Science
Education

Can Engaging in Science Practices Promote Deep Understanding of Them?

DEANNA KUHN, TOI SIN ARVIDSSON, ROSIANE LESPERANCE, RAINIKKA CORPREW
Teachers College Columbia University, New York, NY 10027, USA

Received 10 June 2016; revised 19 September 2016; accepted 16 December 2016
DOI 10.1002/sce.21263
Published online 25 January 2017 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: It is now widely accepted, and indeed emphasized in the Next Generation Science Standards, that science education should encompass scientific practice as well as science content. By participating in an intellectual community engaged in the broad range of activities that constitute scientific inquiry, rather than simply mastering isolated science procedures, it is hoped students will come to better understand and appreciate the norms, goals, and values that govern the conduct of science. We put this expectation to empirical test by engaging a secondary school science class in an extended problem-based activity that included design of investigations, multivariable causal analysis, and argumentation. Compared to students in a nonparticipating control group, in delayed assessments involving new content, participating students showed more advanced investigation, analysis, and argumentation skills, but also superior epistemological understanding regarding science as entailing the evaluation of claims in relation to available evidence. © 2017 Wiley Periodicals, Inc. *Sci Ed* **101:**232–250, 2017

INTRODUCTION

We are among an increasing number of those concerned with science education who advocate an approach to science process skills that goes beyond a long-standing focus on teaching students experimental design and specifically the control of variables strategy. This view in fact has now become explicit in the Next Generation Science Standards

Correspondence to: Deanna Kuhn; e-mail: dk100@tc.columbia.edu

(NGSS), which specify as a key objective of science education acquainting students with science as a practice. Science practice encompasses a range of activities that include posing questions, developing hypotheses, designing and conducting experiments, examining and interpreting data, constructing arguments and counterarguments and debating conclusions. It is believed most effective if students engage in these activities first hand, not as isolated procedures but as interconnected aspects of broad goal-based inquiry that addresses one or more significant questions (Ford, 2012; Kuhn, 2010; Lehrer & Schauble, 2015; Manz, 2014; Sandoval, 2014). By participating in communities of scientific inquiry, it is hoped, students will acquire not only familiarity with and facility in scientific practices but will come to better understand and appreciate the norms, goals, and values that govern the conduct of science (McNeill, 2011; Ryu & Sandoval, 2012).

In the work reported here, we seek to examine this important expectation. Students' understanding regarding science practice is very limited in elementary school (Sandoval, Sodian, Koerber, & Wong, 2014) and has been found not to improve greatly with age (Metz, 2004; Sandoval, 2005; Smith, Maclin, Houghton, & Hennessey, 2000). Does deep engagement in science practice of the broad sort identified above change this picture? Because efforts to implement this sort of engagement are fairly recent and not yet widespread, there does not yet exist a great deal of data that directly address this question.

That deep engagement in science practice foster understanding of the epistemological foundations of science is of critical importance, rather than merely desirable. Students must come to recognize scientific claims not simply as accumulated (unquestioned and unchanging) facts or freely chosen opinions ("I personally don't believe in climate change"), but rather as judgments requiring evaluation in a framework of alternatives and evidence (Greene, Sandoval, & Braten, 2016; Moshman, 2015; Ricco, 2015). Science as argument has come into broad favor in science education (Berland & Hammer, 2012; Kuhn, 1993, 2010; Manz, 2014; McNeill, 2011; Osborne, Erduran, & Simon, 2004), but without the epistemological foundation just indicated, debating scientific claims is a practice that can have only limited meaning to students.

Ideally, deep engagement in science practices and developing deep understanding of their epistemological foundation reinforce one another. Here we report on our effort to explore their relationship by engaging science students in an extended encounter with key science practices, in a format high in demand for reflection, and examining consequences with respect to understanding of science as involving the debate of alternative claims in a context of available evidence.

We begin by characterizing the practice of science as we implemented it in the sequence of activities participating students engaged in, as these differed in several ways from more typical inquiry science activities. The univariable model of a rudimentary science experiment—an independent variable is manipulated and the effect on a dependent variable observed—is the staple of classroom introductions to the scientific method. In the real world, in contrast, outcomes are most often the consequence not of a single cause but of multiple factors acting in concert, a fact that practicing scientists are well aware of and take into account in both their theoretical models and empirical investigations (Sloman, 2005). The logic and execution of a univariable experiment represents at most one narrow slice of authentic scientific inquiry and arguably needs to be enriched by contextualization in a more authentic multivariable model (Kuhn, 2016a).

In addition to having students work within this more authentic multivariable context, we situate activity in the context of what students will see as a meaningful purpose and goal. Furthermore, we draw on social science content that students already have some familiarity with. Although students will feel they already know something about such topics (e.g., teen crime; Jewett & Kuhn, 2016), they likely will not know that they are the stuff of science.

What better way, then, to get them to appreciate its power and relevance? In the course of such activities, students ideally come to see how their (and others') beliefs about the phenomenon are subject to influence by means of application of a scientific method. If students make this discovery for one topic, it may occur to them that the same could be true for other topics, in time leading them to a deeper understanding of science as a practice.

We include the well-studied univariable control-of-variables (COV) strategy as one such foundational practice, but our focus in the present work is on the understanding and practices associated with the coordination of multiple factors as contributors to an outcome (Howard-Jones, Joiner, & Bomford, 2006; Kuhn, 2007; Kuhn & Pease, 2008; Kuhn, Pease, & Wirkala, 2009; Kuhn, Ramsey, & Arvidsson, 2015; Wu, Wu, Zhang, & Hsu, 2013), a critical development in scientific thinking that by comparison has received much less attention. In our earlier studies just cited, we repeatedly observed students who had well mastered the use of COV to identify each of several variables that affected the same outcome. When asked to then predict the outcome of a particular instance based on its standing on all of these variables, however, they typically referred to only a single variable as bearing the explanatory burden. Moreover, the particular variable invoked shifted from instance to instance.

Related to an understanding of multivariable causality is understanding that covariation between two variables need not be perfect in order for a relation to exist between them because effects of other factors, as well as measurement error, are likely to play a role (Kuhn, 2016a; Kuhn & Pease, 2008; Lehrer & Schauble, 2004; Masnick & Morris, 2008; Masnick., Klahr, & Knowles, in press). A data analysis tool for K–12 students, InspireData, we have found productive in promoting this mastery as it allows students to visually represent effects of multiple factors (Kuhn et al., 2015). Such representations enable them to achieve a multivariable understanding they then use to predict outcomes based on multiple variables, in so doing exercising another core purpose of the activity – drawing on evidence, rather than only their own beliefs, as a source of their inferences.

A further overarching dimension of scientific practice that our approach emphasizes is argumentation. As a key means of providing practice and developing argument skill, we engage students in scientific writing in the form of reports to a sponsoring foundation regarding their findings. As well as coordinating multiple kinds of evidence with claims, this activity included addressing challenges to students' claims, thus exercising skills of argument, counterargument, and rebuttal.

Our pedagogical method can be characterized as one of guided inquiry, designed to promote deep conceptual understanding of practices, rather than simply mastery of procedures, as students engage these practices in pursuit of a goal. The phases of the activity are segmented for students into a progression of tasks, with care taken to make clear the purpose and goal of each one and its purpose in relation to the larger objective. For example, with respect to COV, research has shown that students do better if they are guided to pose for investigation an appropriate question regarding the role of one variable at a time (Kuhn & Dean, 2005; Lazonder & Kamp, 2012), progressing through the variables in sequence.

Students are not given direct instruction as to strategies to apply to the component tasks; rather, attention is focused on the task goal and on their coming to recognize the weaknesses of inferior strategies they use in failing to achieve the desired goal, as a first step in devising ways to improve them. Critical to progress, we thus propose, is not simply acquiring new strategies but achieving metalevel awareness of the weaknesses of an initial or habitual approach, As a culminating activity, students reflect on how their final conclusions differ from their initially solicited beliefs about the roles of each of the set of identified factors in contributing to the outcome. Doing so leads to reflection on the task as a whole and on how their evidence-based conclusions provide knowledge central to achieving the best task outcome.

Following this intervention, we conduct a series of assessments of students' ability to extend these practices to new contexts. In addition, and addressing our central research question, assessments are included that probe students' understanding of science as a practice entailing the debate of alternative claims in a context of evidence. Students' performance on these assessments is compared to that of an equivalent group, taught by the same teacher, who delivered their regular science instruction during the intervention period.

METHOD

Participants

Participants were 48 students (equally divided by gender) drawn from three comparable 10th-grade biological science classes in a low-performing urban public high school in the northeast United States. Students were predominantly African American or Latino, with 86% eligible for free lunch and an additional 10% eligible for reduced-cost lunch. For 60% of students, primary home language was not English, with 18% designated as limited English proficient. The school consistently has failed to meet federally set goals for annual yearly progress, and in an assessment of college and career readiness, the school was found to be meeting only 20% of its performance targets.

Intervention

One class was randomly chosen to serve in the intervention condition and students drawn randomly from the other two classes served in a control condition. The intervention was administered over four 80-minute double class periods over a period of 10 days. During this period, control group students experienced their regular science curriculum. The classes had been identified by the school staff as equivalent in academic ability and performance as assessed by standardized tests. An initial paper-and-pencil assessment of the sort commonly used to assess the COV strategy (Jewett & Kuhn, 2016) was administered a month before the intervention began and showed the classes to be equivalent in performance in this regard.

The teacher for both intervention and control classes was the same classroom teacher who had regularly taught this biological science class at this grade level for the past 4 years. During the intervention period, the topic studied by the control group was photosynthesis. The teacher reported incorporating inquiry activities into instruction on this topic, consistent with her practice. These centered on the question of what factors affect rate of photosynthesis. Students were provided tables containing data on several variables (amount of water, amount of CO_2, amount of O_2, and temperature) and resulting rates of photosynthesis and asked to interpret the data.

In the intervention class, the activity was introduced by the teacher at the first session as follows, illustrated by an accompanying Powerpoint graphic:

> A new Astro-World Foundation, funded by some wealthy businessmen, wants to provide money for a space station. Groups of young people would live there for several months. Many young people have applied. The Foundation president needs to choose the best ones. So she asked some applicants to spend a week in a space simulator (picture is shown and function explained). She had background information about each applicant, and each one got a rating on how well they survived in the harsh conditions of the simulator. Some did fine; others okay, and some became sick and had to leave.

> Based on these records, she can decide which things are important to ask new applicants about and which ones aren't. Some of the factors, she noticed, made a big difference to how well an applicant did, some made a small difference, and some made no difference. She found out, for example, that body weight made no difference: Heavy people did as well in the simulator as light ones. But other things about people seemed to make a big difference in how well they did. So now, when she chooses final groups of astronauts to go on the real trips, she'll have a better idea what things to find out about applicants, so she can be pretty sure how an applicant will do and she'll be able to choose the ones who will do best.
>
> But, in order to be sure, she's asked for our help in analyzing their results. Which things are worth asking applicants about and which don't make any difference, like body weight? There are a lot of things that we can ask about but the foundation can't ask about everything. It would take too long. If we know what to ask applicants, we can choose the best team of astronauts.
>
> Here are four things that the foundation thought might make a difference to how well people do in the simulator: Fitness - does how well the person can run or do other exercises matter? - Family size - does the size of the family the person grew up in matter? - Education - does how much education a person has matter? – and Parents' ' health - does the health of the person's parents matter? All the applicants seem healthy, but maybe their parents' health might say something about how healthy they will turn out to be.
>
> Will you help figure out which things are worth asking the applicants about and which ones don't matter? Then you can predict how well they'll do and choose the best ones for the team. Later, you can compare your results with those of your classmates and see who chose the best-performing astronaut team.

Following this introduction, student pairs were formed and each pair asked to record on a form for this purpose which of the four factors they thought would and would not matter. A tally across the class was shown, and in a class discussion it was noted that opinions differ.

Control of Variables Phase. This phase was introduced by the teacher as follows: "These are only opinions and what someone thinks. Now, let's look at the data to find out what actually does matter and whether your hypotheses were right." A reminder of the larger purpose of the activity was then provided and was repeated periodically throughout the activity (a minimum of once per session): "Remember, the goal is to figure out what matters to how well people do in the simulator. Why do we want to know that? Because once we know what matters, we can predict how well people will do. That way, we can pick the best team."

Student pairs were then each provided a set of 24 cards, each containing an applicant's standing on the four factors and a blank space where they could record an applicant's performance rating in the simulator. Students were told that if they studied the records carefully they could determine which factors make a difference to performance and which do not. Pairs were reminded that they needed to agree before making decisions or drawing conclusions.

The only instruction provided to students as to how to proceed was the suggestion that they investigate one factor at a time. Students were invited to choose, from the set, the card(s) they would like to obtain outcome information for and to request these on a "data request form," in addition to explaining on the form what they would find out from examining their choice of case(s). After receiving and recording these outcomes on the chosen card(s), pairs then had the option either to reach a conclusion regarding the factor they had chosen for investigation or to postpone concluding and seek further evidence by repeating the preceding process, which they could do as many times as they wished until they felt ready to draw a conclusion. Once a pair was certain, they had reached a conclusion about a factor's status, they could enter it on a "draft memo" to the foundation director.

The teacher and classroom assistant circulated among students during this process and questioned the student pairs on their decision-making process, specifically asking them to justify their reasoning for claiming a factor as relevant or not. If a pair indicated they had drawn a conclusion without having a controlled comparison as evidence, the teacher or classroom assistant asked probing questions. These were designed to foster recognition of the weaknesses in the pair's investigative approach and, as a result, the inability to reach a definitive conclusion.

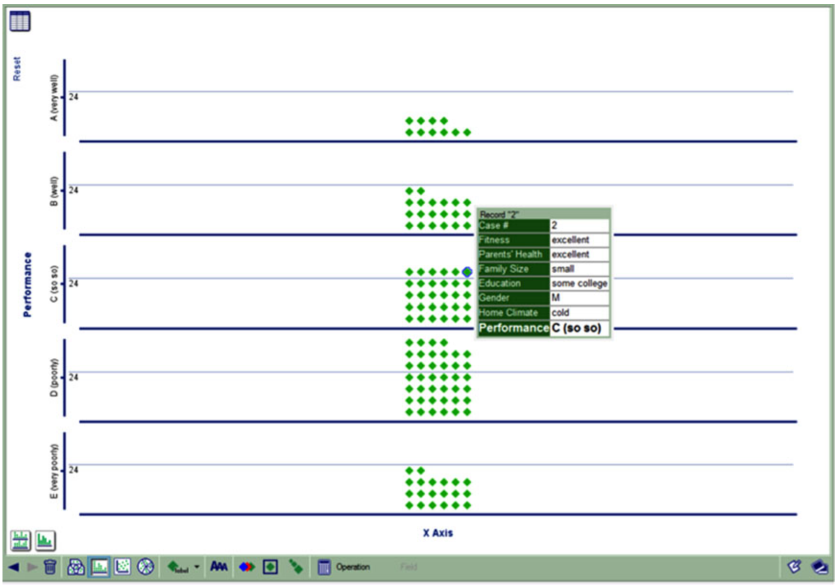
Students were not given advice on strategies or techniques, just probing of their claims and reasoning (e.g., "Couldn't it also be the difference in education that's leading to the different outcomes?"). In the case of valid conclusions, challenging probes were introduced (e.g., "Suppose someone disagrees with you and doesn't think that this factor makes a difference; what could you tell them to convince them?")

Once the majority of pairs had achieved three controlled comparisons showing fitness (a two-level factor) and education (the only three-level factor) effective and family size (a two-level factor) ineffective, pairs completed their final memo to the foundation director. In this memo they indicate which factors applicants should be asked about and which they should not and justify their recommendations with evidence from their investigations.

Multivariable Coordination Phase. The class at this point was ready to transition to the next and principal phase of the intervention, in which students represented and reasoned about the influence of all of the factors operating at once. The skills involved in this multivariable coordination have not been as extensively studied as have univariable experimental design and COV skills and hence warrant some explication. A fundamental understanding students must acquire is that covariation between variables need not be perfect in order for a relation to exist between them because effects of other factors, as well as measurement error, are likely to play a role. This is easiest to understand when the distributions of outcomes for two levels of an independent variable do not overlap (Masnick et al., in press). Commonly, however, such distributions do overlap, such that some of the outcomes for instances of one level of the independent variable will be identical to some of the outcomes for instances of a different level of the independent variable (see Figure 3 below for an illustration), even though the overall outcome distributions for the two levels differ. In this case, it must be recognized that the two distributions overlap because other factors, in addition to measurement error, likely are making their own contributions to outcomes.

The objective of this phase of the intervention was to develop these understandings, making use of InspireData as a tool for this purpose. Students were introduced to this phase thusly, "So far all of our conclusions have been based on comparing just two or maybe three cases. We would be more sure of our conclusions if we looked at more than two or three cases at a time. We have a way to do that." Students were then introduced to the representation of their data using charts generated by the program InspireData and told, "All of the cases that you and your classmates have looked at before are here." It was explained that each diamond represents a case and that the identity of that case can be seen by hovering over the diamond (Figure 1).

It was then illustrated that charts can be generated that separate cases into different categories, for example, in the display shown (Figure 2), only those cases in which the



Record "2"

Case #	2
Fitness	excellent
Parents' Health	excellent
Family Size	small
Education	some college
Gender	M
Home Climate	cold
Performance	C (so so)

Figure 1. InspireData chart showing all cases. [Color figure can be viewed at wileyonlinelibrary.com]

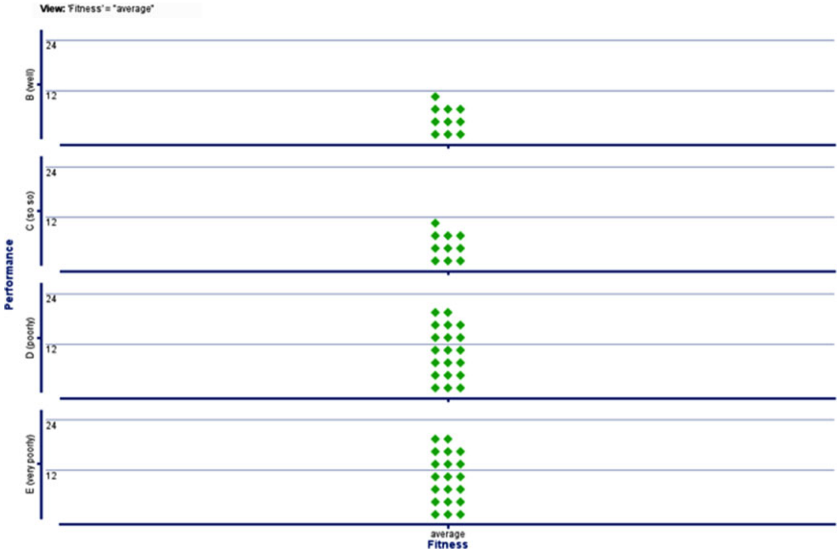


Figure 2. InspireData chart showing only cases with average level of fitness. [Color figure can be viewed at wileyonlinelibrary.com]

applicant's fitness was average rather than excellent are shown. Students were then asked why it was that these applicants all of the same fitness level showed a range of performance outcomes. With a little prompting, students were able to generate the response that other factors besides fitness were contributing to the outcomes.

Students were then shown a third display (Figure 3) in which all levels of the fitness variable are included. They were asked to draw conclusions about whether the fitness variable makes a difference to applicants' performance. Given the ability to see more data

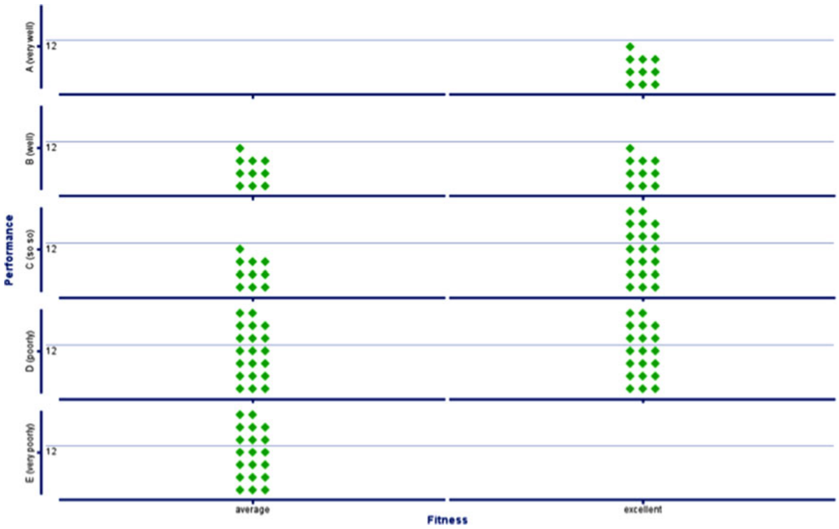


Figure 3. InspireData chart for the fitness factor. [Color figure can be viewed at wileyonlinelibrary.com]

at once, students were asked to reflect on whether they reached the same conclusions as they did earlier when comparing individual cases presented on cards.

Students were then provided InspireData charts for each of five factors, four introduced previously and one new one (home climate, a noninfluencing factor), each of the same form as Figure 3, showing outcomes for all levels of the factor. Students were reminded these charts would give them an opportunity to verify their earlier conclusions. In their pairs, they did this and then wrote memos to the foundation director confirming their earlier conclusions based on a larger sample or revising their conclusions if they thought necessary.

As in the previous phase of the intervention, prompts were introduced in the case of both correct and incorrect conclusions, for example, “Suppose someone disagrees with you and doesn’t think that this factor makes a difference; what could you tell them to convince them?” Once per class session (typically at the end of the session) a whole-class discussion occurred, using one pair’s work as an example.

Application Phase. Students were told that now that they had reached final conclusions, they could try using them to evaluate a new set of applicants. They would then be able to select a set of five applicants to be chosen for the astronaut program and compare their choices to those of their classmates. Students were told that they could select up to four factors about the new applicants that they could receive information on. As students were selecting the factors, the adult reminded them to review the InspireData chart and consider whether knowledge of status on this factor would be informative as to outcome.

Information about 10 new applicants on four factors (including one noninfluencing one, whether or not it was asked for), data for each applicant appearing on a separate card and cards presented one at a time. Students completed the first prediction with guidance and pairs then worked independently. In addition to making each prediction, they were asked for each one, “Which of the four factors you have data on mattered to your prediction?” Students were encouraged to review the InspireData charts to double check their decisions or when there were disagreements within the student pair.

A final discussion occurred when pairs made their selections of the five top-rated candidates and shared these with the whole class. This discussion included remembering the beliefs they had initially held about the factors and noting that they would not have chosen the same applicants before and after the analysis they had conducted.

Postintervention Assessment

Postintervention assessments were administered individually and were not conducted until 5 weeks (35 days) after the conclusion of the intervention, to assess maintenance of new achievements as well as generalization to new content. Assessments for students in the control group were conducted during the same time period.

Maintenance of Skills in Experimental Design and Control of Variables. This delayed assessment of skills achieved during the intervention was situated within the astronaut scenario and extended to two new variables—height and strength. The student was asked to design an experiment in which to test if first height and then strength had an effect on astronaut applicants' performance scores. (Because it involved the intervention content, this assessment was not administered to students in the control condition.)

Extension of Skills in Experimental Design and Control of Variables to New Content. The skills assessed were comparable to those assessed in the preceding maintenance task but with new content unrelated to the intervention, to assess generality of gains and eliminating any advantage of the experimental group in terms of familiarity with the content. The assessment contained three items, all visually represented. For example, one of the items stated that New York City was designing new cars for their subways. The test had four study design conditions each having two subway cars in them to evaluate. Within each condition, the subway cars varied with respect to car size and number of wheels. The student was to select the best study design to test if car size made a difference to how fast the subway train would run.

Extension of Multivariable Analysis and Prediction Skills to New Content. Developed by Kuhn et al. (2015), this multivariable analysis and prediction task presents simplified but authentic data on four factors found to have an effect on average life expectancy (employment, family size as strong contributors, and education and home climate as moderate contributors) across different countries and one noncontributing factor (country size), Students are shown a chart containing a simplified graph for each feature, illustrated in Figure 4 for the employment factor. Instructions were as follows.

> Some people live very long lives. Others die at an early age. What makes the difference? Life Expectancy (**LE**, for short) is the term for how long people on average are expected to live. **LE** differs greatly across different countries. Some countries have a much higher average **LE** than others. What causes the differences? Here are some possibilities that studies suggest. One is employment. As the chart below shows, countries where people have high employment levels have a higher average life expectancy than countries where employment is low (many people are without jobs). As you will see, the chart shows that employment makes a very big difference to **LE.** As you also see, family size, like employment, makes a very big difference. Smaller families on average mean higher **LE.** Other things, like education and climate, make a smaller difference. And some, like country size, seem to make no difference at all—**LE** overall is about the same for small and large

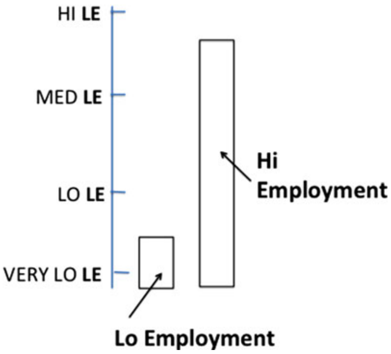


Figure 4. Illustration of information provided for each factor in life expectancy task.
Note. The figure for family size (large vs. small) parallels that for employment. The figures for education and climate depict lesser effects (low vs. medium LE), and the figure for country size depicts no effect (both small and large countries show an average LE between low and medium).
[Color figure can be viewed at wileyonlinelibrary.com]

> countries. Now your task is to make some predictions about **LE** for different countries. You can look back at the charts when you want to. For each country, predict the average **LE** you think that country will have.

The student is asked to predict life expectancy of nine additional countries based on information provided about the country's status on the identified factors (on the four-point scale of Very Lo, Lo, Med, and Hi). The task also asks respondents to indicate which factors they considered in their prediction ("What made you choose this outcome?").

Conducting and Understanding Argumentation as a Practice. The remaining components of the delayed postintervention assessment are adapted from ones reported on by Kuhn et al. (2015) and Kuhn (2016b). They build on a single content theme and progress substantively from constructing a claim to using evidence to support and weaken claims, and, finally, to the task having the most explicit epistemological dimension—identifying appropriate means of reconciling contrasting claims.

1. *Constructing claims for investigation.* Students read, "The Public Health department of Portland, Ohio has noticed that the percentage of residents diagnosed with cancer is much higher in the inner city than in the outlying neighborhoods. The department is undertaking a study to find out why there are more people getting cancer in the inner city than the outlying area."

The student was asked what they would do to investigate the matter. Assessed was whether students would make a causal claim and propose a means of evaluating it.

2. *Constructing a counterargument.* Students read, "John thinks it's because people in the city go to tanning salons. What might someone say to John, if they think that John was wrong?" (This and the next task, note, require the more demanding skill of identifying arguments and evidence that weaken, rather than support, a claim [Kuhn & Moore, 2015], given it was the *weaken* skill that was emphasized during the intervention.)

3. *Assessing evidence strength in relation to a claim.* Students read, "Which of these four types of evidence would be strongest to show that John was wrong?"
   (a) Air pollution is more likely a cause of cancer in the city.
   (b) Many people outside the city also go to tanning salons and don't get cancer.
   (c) Many people who don't go to tanning salons also get cancer.
   (d) There are more tanning salons outside the city than in the city.

The best answer is Option B, evidence showing the presence of the antecedent and absence of the outcome, indicating that the antecedent is not sufficient to cause the outcome. Option C ignores the possibility of alternative causes sufficient to produce the outcome; rejection of Option C, note, requires understanding of the multivariable causality emphasized in the intervention, i.e., that multiple causes can contribute to an outcome (additively or alternatively). Option A is an alternative assertion but does not itself constitute evidence, and Option D does not bear on the claim.

4. *Recognizing and reconciling contrasting claims.* Two contrasting claims are presented, and the student is asked to interpret the discrepancy. Students read, "To investigate why people living in the city are getting cancer more often than people who live outside the city, you tested and found out that air pollution was worse inside the city than outside. You wrote a report of your findings to the Health Department director, telling her that air pollution was a likely cause of the increase in cancer. She also received a report from another person she hired. This report said that a likely cause of cancer increase was not enough stores in the city for people to buy healthy fruit and vegetables that lower risk of cancer. The director isn't sure what to conclude and she has written you asking for advice. What would you write back? Give her the best advice you can."

This final assessment task, newly constructed for the present study, addresses most directly students' epistemological understandings regarding scientific claims, inviting students to explain how contrasting scientific claims should be evaluated. Unlike the first, the second claim in the scenario, note, is not accompanied by any evidence, to assess whether students would detect and identify this difference as relevant to judging their relative strength.

RESULTS

Experimental Design and Inference

Maintenance of COV Skills. The data consisted of the two experiments students designed involving the intervention content, to test if first height and then strength had an effect on astronaut applicants' performance scores. Performance is summarized in Table 1. The two

TABLE 1
Performance of Intervention Group on COV Maintenance Task

Never varied focal variable	0%
Varied focal variable only sometimes	12%
Consistently varied focal variable, but inconsistent control of other variables	12%
Consistent controlled comparison	76%

Note. Entries indicate percent of participants showing. $N = 24$.

TABLE 2
Performance of Intervention and Control Groups on COV Skills with New Content

Performance	Intervention Group (%)	Control Group (%)
Consistently constructed a controlled comparison	96	75
Provided appropriate justification for comparison	75	46

Note. Entries indicate percentage of participants showing. $N = 24$ for the control condition, and 24 for the intervention condition. The difference between groups is statistically significant, $p = .041$ for construction and $p = .038$ for justification, Fisher's exact test.

TABLE 3
Mean Number of Predictions (of nine) for Which Contributing and Noncontributing Factors Were Reported as Having Influenced Prediction

	Intervention Group	Control Group
Contributing factors	7.68	5.22
Noncontributing factor (country size)	0.63	2.67

items could be objectively scored based on the relation of the cases chosen for comparison with respect to the two key criteria for mastery of control of variables (Kuhn, 2016a): Was the focal variable under investigation varied and were other variables controlled (held constant)?

Extension of COV Skills to New Content. As seen in Table 2, when compared to the control group on an assessment unrelated to the intervention, the intervention group outperformed the control group both in constructing a controlled comparison and in providing an appropriate justification.

Multivariable Analysis and Prediction

Students' predictions were compared to a model of correct prediction (weighting the two moderate factors as contributing half as much to outcome as the two strong factors) for each of the nine prediction items. Against this model, the number of correct choices among the four outcome prediction options averaged 6.42 ($SD = 1.67$) for the intervention group and 3.65 ($SD = 1.46$) for the control group, $t = 5.51$, $p < .001$ (with chance correctness in choice among the four outcome prediction choices 25%, or a score of 2.25 across the nine items). Among the intervention group, modal response was a correct prediction for 23 of the 24 students (96%), compared to 27% for the control group.

Table 3 summarizes students' explicit representations of the factors they maintained had influenced each of their predictions.

In terms of individual performance patterns, correct attribution of all effective factors was shown by intervention students on an average of 6.00 of the nine items, compared to an average of 1.60 items by the control group, $t = 4.37$, $p < .001$, reflecting the greater consistency of attributions across items by intervention students. In the intervention group, 15 of 24 students (63%) chose the four effective factors consistently across all nine predictions, whereas in the control group only one student (4%) did so (a significant difference, $p = .0009$, Fisher's exact test).

Argumentation

Constructing claims for investigation. In response to the request to design an investigation to explain the differing cancer rates in the two locations, control group participants all simply offered a single-factor explanation, without identifying a claim to put to empirical test. Here are two examples:

> The inner city is full of more people and it might be the reason why, because when there are more people you get more diseases.
>
> The people in the city have more cancer then the people out in the outlying area because they have more things to give people cancer than the outlying people have more natural things.

Among the intervention group, 29% of responses were of a similar nature. The remaining 71% of intervention group students proposed an empirical investigation having a comparative design (a significant difference from the control group, Fisher's exact test, $p = .0001$). Following are two examples:

> People in the city may be breathing in toxic air. So take people from each neighborhood and switch them with the other. Study change in people.
>
> If there is a great amount of pollution the more people will get sick. I can test by having two groups of healthy people. The 1st group of healthy people stay in the inner city, while the other group of people go to the outlying city (not polluted place).

Constructing a Counterargument. This component of the assessment examined whether when the causal claim was provided, the student could identify a counterargument. Responses were of three types:

(a) showing the failure of the alleged cause to produce the outcome (tanning salon use does not lead to cancer),
(b) securing (unspecified) evidence to establish the true cause, and
(c) making a counterclaim of an alternative causal agent (it is X that causes cancer).

As seen in Table 4, the modal response in the intervention group was to weaken the claim by producing evidence falsifying it. Among the control group, the modal response was to make an alternative claim, leaving the original claim unaddressed. Those in the intermediate group (Establish true cause) recognized a need for evidence but did not identify what it would be.

Assessing Evidence Strength. As seen in Table 5, in choosing the strongest of four types of evidence presented, all students recognized option D as irrelevant to the claim, but the two groups differed in their choices among the remaining three options. The intervention group was most likely to select Option B, the most powerful evidence in falsifying a causal claim. The control group was most likely to select Option C.

TABLE 4
Performance of Intervention and Control Groups on Counterargument Construction

Type of Counterargument Proposed	Intervention Group (%)	Control Group (%)
Show cause fails to produce outcome	50	13
Establish true cause	33	8
Make counterclaim of an alternative causal agent	17	79

Note. Entries indicate percentage of participants showing. $N = 24$ for the intervention group and 24 for the control group. This difference between the two groups in proportion showing the strongest counterargument type is significant, $X^2(1) = 7.85$, $p = .005$.

TABLE 5
Performance of Intervention and Control Groups on Assessing Evidence Strength

Response Choice	Intervention Group (%)	Control Group (%)
A. Air pollution is a more likely cause of cancer in the city	12	21
B. Many people outside the city also go to tanning salons and don't get cancer	54	21
C. Many people who don't go to tanning salons also get cancer.	34	58
D. There are more tanning salons outside the city than in the city.	0	0

Note. Entries indicate percentage of participants showing. $N = 24$ for the intervention group and 24 for the control group. The difference between groups in proportion choosing Option B was significant, $X^2(1) = 5.69$, $p = .017$.

Reconciling Contrasting Claims. As this was a new task, two of the authors examined responses from a larger sample drawn from the same population and in an iterative process developed a coding scheme designed to capture whether respondents recognized the discrepancy between the claims as the issue to be addressed and, if so, how they addressed it. This coding scheme, shown in Table 6, was then applied to responses of the present sample. A percentage agreement of 83% was achieved. Types of responses offered by the two groups and the percentages showing each type appear in Table 6.

As reflected in Table 6, not all students addressed the question posed—how to account for and resolve the contrasting claims the director was left with. Among control group students, the majority did not directly address the epistemological question regarding contrasting claims and instead addressed only how the director should use the information she had to address the practical problem (Option d; Table 6); in other words, they adopted an engineering rather than scientific perspective. The following are examples.

> The advice I can give is to sell more fruit and vegetables that can lower the cancer rate because if that's the problem which is less fruits and vegetables you should sell more.

Sometimes both potential factors were mentioned, but the approach remained achieving a practical solution:

> Make more stores that sell fruit and vegetables and lessen the air pollution. That way people can eat healthier and aren't breathing in toxic fumes.

Among students in the remaining three categories, who addressed the two possible causes as scientific claims, a few confined themselves to expressing an opinion (Option c; Table 6) as to their merits. For example:

> I do think it's the air pollution because in the city there's lots of cars and stores. I don't think it would be not enough healthy fruit and vegetables stores, because if you want it you can go somewhere else to buy it.

A third group of students said that further investigation (Option b; Table 6) would be needed to evaluate the claims and made recommendations. For example,

> You can get two people and they did have cancer, and one has to be inner city and other one has to live in neighborhood. And the inner city person has to eat the neighborhood's food and the neighborhood person have to eat the inner city's food. Then you can understand that if the inner city person got cancer then it's because of the air and the neighborhood person got cancer that because of the food. Then you will see why the inner city's people get cancer and neighborhood people don't.

Finally, only students from the intervention group expressed awareness that the two claims are *not contradictory* (Option a; Table 6) and both could be correct—50% did so. For example,

> It could be both air pollution and stores within the city that are highly effective of causing this situation.

Or

> I think both these reports are correct because both make sense and could be just a few of the reasons to why any have cancer. Studies can have not only one but many outcomes.

TABLE 6
Performance of Intervention and Control Groups on Reconciling Contrasting Claims

Reconciliation Strategy	Intervention Group (%)	Control Group (%)
a. Resolve apparent scientific conflict by recognizing contrasting claims as compatible	50	4
b. Advocate further investigation	21	17
c. Express opinion regarding strength of claim(s)	17	33
d. Advocate practical solutions rather than addressing causal claims (Engineering strategy)	12	46

Note. Entries indicate percentage of participants showing *N*s were 24 and 24 for intervention and control groups, respectively. Groups differ significantly with respect to both overrepresentation of the intervention group in the first category, $X^2(1) = 12.76$, $p = .0004$, and overrepresentation of the control group in the final category, $X^2(1) = 6.45$, $p = .011$.

A few students combined this awareness of multivariable causality with suggestions for further investigation. For example,

> You can analyze both reports. For the lack of healthy fruit stores report, you can ask someone to search for information of another city that has enough healthy stores and analyze if those cities have the same problem with cancer or not. If it seems that their information they give you shows that fruits are very necessary to low the risk of cancer then you can conclude that both air pollution and lack of healthy fruit stores are factors that increase the risk of cancer in the city.

DISCUSSION

The present work corroborates the studies cited earlier in establishing that extended engagement in inquiry in a multivariable context promotes understanding and skill in coordinating multiple causes contributing to an outcome. Here we demonstrate this result following an intervention of considerable length yet significantly shorter than the multiyear intervention reported on by Kuhn et al. (2015) with students several years younger. The contrasts between intervention and control groups on the life expectancy posttest assessment of multivariable causal understanding are particularly notable. In contrast to control group students, a majority of intervention group students recognized that standing on all operative variables needed to be considered and were able to coordinate their effects with consistency and correctly predict outcomes.

The case for the importance of multivariable skills and understanding was made earlier and in the studies cited previously. If students are to engage in activities having characteristics of authentic science practice they must be ones in which multiple variables play a role in determining outcomes. Not emphasized in earlier work, however, is the connection between understanding of multivariable causality and argumentation. This connection is highlighted in the assessment included in this study in which students must judge which of four options constitutes the strongest evidence against a particular causal claim. If the possibility of multiple causal contributors to an outcome is not recognized, a respondent will wrongly judge the mention of a new possible cause as more damaging evidence than it in fact is against the causal efficacy of the initially named factor. This error in causal reasoning has implications for argumentive discourse. If a single cause is regarded as sufficient to bear the explanatory burden of accounting for an outcome, alternative causes will be seen as contradictory: Either my cause or your cause must be the correct one. An affective component enters in and reasoning becomes motivated by allegiance to one's preferred cause, with the alternative cause seen as threatening to replace it, when in fact it may be unnecessary to choose between the two.

Intervention students' favoring of option B—the cause fails to produce the effect—as the most decisive counterargument evidence in the Assessing Evidence Strength component of the argumentation posttest assessment is particularly notable in comparison to the performance of a sample of average adults (Kuhn, 2016b), among whom only a quarter chose that option. The present intervention participants also appear to have gained a respect for the power of evidence in rejecting the option that invokes simply an alternative cause as a counterargument, an option that leaves the initial claim unexamined. Among our control group, in contrast, students most often chose the option of an alternative cause producing the outcome as the strongest in weakening the claim of causal power of the original cause, despite its nondefinitive and at best indirect implications in this regard.

The major research question we posed in the present study is how engagement in broad scientific practice emphasizing multivariable investigation, analysis, and argumentation stands to influence epistemological understanding regarding science practice, specifically understanding of science as constituting the debate of contrasting claims in a framework of available evidence. Essential to understanding and valuing argument is the slow-to-develop epistemological understanding of claims as neither nonrevisable facts nor unconstrained opinions but rather judgments requiring evaluation in a framework of alternatives and evidence (Greene, Sandoval, & Braten, 2016; Moshman, 2015). Key is therefore the understanding of the critical role of counterargument and of evidence as means of examining and evaluating a claim (with evidence capable of weakening as well as supporting claims). Without this conceptual underpinning, we cannot expect students to grasp the practices of science that science educators increasingly have regarded as a key dimension and objective of science education (Sandoval, 2014).

Causal analysis, argument, and epistemological understanding bear close connections to one another, and, we have suggested, develop in ways that are likely mutually reinforcing. Other authors who have highlighted this link between practices and epistemological understanding (Duschl, 2008; McNeill, 2011; Ryu & Sandoval, 2012) have worked largely with students younger than those in the present study and hence have focused on what Sandoval (2005) has called practical epistemologies, i.e., epistemological understandings that are revealed indirectly by how students behave in science inquiry activities.

The intervention group in the present study made clear progress in this respect. They showed understanding with respect to both counterargument and evidence as key components of science as argument. Relative to the control group, intervention students demonstrated clear advantages in (a) constructing a claim amenable to empirical investigation, (b) constructing a weakening counterargument, and (c) identifying types of evidence in terms of their counterargument power.

It is, however, our final postintervention assessment task that asks students explicitly how differing claims are to be evaluated and reconciled, thus providing the most direct evidence with regard to their understandings of the epistemological foundations of science practice. Intervention students' achievements, recall, were evident 5 weeks after the intervention concluded. In contrast, performance by control group students corroborates that even by high school age, and even in the context of a science class, an engineering frame dominates and students are not naturally disposed to treat diverging scientific claims as warranting scrutiny and potential reconciliation (Barzilai & Eshet-Alkalai, 2015; Schauble, Klopfer, & Raghavan, 1991). Here again multivariable causality understanding appears to have played a role in intervention students' understanding, and again we see a particularly decisive effect of the intervention, with control group students failing to exhibit the understanding intervention students did regarding the central role that evaluating contrasting claims plays in the practice of science.

What led intervention group students responding to this final task to better recognize the presence of contrasting scientific claims as a concern demanding attention? As is always the case, it is impossible to be sure which specific components of a multicomponent experience were most critical to its outcomes. Our conjecture, however, is that a key factor was the emphasis during the intervention on counterargument and evidence to weaken claims (*"Suppose someone disagrees with you . . . "*). Reasoning of this sort fosters awareness that claims are subject to scrutiny and potentially to falsification—an understanding foundational to the epistemology of science.

We conclude with acknowledgement of two limitations of the intervention reported on here—one pertaining to process and the other to content. We earlier characterized authentic scientific practice as involving communities of practice. Members of scientific communities develop shared norms that govern their activities and progress, and the evolution of such norms can be observed as well in young people who collaborate in an intellectual community of peers (Kuhn & Zillmer, 2015). Although extended over 2 weeks of engagement that took place most often with a partner, the present peer collaboration was not long enough or varied enough in form to clearly observe these norms, e.g., of what counts as evidence, emerge.

The intervention was also limited with respect to scientific content. We proposed that social science content is not only worthy as content for scientific investigation but perhaps particularly so for the population we worked with. Nevertheless, the variables students investigated were simplistic ones and the relations students identified among them similarly oversimplified. We advocate this "content-lean" context as the right place to start with students in introducing them to the demanding conceptual understandings we sought to help them develop. Yet, certainly, once some success in this respect is achieved, there is every reason to proceed to examine richer scientific content models. Following this path, there is hopefully diminished likelihood of students confusing a theoretically rich account of plausible mechanism connecting cause and effect with empirical evidence that addresses its correctness (Kuhn & Katz, 2009).

An additional limitation of the present work that warrants mention is the restricted population of student participants. There are multiple reasons to work with a chronically underachieving, underserved population. Here we perhaps need only say that if the effort is successful, it is highly likely to be at least as successful among more advantaged populations.

Most broadly, the present results support the view introduced at the outset that scientific practices are interrelated, as are understandings regarding science practice, and students stand to benefit from engaging in science practices not as isolated procedures but as an integrated whole. These interrelated components contribute to the sense that can be made of the whole. The most significant benefit to students will come from their seeing how they fit together and hence able to see what makes the enterprise they are part of worth valuing.

REFERENCES

Barzilai, S., & Eshet-Alkalai, Y. (2015). The role of epistemic perspectives in comprehension of multiple author viewpoints. Learning and Instruction, 36, 86–103.

Berland, L. K., & Hammer, D. (2012). Framing for scientific argumentation. Journal of Research in Science Teaching, 49, 68–94.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. Review of Research in Education, 32, 268–291.

Ford, M. J. (2012). A dialogic account of sense-making in scientific argumentation and reasoning. Cognition and Instruction, 30, 207–245.

Greene, J., Sandoval, W., & Braten, I. (Eds.) (2016). Handbook of epistemic cognition. New York, NY: Routledge.

Howard-Jones, P., Joiner, R., & Bomford, J. (2006). Thinking with a theory: Theory-prediction consistency and young children's identification of causality. Instructional Science, 34, 159–188.

Jewett, E., & Kuhn, D. (2016). Social science as a tool in developing scientific thinking skills in underserved, low-achieving urban students. Journal of Experimental Child Psychology, 143, 154–161.

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

Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only challenge. Science Education, 91, 710–726.

Kuhn, D. (2010). Teaching and learning science as argument. Science Education, 94, 810–824.

Kuhn, D. (2016a). What do young science students need to know about variables? Science Education, 100, 392–403.

Kuhn, D. (2016b). A role for reasoning in a dialogic approach to critical thinking. Topoi. doi 10.1007/s11245-016-9373-4

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

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? Journal of Experimental Child Psychology, 103, 386–394.

Kuhn, D., & Moore, W. (2015). Argument as core curriculum. Learning: Research and Practice, 1, 66 – 78.

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? Cognition and Instruction, 26, 512 – 559.

Kuhn, D., Pease, M., & Wirkala, C. (2009). Coordinating effects of multiple variables: A skill fundamental to causal and scientific reasoning. Journal of Experimental Child Psychology, 103, 268 – 284.

Kuhn, D., Ramsey, S., & Arvidsson, T. S. (2015). Developing multivariable thinkers. Cognitive Development, 35, 92 – 110.

Kuhn, D., & Zillmer, N. (2015). Developing norms of discourse. In L. Resnick, C. Asterhan, & S. Clarke (Eds.), Socializing intelligence through academic talk and dialogue. Washington, DC: American Educational Research Association.

Lazonder, A., & Kamp, E. (2012). Bit by bit or all at once? Splitting up the inquiry task to promote children's scientific reasoning. Learning and Instruction, 22, 458 – 464.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. American Educational Research Journal, 41, 635 – 680.

Lehrer, R., & Schauble, L. (2015). The development of scientific thinking. In L. Liben & U. Mueller (Vol. eds.) & R. Lerner (Series ed.), Handbook of child psychology and developmental science, Vol. 2: Cognitive process(7th ed.). Hoboken, NJ: Wiley.

Manz, E. (2014). Representing student argumentation as functionally emergent from scientific activity. Review of Educational Research, 1 – 38.

Masnick, A., & Morris, B. (2008). Investigating the development of data evaluation: The role of data characteristics. Child Development, 79, 1032 – 1048.

Masnick, A., Klahr, D., & Knowles, E. (in press). Data-driven belief revision in children and adults. Journal of Cognition and Development.

McNeill, K. L. (2011). Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year. Journal of Research in Science Teaching, 48, 793 – 823.

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

Moshman, D. (2015). Epistemic cognition and development: The psychology of justification and truth. New York, NY: Psychology Press.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. Journal of Research in Science Teaching, 41, 994 – 1020.

Ricco, R. (2015). The development of reasoning. In L. Liben & U. Mueller (Vol. eds.), R. Lerner (Series ed.), Handbook of child psychology and developmental science, Vol. 2: Cognitive process (7th ed.).Hoboken, NJ: Wiley.

Ryu, S., & Sandoval, W. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. Science Education, 96, 488 – 526.

Sandoval, W. (2005). Understanding students' practical epistemologies and their influence. Science Education, 89, 634 – 656.

Sandoval, W. (2014). Science education's need for a theory of epistemological development. Science Education, 98, 383 – 387.

Sandoval, W., Sodian, B., Koerber, S., & Wong, J. (2014). Developing children's early competencies to engage with science. Educational Psychologist, 49, 139 – 152.

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

Sloman, S. (2005). Causal models: How people think about the world and its alternatives. New York, NY: Oxford University Press.

Smith, C., Maclin, D., Houghton, C., & Hennessey, M. (2000). Sixth-grade students' epistemologies of science: The impact of school science experiences on epistemological development. Cognition & Instruction, 18, 349 – 422.

Wu, H.-K., Wu, P. H., Zhang, W. X., & Hsu, Y. S. (2013). Investigating college and graduate students' multivariable reasoning in computational modeling. Science Education, 97, 337 – 366.