencing their predictions as they had earlier identified as causal when asked to make explicit judgments of the causal roles of each variable (in a multivariable context in which the participant is asked to identify causal and noncausal effects and must use a COV method to do so successfully). For almost half the sample, the median number of variables implicated as influencing an outcome prediction (and hence at least implicitly judged causal) in the context of the MVP task was one (of a possible five). Both of these patterns clearly violate the assumptions of consistency and additivity of effects. Kuhn and Dean (2004) thus characterized these patterns as reflecting an immature mental model of multivariable causality.

One conceivable explanation for this poor performance is that the individuals performing poorly have not mastered a core element of scientific method, namely COV. Perhaps, once they understand how to analyze multivariable constellations into component individual effects, they should have no trouble taking into consideration and aggregating these individual effects so as to perform the MVP task.

One method for testing this hypothesis is correlational, i.e., examine individuals who have or have not achieved mastery of scientific method (defined for these purposes as COV) and assess their ability to make correct MVP judgments. We chose instead to test the hypothesis using a stronger, experimental, rather than correlational, method, one in which we induce the scientific method skill (COV) and assess any resulting effects on performance on the MVP task. (The method is termed experimental not in the sense of random assignment but rather in the more rudimentary sense of undertaking to induce, rather than merely observe, a skill in order to assess its implications or effects.)

If immature mental models of multivariable causality are an epiphenomenon of immature scientific method skill, the previously noted weaknesses (inconsistency and nonadditivity) in the former, as manifested in the MVP task, should disappear.

The experimental design also allows testing of a second hypothesis: that weaknesses in mental models of multivariable causality (as manifested in the MVP task) observed in initial assessments are attributable to uncertainty regarding the causal structure of the domain and the nature of the different possible effects, and inconsistency is therefore a result of vacillation as different ideas are explored. Students investigate the domain over multiple sessions and become familiar with the variables and their effects, following which their MVP skills are assessed. At this assessment, lack of familiarity with the effects of the individual variables can therefore be eliminated as an explanation.

METHOD

Participants

Participants were 30 fourth-grade children in an independent school affiliated with a university in a large urban setting. The school was attended by a mix of children of families affiliated with the university and families from the surrounding community. Students were equally divided by gender and were of heterogeneous ethnic and socioeconomic backgrounds.

Design

Participants were of an age level (fourth grade) at which previous work had shown the feasibility of inducing scientific method skills, in particular COV, during a period of repeated engagement with a problem environment requiring these skills (Kuhn et al., 1995; Kuhn, Black, Keselman, & Kaplan, 2000; Kuhn, Schauble, & García-Mila, 1992). Participants were in the initial months of a 3-year instructional program designed to develop inquiry skills. As a means of promoting COV, students worked in pairs in repeated sessions over a period of 3 months on a problem requiring them to identify which of five potential dichotomous variables were causal and which were noncausal in influencing an outcome. Later in the period, a second task was introduced (MVP) that required predicting outcomes for novel variable constellations and justifying these predictions. On the basis of the latter task, students' conceptions of causal consistency and causal additivity (of multiple factors), i.e., their mental models of multivariable causality, were assessed. The major question to be addressed is whether students who are successful in the first activity will show success in the second, i.e., will have achieved scientifically correct mental models of multivariable causality.

Procedure

Investigation/Inference (COV). Students worked with software designed for the purpose of developing skills in scientific investigation and inference, in particular controlled comparison (COV). The software depicts multivariable causal systems resembling those used in earlier research (Kuhn et al., 1992, 1995, 2000; Schauble, 1990, 1996). The students' task is to choose cases for examination that represent different combinations of variables and draw appropriate inferences regarding which variables do and do not make a difference to the outcome.

In the *Earthquake Forecaster* version, students play the role of junior earthquake forecaster and are introduced to a set of five dichotomous variables that may or may not affect earthquake risk. They are able to choose cases to investigate (consisting of combinations of variable levels of each variable) and observe resulting risk levels. In the *Ocean Voyage* program, similarly students investigated which of a set of five dichotomous variables influence the progress of ancient ships across the ocean. In each case, of five potential causal

TABLE 1
Causal Structure Represented in *Earthquake Forecaster*

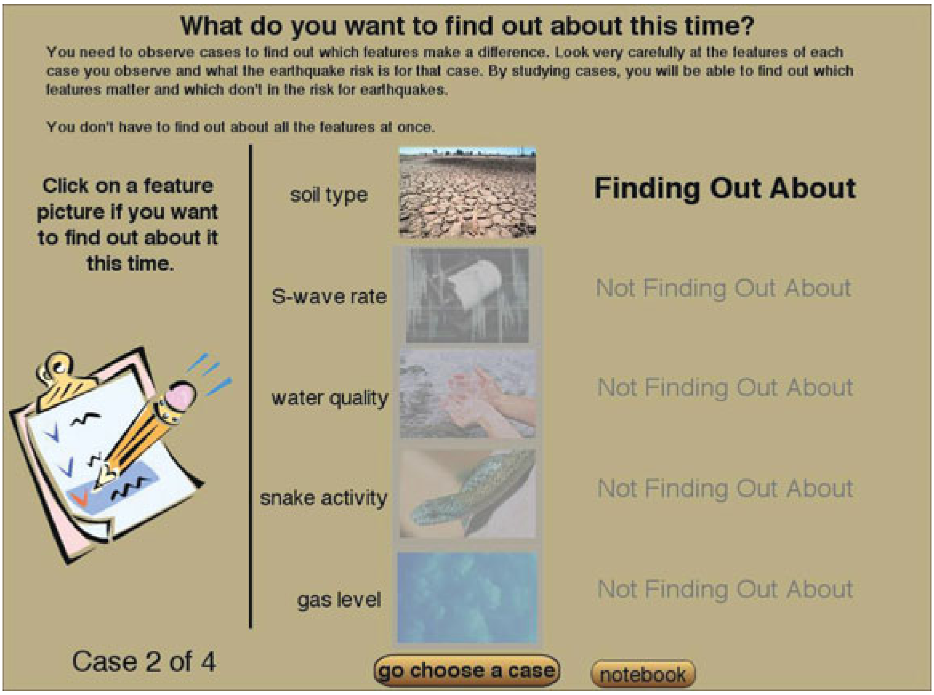
Feature	Effect	Outcome
Soil type (igneous and sedimentary)	Noncausal	Both yield identical outcomes
S-wave rate (fast and slow)	Noncausal	Both yield identical outcomes
Water quality (good and poor)	Causal	Good yields one level of greater risk
Snake activity (high and low)	Causal	High yields one level of greater risk
Gas level (heavy and light)	Causal	Heavy yields one level of greater risk

Outcome levels: Extreme, high, medium, or low earthquake risk.

variables, three in fact have equal additive causal effects on outcome and two have no effect. The causal structure and variables represented in *Earthquake Forecaster* appear in Table 1.

Pretest. All students participated in an individual pretest assessment session with *Earthquake Forecaster*. This session took place in a room adjacent to the classroom and was conducted individually by a young adult who was a member of the research team and not part of the school staff. This adult provided any guidance the student needed with the program. One cycle of this module allows the student to select a sequence of four cases for examination, observe outcomes, and, following each, to draw inferences about the causal, noncausal, or indeterminate status of each of the five variables, to provide justifications for inferences, and to record any information desired in an electronic notebook, which remained available for their consultation.

After an initial introduction, students are asked to choose what they will find out about in their first case selection (see Figure 1). Students identify whether they are or are not finding



What do you want to find out about this time?

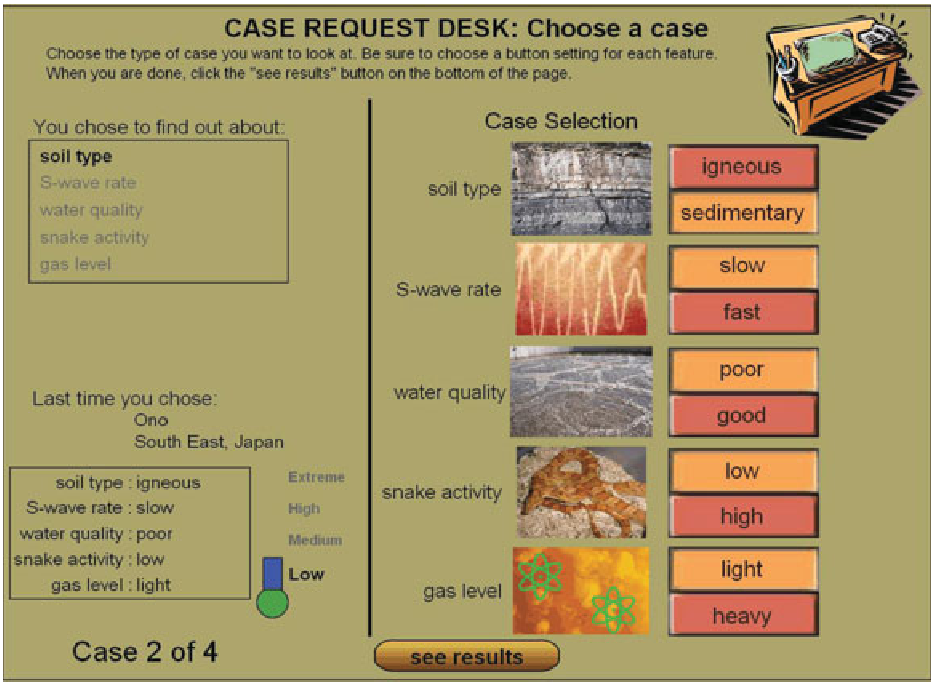
You need to observe cases to find out which features make a difference. Look very carefully at the features of each case you observe and what the earthquake risk is for that case. By studying cases, you will be able to find out which features matter and which don't in the risk for earthquakes.

You don't have to find out about all the features at once.

Click on a feature picture if you want to find out about it this time.

soil type	Finding Out About
S-wave rate	Not Finding Out About
water quality	Not Finding Out About
snake activity	Not Finding Out About
gas level	Not Finding Out About

Figure 1. Find out screen. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]



CASE REQUEST DESK: Choose a case

Choose the type of case you want to look at. Be sure to choose a button setting for each feature. When you are done, click the "see results" button on the bottom of the page.

You chose to find out about:

soil type
S-wave rate
water quality
snake activity
gas level

Last time you chose:
Ono
South East, Japan

soil type : igneous
S-wave rate : slow
water quality : poor
snake activity : low
gas level : light

Case 2 of 4

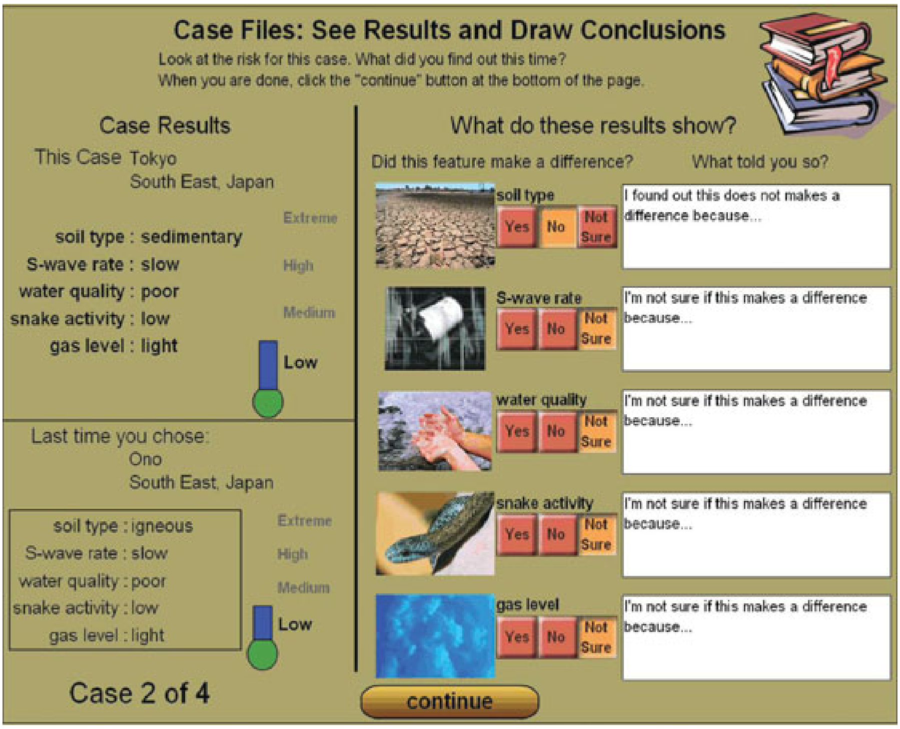
Figure 2. Case request screen. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

out about a feature by clicking the feature picture(s) corresponding to their choice(s). Then, students construct a case of their own choosing, by selecting the level (Table 1) of each feature (see Figure 2). These choices yield an outcome displayed in the form of a gauge representing the earthquake risk level. Students are then asked to make any inferences they believe to be justified regarding the causal or noncausal status of any of the features (Figure 3). The final screen for each case prompts the student to enter any notes they wish to (Figure 4).

Each of the screens presented here is depicted as they would appear during a second case investigation. The first-case screen includes no reference to a previous case; subsequent screens include not only the outcome for the current case but also show results for the immediately preceding case. After answering questions regarding the outcome of the fourth case and allowing the student to make any additional notes on the final notebook screen, the program thanks the student for participating and shuts down.

Investigation/Inference Exercise. Approximately 1 month after the pretest session, following a midyear vacation, students began work with the parallel *Ocean Voyage* program. This program is identical in all respects to *Earthquake Forecaster* except for content, which involves the variables that affect the success of an ancient ocean voyage across the sea. The five variables were captain's age (young or old), crew size (large or small), navigation (compass or stars), sail type (latteen or square), and ship hull shape (round or V).

Work on *Ocean Voyage* was integrated into students' regular classroom work. Sessions lasted from 30 to 45 minutes and took place once or occasionally twice per week, depending on the class schedule, over a period of 9 weeks, interrupted by a 2-week school vacation midway through. The activity was introduced to students by the classroom teacher as a unit on inquiry, defined as "how to find out things." The *Ocean Voyage* content was designed to complement students' work on a unit on the seas, which had been designed by



Case Files: See Results and Draw Conclusions

Look at the risk for this case. What did you find out this time?
When you are done, click the "continue" button at the bottom of the page.

Case Results

This Case Tokyo
South East, Japan

soil type : sedimentary
S-wave rate : slow
water quality : poor
snake activity : low
gas level : light

Last time you chose:
Ono
South East, Japan

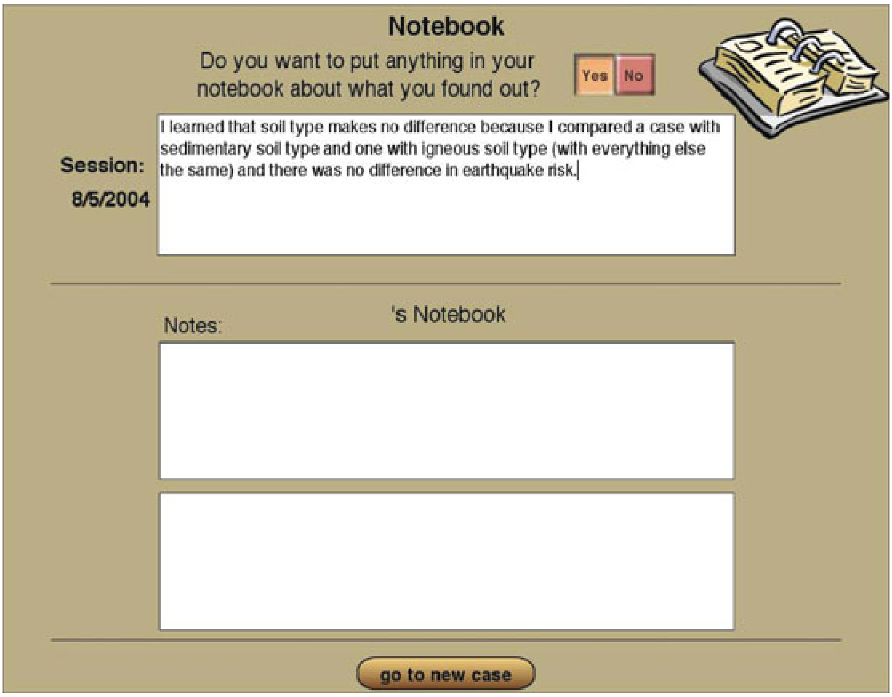
soil type : igneous
S-wave rate : slow
water quality : poor
snake activity : low
gas level : light

Case 2 of 4

What do these results show?

Did this feature make a difference?	What told you so?
soil type	I found out this does not makes a difference because...
S-wave rate	I'm not sure if this makes a difference because...
water quality	I'm not sure if this makes a difference because...
snake activity	I'm not sure if this makes a difference because...
gas level	I'm not sure if this makes a difference because...

Figure 3. Results and conclusions screen. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]



Notebook

Do you want to put anything in your notebook about what you found out?

Session:
8/5/2004

I learned that soil type makes no difference because I compared a case with sedimentary soil type and one with igneous soil type (with everything else the same) and there was no difference in earthquake risk.

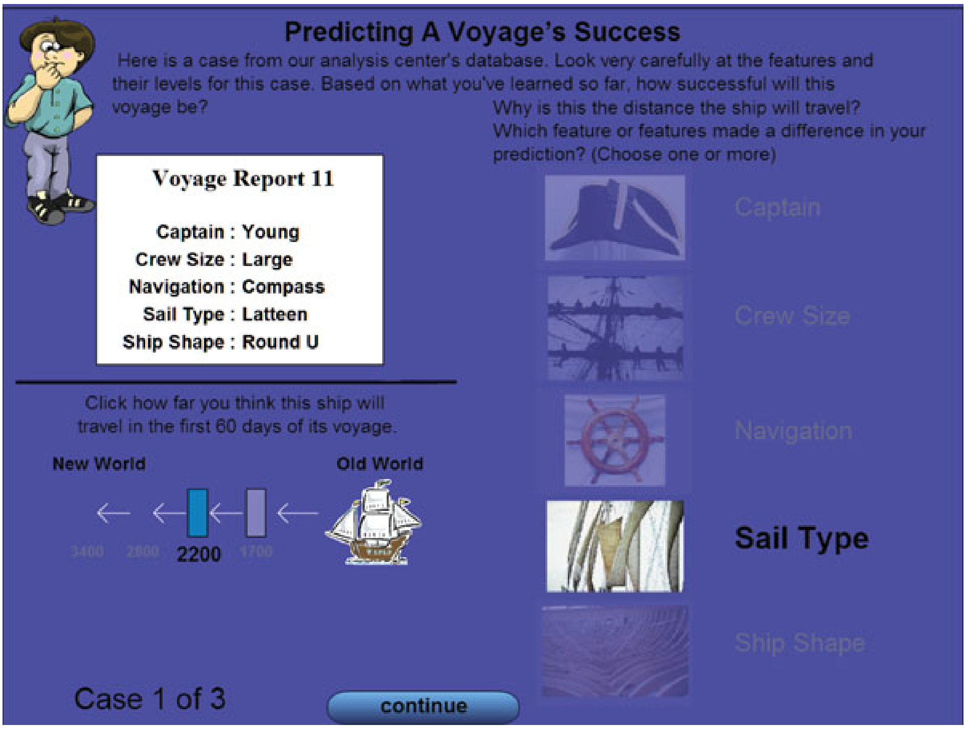
Notes: 's Notebook

Figure 4. Notebook screen. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

their classroom teachers and took place contemporaneously but did not include any of the specific content of *Ocean Voyage*. Except for occasional brief departures for some purpose, the classroom teacher generally remained in the classroom and supervised the activity, with the assistance of one or two members of the research team. The teacher also took charge of procedures the class had developed for assigning partners.

Because of student absences and other scheduling issues, the number of times a student worked on the *Ocean Voyage* program varied somewhat, with a range from 5 to 9 and a mean of 7.25. Students worked in pairs, and only occasionally alone (when an uneven number of students was present) to encourage them to externalize their thinking and deliberate regarding their judgments, a method we had found productive in previous work. They worked with a new partner at each session.

Prediction/Attribution (MVP). After students had worked with the *Ocean Voyage* program for 1 month, the prediction/attribution module of *Ocean Voyage* was introduced as a separate "assess your skill" module. It requires the student to predict an outcome based on a combination of levels of the five dichotomous variables and to indicate which variables influence the prediction. This task served as the basis for inferring a student's model of multivariable causality. Students worked individually on this module over repeated sessions for a period of approximately 1 month, coincident with the final month of their work on the investigation module. The number of sessions varied, with a range from 2 to 6 and a mean of 4.5. Sessions lasted about 5 minutes and occurred at the conclusion of the investigation/inference session. Students had access to their electronic notebooks if they wished to consult them.



Predicting A Voyage's Success

Here is a case from our analysis center's database. Look very carefully at the features and their levels for this case. Based on what you've learned so far, how successful will this voyage be?

Voyage Report 11

Captain : Young
Crew Size : Large
Navigation : Compass
Sail Type : Latteen
Ship Shape : Round U

Click how far you think this ship will travel in the first 60 days of its voyage.

Why is this the distance the ship will travel? Which feature or features made a difference in your prediction? (Choose one or more)

Case 1 of 3

Figure 5. Prediction (MVP) screen (*Ocean Voyage* program). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

At each session, a series of three instances were presented consecutively on the screen, each consisting of a particular constellation of variable levels, without any outcome depicted. (See Figure 5.) The student was asked to predict the outcome for that case (of four levels ranging from least to greatest) and then in the list of variables that followed, indicate those that affected the prediction. ("Why is this the outcome? Which feature or features made a difference in your prediction?") Each of the variables was listed and the instruction indicated "Choose one or more." No feedback was provided with respect to correctness. At subsequent sessions, students were given the general feedback that their previous predictions had not been entirely correct and that they should keep working on their prediction skills.

Investigation/Inference (COV) Posttest. To ascertain that students' progress in investigation/inference skills is not limited to the particular content in which the skills developed, it is necessary to establish that the skills generalize to new content. For this purpose, the *Earthquake Forecaster* pretest was readministered individually to all students, during the week following completion of work on the investigation/inference and prediction modules.

RESULTS

Despite their relatively young age, a majority of participants, though not all, made substantial progress in the development of investigation and inference skills, as the result simply of exercise in an environment that required them and despite the absence of specific instruction—a result consistent with previous research (Kuhn et al., 1992). In contrast, progress was minimal with respect to exhibition of mental models of multivariable causality having the fundamental analysis-of-variance characteristics of consistency and additivity.

Pretest

No student showed competence in controlled comparison and identification of causal and noncausal factors in the pretest assessment on the *Earthquake Forecaster* program. Specifically, no student chose two consecutive cases that represented a controlled comparison (i.e., were composed of the same variable level for all variables except one) and drew an appropriate conclusion regarding the effect of the varied variable. The typical strategies, consistent with earlier research on scientific investigation (Kuhn, 2002; Zimmerman, 2000, in press), were to set out to investigate the effects of multiple factors at once, to compare cases that varied on multiple dimensions, and to draw invalid inferences justified by reference to compatible pieces of evidence in conjunction with theoretical expectation or justified entirely by theoretical belief. Thus, for example, in *Earthquake Forecaster* a student might in observing an initial instance attribute the high-earthquake risk entirely to the high-snake activity associated with this instance and justify her inference entirely on the basis of her prior belief that a high level of this variable is associated with earthquake risk. The next instance she chooses to observe is characterized by low-snake activity and low-earthquake risk, and this time she justifies her causal inference regarding snake activity based on the covariation of this variable with outcome over the two instances. She ignores, however, two other variables she has also chosen to vary over the two instances and that also covary with outcome.

Investigation/Inference Exercise

Also consistent with earlier research (Kuhn et al., 1992, 1995, 2000), after several sessions of practice with the *Ocean Voyage* program, the investigatory and inference strategies of the majority of students began slowly to improve. As microgenetic analyses of the nature of such improvement have been presented elsewhere (Kuhn et al., 1995) and are not central to the purpose of the present research, they are not presented here. Instead, participants are categorized into two broad groups that fit the purposes of the present study, those who made substantial progress in the development of scientific reasoning skill and those who did not.

Posttest Classification

Posttest performance on *Earthquake Forecaster* was used as the criterion for classification of a participant in the successful category. However, all students classified as successful on the *Earthquake Forecaster* posttest did begin to show success on the *Ocean Voyage* program at least by the later sessions and all had correctly identified the three causal and two noncausal *Ocean Voyage* variables by the final session.[2]

Students classified as successful at a minimum demonstrated two sequences of consecutive cases in their posttest performance on *Earthquake Forecaster* that met these criteria:

1. Indication of an intent to find out about the effect of a single variable (see Figure 1).
2. Selection of two instances that varied with respect only to the variable indicated in #1 (see Figure 2).
3. An appropriate inference of causality or noncausality for the variable indicated in #1 based on the results of the comparison constructed in #2 (see Figure 3).

(A criterion of consistent, i.e., uniform, display of these characteristics across the entire session was not imposed since a mixture of usage of advanced and less advanced skills has been found the norm; Kuhn et al., 1995; Siegler, 2006.)

Using these criteria, 19 of the 30 participants were categorized as successful and 11 were not, a proportion in accord with what would be expected at this age level following dense engagement with a problem environment requiring scientific reasoning skills (Kuhn, 2002).

Prediction/Attribution (MVP)

Our interest centers on the 19 participants classified as successful. Does their mastery of scientific reasoning skill beyond what would be expected for their age level, as well as their familiarity with the *Ocean Voyage* content, show evidence of improving their mental models of multivariable causality to a more scientifically adequate level than that observed among preadolescents and many adults in previous research (Kuhn & Dean, 2004)? To

[2] Considerable evidence is now available that children of this age do not conceptualize interaction effects among variables (Kuhn, 2002; Kuhn et al., 1995; Zimmerman, 2000). A concern about the possibility of such effects was never voiced during students' *Ocean Voyage* investigations, and it is hence unlikely that a concern about interactions among variables had a detrimental effect on their predictions. In any case, it would not account for inconsistency from one prediction to the next in the variables implicated as influencing the prediction.

TABLE 2
Individual Averages (Over Three Final MVP Sessions) of Inconsistency of Causal Attributions Within a Session

	Causal Variables[a]	Noncausal Variables[b]	All Variables[c]
TD	0.00	0.00	0.00
OD	0.33	0.00	0.33
TM	0.33	0.00	0.33
MB	0.33	0.33	0.66
OG	0.66	0.66	1.33
FB	1.33	0.00	1.33
PG	1.33	0.00	1.33
SF	2.00	0.00	2.00
KN	1.33	1.00	2.33
MM	1.66	0.66	2.33
SS	1.66	0.66	2.33
LD	1.66	1.00	2.66
CN	1.66	1.00	2.66
DM	2.00	0.66	2.66
MF	2.66	0.00	2.66
DT	1.66	1.33	3.00
MR	2.00	1.33	3.33
CG	2.00	1.66	3.66
TA	2.33	1.33	3.66

[a] Mean number of variables (of 3) for which inconsistency appears.
[b] Mean number of variables (of 2) for which inconsistency appears.
[c] Mean number of variables (of 5) for which inconsistency appears.

answer this question, we assessed these models over several occasions, rather than just once, for reliability and to eliminate unfamiliarity with the MVP task as a possible source of difficulty.

At each *prediction/attribution (MVP)* session,the student was asked to make predictive judgments about three cases. For each case, from zero to five variables could be implicated as having played a role in the outcome. Our primary interest is in the degree of consistency of these causal attributions within a session over the three cases. Inconsistency for a particular variable is defined as not consistently implicating the variable as either causal or noncausal (i.e., implicating the variable as causal with respect to one or two of the three instances that make up a session of the MVP task and noncausal with respect to the others). From one session to the next, it is plausible that conclusions regarding causal status of a variable would change, especially as students were working on the investigation module during the same period, but one would not expect these conclusions to change within a session, especially in the absence of any feedback.

Table 2 presents for each of the 19 students who had been classified in the successful category the mean number of variables for which inconsistency appears during the student's three final sessions of the MVP task. At each session, inconsistency can be shown for any number from 0 to 5 of the five variables presented in the task. We average over the last three sessions, given the variability across sessions that is the norm in microgenetic studies (Kuhn et al., 1995; Siegler, 2006). Students are identified by initials and listed individually in Table 2 in an order reflecting overall performance, with the better performing (lower inconsistency scores) appearing first. Only the 19 students previously classified in the successful category are included.[3]

Results are displayed by individual since it is individual patterns that are of concern, not group averages that could mask inconsistency, as well as separately for causal and noncausal variables, i.e., the three that in fact had causal effects on outcomes vs. the two that did not. Students, especially the 19 ultimately successful ones considered here, would have had ample opportunity by this point to have identified the variables that had causal effects on outcomes and those that had no effects, and we confirmed from the individual records of each participant that all 19 had indeed done so.

As seen from comparison of the first two columns in Table 2, and consistent with previous research (Kuhn et al., 1995), causal variables pose more of a challenge to sound reasoning than do noncausal variables. For causal variables, the mean of the individual means (of number of variables for which inconsistency is shown) is 1.42 (of a possible 3). For noncausal variables, the mean of individual means is 0.61 (of a possible 2). Thus, inconsistency tends to be more frequent when the variable is causal (almost half of the time, versus less than a third). Nonetheless, inconsistency in causal attribution, we see, remains a significant limitation with respect to both causal and noncausal variables for a majority of these students, despite the progress they have made in scientific reasoning.[3]

Only the first two (TD and OD) of the 19 students in Table 2 showed clear progress in the consistency of their causal attribution during the course of their engagement with the activity. TD showed a sharp transition between the second and third sessions, with initial inconsistency (mean for first two sessions) of 2.00 for causal variables and 1.00 for noncausal variables (hence 3.00 for both). At the next session, inconsistency dropped to zero for both causal and noncausal variables and remained there (see Table 2). OD showed a more gradual transition from an initial inconsistency (mean for first three sessions) of 1.33 for causal variables and 0.33 for noncausal variables, dropping to 0.33 and 0.00, respectively, by the final sessions, as shown in Table 2. All other students showed negligible improvement in consistency over sessions.

The other major question we wish to ask is the extent to which students' responses in the MVP task appropriately incorporate the roles of all three causal variables. In contrast to Table 2, we look here at their performance only at the final MVP session, when their knowledge and skill should have been at its maximum. Again, despite the fact that the 19 participants considered here had by this point all successfully identified the three variables that had causal effects on outcomes and the two that did not, underattribution of causality remains a significant constraint on their causal reasoning. Among these 19, only seven consistently implicate all three causal variables as having contributed to the outcome. Of the remaining 12, four implicate a median of 2 as causal (over the three prediction judgments that constitute the session), and the remaining eight implicate a median of 1 as causal.

When the two characteristics, inconsistency and underattribution, are combined, rather than the normative scientific model of multivariable causality, what emerges as characteristic is a model in which the explanatory burden in a multivariable context shifts from one single variable to another single variable over time. A student we have called Dora provides a typical example. She was successful in the investigation/inference task, identifying the three causal and two noncausal variables in *Ocean Voyage*, as well as meeting the stipulated criteria in the *Earthquake Forecaster* posttest. Her fourth attempt at the MVP task, however,

[3] Comparable levels of performance were observed on the part of the 11 participants who did not meet the criteria indicated and were not classified in the successful category. Mean inconsistency scores were 1.75 for causal variables and 0.85 for noncausal variables. Hence, students in the successful category did not show significantly greater mastery in the MVP task, as a result of their achievement in scientific reasoning, than did students who did not show this scientific reasoning achievement.

after she was well experienced with it, proceeded as follows. Her first prediction was incorrect (by one level, of the four possible prediction outcomes). In response to the accompanying query ("Why is this the outcome?," followed by the opportunity to select any number of the five variables listed), she selected only causal variable 1 (CV1). A new, second variable constellation appeared and Dora was prompted to predict an outcome. Her second prediction was incorrect (by two levels), and she selected only noncausal variable 1 (NCV1) as having played a role in her prediction. Dora's third prediction was correct; however, she still implicated only a single variable (this time CV2) as having played a role in her prediction.

Dora's next set of predictions was similar. The first was in fact correct, but she implicated only CV3 in her prediction. Her second prediction was incorrect (by one level) and this time she returned to CV1 as having influenced her prediction. Her third prediction was incorrect (by two levels) and this time she again implicated CV2 as the only influence. Dora's only consistent implicit causal attribution over this set of predictions was NCV2, which she never implicated as causal. Moreover, never did she implicate more than one variable as having influenced her prediction, despite the fact that the instructions stated clearly (and were verbally reinforced by the interviewer) that she could choose as many of the five as she wished.

Again, Dora's case is entirely typical, and what is striking about it is the fact that she has demonstrated the correct causal knowledge (of which variables are causal) that would enable her to perform correctly on the MVP task. Yet she does not do so. Her problem is not isolation of variables and identification of causal effects, it appears, but the quite different problem of coordination of the multiple effects she has identified. Constraining this coordination effort, it is proposed here, is an inadequate model of multivariable causality, one in which the principles of consistency and additivity do not apply.

A final question is the relation between the two characteristics, inconsistency and under-attribution. To investigate this question, the 19 successful students were grouped according to degree of inconsistency shown over the last three sessions (as in Table 2) and the resulting attribution frequencies were examined by group. As Table 3 shows, those students having lower inconsistency scores come closer to correctly attributing causality to all three causal variables (a score of 3.0). The pattern is similar when one examines attributions as a function of consistency for only the three causal variables for which causality should have been attributed. This association between the two characteristics, inconsistency and under-attribution, supports the interpretation that they are joint manifestations of an immature understanding of multivariable causality.

TABLE 3
Mean Number of Variables Implicated As Causal (Over Three Final MVP Sessions As a Function of Degree of Consistency of Causal Attributions

Mean Number of Variables for Which Inconsistency Appears (of a Possible 5)	Mean Number of Variables Implicated as Causal (of a Possible 3)	*N*
<0.50	2.70	3
$0.51 < 1.99$	2.25	4
$2.00 < 2.99$	1.86	8
>3.00	1.80	4

DISCUSSION

The research on multivariable causal inference cited at the outset of this article overwhelmingly emphasizes impressive reasoning competence exhibited by both children and adults. The present results, based on the performance of fourth graders, indicate the need for qualification, as do results for adults reported elsewhere (Kuhn, 2007; Kuhn & Dean, 2004). A mental model of multivariable causality that conforms to a normative, scientific standard cannot be taken for granted as underlying children's (or adults') causal reasoning. Nor is mastery of the scientific method of controlled comparison and isolation of individual causal effects a sufficient condition to insure such a model. It appears to be a cognitive achievement of its own and therefore deserving of attention in its own right.

The term "mental model" traditionally has been used to characterize an individual's conception of some particular objects or events. But it may be useful as well to characterize more generic conceptions, such as mental models of causality, that extend across different phenomena. Acquisition of a sound mental model of multivariable causality is of at least as great practical import as is the cognitive skill that has been investigated so extensively under the heading of COV. Relatively few people have the opportunity or inclination to design experiments, whereas natural-experiment observations are commonly available (Kuhn & Brannock, 1977). And people make causal attributions all of the time. Doing so entails coordination of theoretical expectations and new information (Keil, 1998; Kuhn, 1989). Limiting oneself to a single explanatory variable constrains explanation. Doing so has been studied under the heading of "discounting" in the social psychology literature, with the implication that other potentially explanatory variables have been considered and rejected, and discounting has been noted to appear in children by age 8 or 9 (Sedlak & Kurtz, 1981). For many individuals at least, it may be more accurate to characterize the inference process as based on a model in which single factors most often suffice to explain outcomes and need not remain consistent across instances.

Once the consistency rule is abandoned, the resulting freedom allows the illusion of drawing on new evidence, but it is done selectively in a way that protects one's theories without ever subjecting them to serious test. If a new piece of evidence threatens one theory, its implications can be avoided simply by shifting the explanatory burden to another variable (Kuhn et al., 1995). Allowed free reign, this mental model of causality leads to the fallacies in attribution that are all too familiar in everyday thinking: Superior skill accounted for our team's victory, but the other team's win was due to luck.

How do we explain the failure of students in the present study to utilize the knowledge they gained in the first (investigation and inference) task in performing the second (prediction) task? Failure of cognitive skills to transfer to new contexts is a well-documented phenomenon (Detterman & Sternberg, 1993), and in this sense the present findings perhaps need no particular explanation. Yet, we know that children younger than the age of those in this study can under certain conditions appropriately integrate information from at least two sources in an additive fashion (Anderson, 1991; Dixon & Tuccillo, 2001; Wilkening, 1982), so their failure to do so in the present causal context cannot readily be attributed to processing limitations. In the present case, over the course of repeated occasions that provided familiarity with the task format, they needed to integrate the effects of three variables, whose individual effects they had already ascertained and by this point well understood. If the problem were one of processing overload, they might have simply ignored one of the variables and focused on integrating the other two. The results, however, do not support such a model.

Another possible interpretation is one of task interference. Perhaps the first task biased children to think in terms of the effects of individual variables in isolation and this interfered with their ability to consider multiple effects in the second task. There is little evidence to support this interpretation, however. The two task formats were similar (compare Figures 3 and 5) in presenting all five variables for consideration and asking for judgments with respect to all of them. Difference between the two lies rather in the task itself: in the first case asking the respondent to identify the effect of each variable and in the second to integrate these individual effects (and indicate the basis for doing so). The first task is clearly a prerequisite for performing the second. Rather than the first task's ever being conceived of as interfering with the second, typically all research attention has been focused on the first task while the second has been considered a nonissue. In other words, it has simply been assumed that once an individual was able to isolate and identify the role of individual variables, making outcome predictions based on this knowledge would be no problem. This is an important assumption, given the ubiquitous occasions people have to make such predictions in a multivariable world—an assumption that the present results indicate is not warranted.

If skill in the MVP task does not emerge unaided, we are left with the question of what is necessary to master it. The MVP task poses two challenges, corresponding to the two assumptions of the ANOVA model underlying multivariable causality. First, all variables that affect the outcome must be taken into account, rather than only those that happen to be the momentary focus of attention. This skill, as suggested above, at least for the three-variable case examined here is more likely to rest on disposition than competence. As such it is likely to be influenced by the increasing metalevel mental self-monitoring that is characteristic during the second decade of life (Klaczynski, 2004; Kuhn, 2005; Kuhn & Franklin, 2006) and that supports the consideration of multiple alternatives. Attention to alternatives is critical in scientific reasoning (in recognizing that uncontrolled alternative variables may affect outcome), as well as in recognizing counterexamples in conditional reasoning (Klaczynski, 2004; Markovits & Barrouillet, 2002), both forms of reasoning that show improvement in early adolescence (Kuhn & Franklin, 2006).

The second challenge posed by the MVP task, however, corresponding to the other ANOVA assumption, is more difficult to accomplish, and that is a revision of one's model of multivariable causality to incorporate consistency across occasions as a constraint. If students are operating under a model in which no (or inadequate) constraints exist requiring that causal variables have to operate consistently across different occasions, the factors that would lead them to impose such constraints are not obvious, and, as has been suggested here, even many adults may not have incorporated this constraint in their causal reasoning. Keselman (2003) found that longer term practice (in predicting outcomes based on multiple causal variables) was as effective as the direct instruction she undertook, but the gains she observed among young adolescents were only small ones. In current work, we are examining collaborative practice, where partners must justify their predictions to one another, given its success in developing investigative and inference skills, as well as feedback (from predictions), which may serve both cognitive and affective functions.

For now, the point to be made is that this educational objective is one essential to address in designing curricula that seek to promote the development of students' scientific thinking skills. The skills of COV and MVP, as they are assessed in the tasks employed in the present work, are of course only greatly simplified pieces of a complex network that represents scientific thinking. Yet the present work suggests that each of these pieces must be identified and examined if we are eventually to gain better understanding of the interrelated whole.

REFERENCES

Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N., Mamiok-Naaman, R., Hofstein, A., Niaz, M., Treagust, D., & Tuan, H. (2004). Inquiry in science education: International perspectives. Science Education, 88, 397–419.

Anderson, N. (1991). Contributions to information integration theory: Vol. III. Development. Hillsdale, NJ: Erlbaum.

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

Dean, D., Jr., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. Science Education, 91, 384 – 397.

Detterman, D., & Sternberg, R. (1993). Transfer on trial: Intelligence, cognition, and instruction. Norwood, NJ: Ablex.

Dixon, J., & Tuccillo, F. (2001). Generating initial models for reasoning. Journal of Experimental Child Psychology, 78, 178–212.

Downing, C., Sternberg, R., & Ross, B. (1985). Multicausal inference: Evaluation of evidence in causally complex situations. Journal of Experimental Psychology: General, 114, 239–263.

Duschl, R., & Grandy, R. (2005). Reconsidering the character and role of inquiry in school science: Framing the debates. Unpublished manuscript, Rice University.

Glymour, C. (2001). The mind's arrows. Cambridge, MA: MIT Press.

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.

Hilton, D., & Slugoski, B. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. Psychological Review, 93, 75–88.

Keil, F. (1998). Cognitive science and the origins of thought and knowledge. In W. Damon (Series Ed.) & R. Lerner (Vol. Ed.), Handbook of child psychology (5th ed., Vol. I). New York: Wiley.

Keselman, A. (2003). Promoting scientific reasoning in a computer-assisted environment. Journal for Research in Science Teaching, 40, 898–921.

Klaczynski, P. (2004). A dual-process model of adolescent development: Implications for decision making, reasoning, and identity. In R. Kail (Ed.), Advances in child development and behavior (Vol. 31). San Diego, CA: Academic Press.

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

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

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

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

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

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. Science Education, 77(3), 319–337.

Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), Handbook of childhood cognitive development. Oxford, England: Blackwell.

Kuhn, D. (2005). Education for thinking. Cambridge, MA: Harvard University Press.

Kuhn, D. (2007, February/March). Jumping to conclusions: Can people be counted on to make sound judgments? Scientific American–Mind, 18(1), 44–51.

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 and Development, 5, 261–288.

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

Kuhn, D., & Franklin, S. (2006). The second decade: What develops (and how)? In W. Damon & R. Lerner (Series Eds.) & D. Kuhn & R. Siegler (Vol. Eds.), Handbook of child psychology: Vol. II. Cognition, perception, and language (6th ed.). Hoboken, NJ: Wiley.

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

Kuhn, D., Katz, J., & Dean, D. (2004). Developing reason. Thinking and Reasoning, 10, 197–219.

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

Lehrer, R., & Schauble, L. (2006). Scientific thinking and scientific literacy: Supporting development in learning contexts. In W. Damon & R. Lerner (Series Eds.) & K. A. Renninger & I. Sigel (Vol. Eds.), Handbook of child psychology (6th ed., Vol. IV). Hoboken, NJ: Wiley.

Markovits, H., & Barrouillet, P. (2002). The development of conditional reasoning: A mental model account. Developmental Review, 22, 5–36.

National Research Council (1996). The National Science Education Standards. Washington, DC: National Academy Press.

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

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.

Schulz, L., & Gopnik, A. (2004). Causal learning across domains. Developmental Psychology, 40(2), 162–176.

Schustack, M., & Sternberg, R. (1981). Evaluation of evidence in causal inference. Journal of Experimental Psychology: General, 110, 101–120.

Sedlak, A., & Kurtz, S. (1981). A review of children's use of causal inference principles. Child Development, 52, 759–784.

Siegler, R. (2006). Microgenetic studies of learning. In W. Damon & R. Lerner (Series Eds.) & D. Kuhn & R. Siegler (Vol. Eds.), Handbook of child psychology: Vol. II. Cognition, perception, and language (6th ed.). Hoboken, NJ: Wiley.

Sloman, S., & Lagnado, D. (2005). Do we "do?" Cognitive Science, 29, 5–39.

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

Spellman, B. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. Psychological Science, 7, 337–342.

Steyvers, M., Tenenbaum, J., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. Cognitive Science, 27, 453–489.

Waldmann, M., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 216–227.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), Proceedings of the Twentieth Annual Conference of the Cognitive Science Society (pp. 1102—1107). Mahwah, NJ: Erlbaum.

Wilkening, F. (1982). Integrating velocity, time, and distance information: A developmental study. Cognitive Psychology, 13, 231–247.

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

Zimmerman, C. (in press). The development of scientific thinking in elementary and middle school. Developmental Review.