Now They See the Point: Improving Science Reasoning Through Making Predictions

Christian D. Schunn (schunn@gmu.edu) Christine J. O'Malley (comalley@gmu.edu)
Department of Psychology; George Mason University
Fairfax, VA 22030 USA

Abstract

Previous research on scientific reasoning has found that many students find it difficult to think about the theoretical level when asked to design experiments. Two studies are reported that explore whether forcing students to make predictions before running their experiments improves their scientific reasoning performance. Both studies find that, if they do make predictions, students become more focused on the theories they are being asked to test. The students become more likely to make conclusions about the theories under test and they design experiments more relevant to the theories under test.

Introduction

Science education is a core component of education throughout the industrialized world, and the ability to reason scientifically is a generally valued skill. Nevertheless, relatively little is known about the details of how students become good scientific reasoners. There is one clear fact, however, about the developmental process that has been frequently documented: students do not come naturally to many aspects of scientific reasoning, and it is not easy to teach those skills (e.g., Kuhn, 1989; Lehrer, Schauble, & Petrosino, in press; Schauble, 1990). Even towards the end of a students' college education, many basic scientific reasoning skills are weak or missing (Schunn & Anderson, 1999, In press).

Developing an understanding of what can improve scientific reasoning skills is an important problem for cognitive science. It tends to involves many disciplines of cognitive science because it is a difficult problem that requires resolving: 1) what it means to reason scientifically (philosophy, history), 2) what cognitive processes are involved (psychology), 3) what kinds of interventions are successful (education), and 4) constructing complex computer environments that model and support it (computer science). What makes it an inherently cognitive science-like problem is that these four components are intimately interconnected.

This paper explores one simple method for improving scientific reasoning: forcing students to make predictions before running an experiment. It presents two empirical studies conducted in the psychology laboratory (as opposed to a classroom situation). The studies compare the scientific reasoning performance of students forced to make predictions before each experiment with students not asked to make predictions before each experiment. Before turning to the empirical studies, we will provide additional background on this particular issue.

To make a prediction, one needs a hypothesis or theory. Science textbooks generally recommend that one should always have a hypothesis before running an experiment. However, philosophical, historical, and, more recently, psychological accounts of science agree that one need not always have a hypothesis before running an experiment (see Okada and Shimokido, in press, for a review).

We acknowledge that there are plenty of situations in which people do not have a theory before conducting the experiment (Klahr & Dunbar, 1988). A central aspect of doing science, however, is the development and testing of formal theories—unifying or explanatory accounts that sit at a level above simple beliefs about the effects of particular variables.[1] Thus, having a theory to test is a common and important situation.

A separate question (and the one we examine) is whether one should always make concrete predictions before running the experiment *when one does have a theory to test*. There are plenty of situations in which one does have a theory to test. What role do explicitly made predictions serve in those situations?

Several recently developed science education computerized training environments have components that prod students into making predictions before running the experiments (e.g., Loh et al., in press; White, 1993, 1995). While these systems as a whole have been demonstrated to be effective, the value-added of the prediction-making component of these complex systems has not been tested in isolation. Thus, little is known from that research about whether forcing predictions actually improves reasoning.

There are several reasons to think that making predictions will help scientific reasoning. First, making predictions may remind students to focus on the theories that they are supposed to be testing. Schunn and Anderson (1999, In press) found that even undergraduates pay little attention to the theories they are supposed to be testing and instead simply explore the effects of different variables.

Second, making predictions may lead participants to consider alternative theories, and thus design experiments that more uniquely target the theory under test. On a related point, Koehler (1994) found that generating ones own hypothesis rather than being given the hypothesis leads to more accurate evaluations of the likelihood that the hypothesis is correct (however, see Schunn and Klahr (1993) for the exact opposite finding).

[1] We will use the term *theory* to refer to these general accounts and the term *hypothesis* to refer to beliefs about particular concrete variables. For example, ACT-R (Anderson & Lebiere, 1998) or SDDS (Klahr & Dunbar, 1988) are theories; "making predictions should improve reasoning" is a hypothesis.

There are also several reasons to think that making precise predictions will hurt scientific reasoning. First, it could be that making predictions would direct students away from the theoretical level that they are supposed to be testing and instead focus on simple empirical effects of particular concrete variables.

Second, getting students to make precise predictions could push students into an engineering rather than scientific mode (Schauble, Klopfer, & Raghavan, 1991; Tschirgi, 1980). In other words, it could lead students to focus on how to produce a particular outcome rather than on finding out why certain outcomes occur. As a variant of this theme, focusing on concrete predictions might lead students adopt a goal of trying to maximize their prediction accuracy (i.e., see how well they can predict outcomes). This new goal could be seen as a kind of engineering goal that is potentially at odds with the scientific goal of testing the theory.

In sum, there is a general belief that making predictions is important to scientific reasoning, possible reasons for it to help or hurt scientific reasoning, and little evidence one way or the other. We examine the role of making predictions on scientific reasoning in two different situations: when students are designing an experiment to choose between two alternative theories (Study 1); and when students are designing an experiment to test only one given theory (Study 2).

The Simulated Psychology Lab

To examine the influence of making predictions on scientific reasoning skills we selected a real scientific question from cognitive psychology: what is the cause of the spacing effect in memory? The spacing effect itself is intuitively understood by undergraduates—that spaced practice produces better memory performance than massed practice (i.e., cramming is bad). The advantage of using this particular question is that it is relatively easy to explain to undergraduates without the use of complex domain-specific jargon and yet it is an authentic scientific problem rather that a toy problem. Recent work in the psychology and education of science suggests that it is important to use realistically complex problems (Chinn & Malholtra, in press).

As we noted earlier, not all situations require a theory to be tested in the experiment. However, since we wanted to examine the role of predictions, it was important to place students in a theory-testing situation. The spacing-effect problem may be too complex for students to quickly develop their own theories to test from the start. For this reason, we gave students theories to test. In particular, the students were presented with two theories that had been proposed to account for the spacing effect and their goal was to develop experiments to tease the theories apart (i.e., determine if either, both, or neither of these theories adequately explained the spacing-effect phenomenon).

Briefly, the first theory was the shifting context theory, which stated that memories were associated with the context under study and that context gradually shifted with time. Thus, the spacing effect occurs because spaced practice produces associations to more divergent contexts, which in turn are more likely to overlap with the test context. The second theory was the frequency regularity theory, which stated that the mind tries to estimate how long memories will be needed based on regularities in the environment and, in particular, adjusts forgetting rates according to the spacing between items. The students were given longer descriptions of the theories (and the spacing effect itself) with concrete examples, could look at the descriptions of the theories throughout the task, and had several opportunities to ask the experimenter questions about the theories. (In Study 2, participants were only given the shifting context theory to test).

With the spacing-effect phenomenon and two theories in hand, we could have then given the students paper and pencil and asked them to describe an appropriate experiment. However, science is more than just experimental design. It also involves data analysis (among many other things). Moreover, few scientific questions are answered in the first experiment. Instead, scientists iterate and refine their methodology in response to experimental results. In order to place students in such a more realistic iterative situation that also included a data analysis process, we asked the students to design and interpret experiments using an environment called the Simulated Psychology Lab (Schunn & Anderson, 1999).

The Simulated Psychology Lab is a computer environment that allows students to design a wide variety of experiments and examine the results of those experiments. Students create experiments by selecting values for six factors, of which up to four could be simultaneously manipulated for any single experiment. They are told that the computer had been given the results of many actual experiments, and that it will display the results of any type of experiment they chose to generate.[2]

There were two groups of factors, source task factors and test factors, that the participants could manipulate. The source task factors included 1) repetitions—the number of times that the list of words was studied; 2) spacing—the amount of time spent between repetitions; and 3) source context—whether the participants were in the same context for each repetition or whether they changed contexts on each repetition. The test factors included 1) the test task—free recall, recognition, or stem completion; 2) delay—the amount of time from the last study repetition until the test was given; and 3) test context—whether the participants were in the same context or a different context at test relative to study. Only three of the factors are highly relevant to testing the two theories: spacing, source context, and test context. (In Study 2, since participants were asked to investigate only the shifting context theory, then only two factors are relevant: source context and test context).

For each of these factors, the participants could either vary it or hold it constant. Values had to be chosen for all of the factors before participants were allowed to continue. There were no confines on the order of value selection, and

[2] In fact, in order to produce numbers for the large number of possible combinations that the students could generate, the computer uses a mathematical model based on ACT-R (Anderson & Lebiere, 1998) that is very consistent with previous memory and spacing effect results, and includes a small level of random noise for added realism. See Schunn and Anderson (1999) for details.

the participants could change any of their selections at any time up until they chose to run the experiment.

The results were displayed using a table format and the participants could decide how to organize their tables. If participants were in an experimental condition that asked them to make predictions, then they made numerical predictions in a table. For each cell in the designed experiments, the participant must predict the percent correct of the hypothetical subjects. For example, Figure 1 presents an example table in which source context, spacing, and delay are manipulated and predictions have been already made for the first 8 cells (the bold 5 is currently being entered). Note that the table also contains information about the settings of the factors not being manipulated.

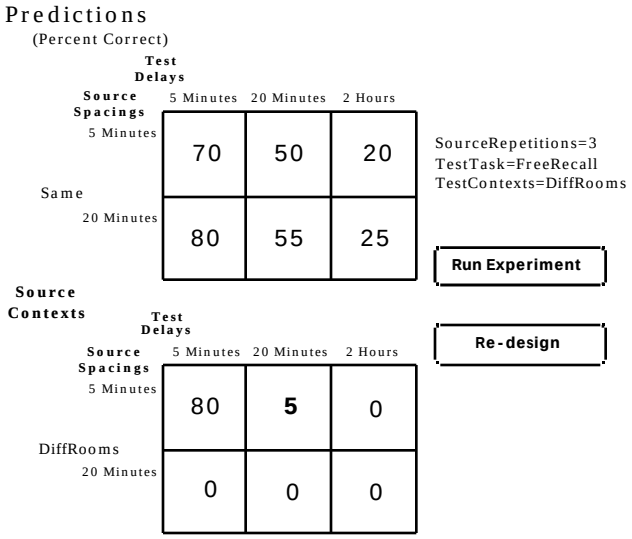

Figure 1. The interface used for making predictions.

A few words should be said about the form of the prediction task. In psychology, scientists are rarely asked to make precise numerical predictions. However, there are sciences in which one does make precise numerical predictions (indeed, in some sciences, predictions can only be made in quantitative terms because of the complexity of the theories). Moreover, it is not clear whether there is a simple method in a computer interface for asking students to make qualitative predictions for each of the factors (especially factors with 3 levels) and their interactions.

After making predictions, participants clicked on the 'Run' button and were shown the results of their experiments. Participants in an experimental condition that did not ask them to make predictions simply jumped straight to the experimental results. The results were shown in a table of the same format as was used to make predictions. If participants made predictions, the results table also showed their predictions (in smaller, italic text). Figure 2 presents an example results table (along with sample predictions). The correlation coefficient in the upper right is the Spearman correlation between the predictions and the actual outcomes, and was given to participants to provide a rough estimate of the accuracy of their predictions.

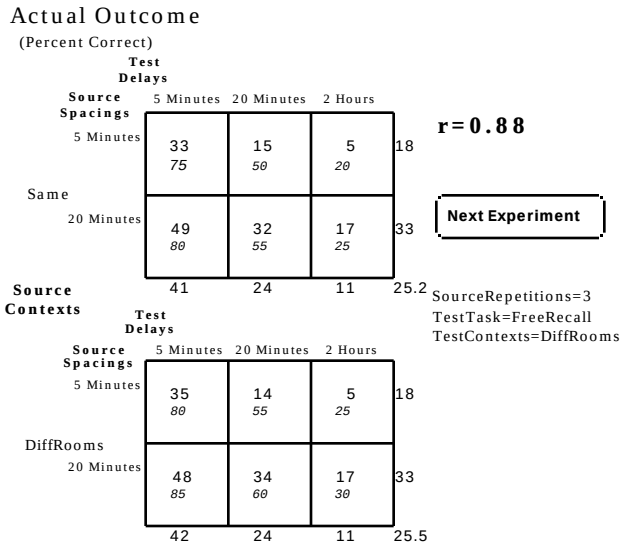

Figure 2.The interface used for displaying results (predictions, in italics, occurred only in the Prediction conditions).

For the purposes of this paper, there is one crucial performance dimension in this task: do participants focus on the theories under test? Previous research has shown that the majority of students in this task completely ignore the theories under test and simply focus on testing the effects of the 6 factors (Schunn & Anderson, 1998, 1999, In press). This focus on theories versus factors can be examined in two different ways. First, one can examine what types of conclusions the students make at the end of their experimentation: do the students make conclusions about the theories or the factors? Second, one can examine what types of experiments they design: do they focus on the factors that are actually relevant to the theories under test?

Study 1

Methods

Participants 56 George Mason undergraduates participated for course credit, of which 6 were removed due to computer problems. None of the participants had completed a research methods course, although a few (<10%) were currently enrolled in a research methods course.

Procedure Participants were randomly assigned to one of two conditions. Participants in the Prediction condition had to make numerical predictions for each cell in their experiments before viewing the outcomes of the experiment. By contrast, participants in the No Prediction condition skipped the numerical prediction phase entirely, both in the instructions and in the experiment itself.

Participants in both conditions were given a 15-minute tutorial on the computer that covered the spacing effect, the two theories, and how to use the Simulated Psychology Lab. The experimenter then reiterated the goals of the experiment (which had been presented on multiple computer screens including the very last one): to test the two theories of the spacing effect to determine whether one, both, or neither could account for the spacing effect. Participants worked on the task until they felt they understood the cause

of the spacing effect or until time had expired (40 minutes). Once finished, participants were asked what they found and their responses were recorded. They then answered a series of questions about the theories and any conclusions they came to about the effects of the six factors.

Results & Discussion

Overview The results are broken into 3 sections. First, we verify that there were no background differences between the groups. Second, we examine the effects of the manipulation on what kinds of experiments the students generated. Third, we examine the effects of the manipulation on what kinds of conclusions the students made at the end of the task.

Background Differences To verify that the groups were roughly equivalent, we compared their reported SAT and status. There were no differences by group in either measure. For status, 18% and 21% of the undergraduates were upperclassmen in the Prediction and No Prediction groups respectively, $\chi^2(1)<1$. For SAT, the combined Verbal + Quantitative scores were 1048 and 1052 for the Prediction and No Prediction groups respectively, $F(1,53)<1$.

Types of Experiments Conducted The participants in the Prediction group ran marginally fewer experiments than did the participants in the No Prediction group, with means of 5.8 and 8.4 experiments respectively, $F(1,55)=3.4$, $MSE=29.4$, $p<.1$. This result is not surprising because the No Prediction subjects had more time to run experiments since they did not have to make predictions for each one.

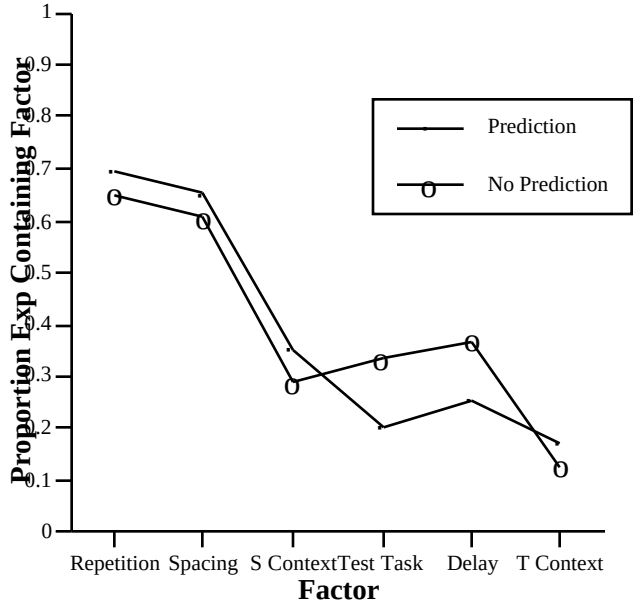

Figure 3. Mean proportion of experiments containing each of the factors within each group of Study 1.

More important than the number of experiments conducted are the types of experiments conducted. Figure 3 presents the proportion of experiments involving each factor. As one can see, the students in the Prediction group were generally more likely to focus on the factors relevant to the theories under test (with the exception of the repetitions factor, which is the first option in the interface). Let Appropriateness be defined as the mean proportion of experiments involving Spacing, Source Context, and Test Context minus the mean proportion of experiments involving Repetitions, Test Task, and Delay. Then students in the Prediction group had a significantly higher Appropriateness score than the No Prediction students, with means of 0.01 and -0.11 respectively, $F(1,53)=4.1$, $MSE=.05$, $p<.05$.

Types of Conclusions Made When the time was up or the students announced they were done, the experimenter asked the students what they had found. We coded whether the students responded to that question with a discussion of the factors that could be manipulated or a discussion of the theories under test. In the Prediction group, 31% of students mentioned the theories first, whereas in the No Prediction group, only 8% of students mentioned theories first, $\chi^2(1)=4.5$, $p<.05$. Thus, the manipulation did have a significant impact on whether the students focused on the theory testing nature of the task.

If they did not volunteer information about the factors at the end of the task, then the students were explicitly asked about each factor. There was no effect of the manipulation on the number of factors for which the students had correct statements about their effects, with means of 3.0 and 3.1 in the Prediction and No Prediction conditions, respectively, $F(1,53)<1$. Thus, the difference in propensity to make conclusions about the theories at the end of the task was not a function of having learned less about the factors.

Summary Study 1 found that forcing students to making predictions did improve scientific reasoning in that problem. In particular, it led students to actually focus on the theories under test and manipulate factors relevant to those theories.

Study 2 examines whether these results generalize to a situation in which students have been given only one theory to test. Making predictions may only be helpful when it leads students to realize the key differences between theories and thus generate experiments that would tease the theories apart. Additionally, the frequency regularity theory is somewhat subtle and it may be that many of the students either did not understand it or did not know how to test it. Thus, in Study 2, students were only asked to test the shifting context theory.

Study 2

Methods

Participants 69 undergraduates participated for course credit, of which 2 were removed due to computer problems. None of the participants had completed a research methods course, although a few (<10%) were currently enrolled in a research methods course.

Procedure The procedure for Study 2 was identical to Study 1 with two exceptions. First, participants were never told about the frequency regularity theory and were given only the shifting context theory to test. Second, we did not

collect background information about the students (SAT, major, year, etc) since it did not prove predictive of performance in Study 1.

Results & Discussion

Types of Experiments Conducted In study 2 the participants in the Prediction group ran approximately half as many experiments than did the participants in the No Prediction group, with means of 6.4 and 11.7 experiments respectively, $F(1,65)=8.8$, $MSE=52.6$, $p<.01$. Once again, this result is not surprising, since the No Prediction subjects had more time to run experiments because they did not have to make predictions for each one.

The size of the difference in number of experiments is larger than what was found in Study 1 and causes some problems for subsequent analyses. Specifically, it raises the question: are the differences in groups due to the number of experiments conducted or the cognitive consequences of the manipulation? Moreover, it appeared that in this Study, there were a significant number of participants running a very large number of experiments without much understanding (as many as 36 experiments in 40 minutes!)—they were simply clicking buttons. Therefore, we decided to remove from the remaining analyses all participants who ran more than 10 experiments (3 participants in the Prediction group and 6 in the No Prediction group). One consequence of this unequal reduction in condition Ns is that the subsequent condition comparisons should be more conservative tests of the manipulation: the ones removed from the analyses are more likely to not have understood the task and we have removed more of them from the No Prediction condition.[3]

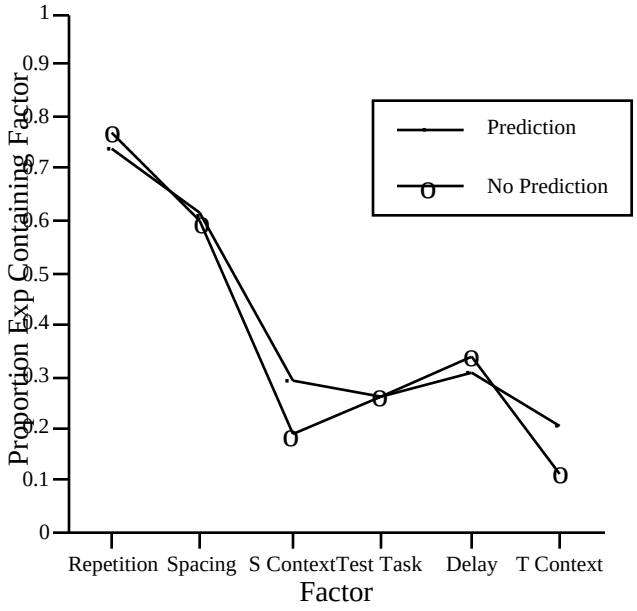

Figure 4. Mean proportion of experiments containing each of the factors within each group of Study 2.

[3] Moreover, the same patterns of results were found if all subjects were included in the subsequent analyses

The more central analysis was of the types of experiments conducted. Figure 4 presents the proportion of experiments involving each factor. As in Study 1, the students in the Prediction group were generally more likely to select the factors relevant to the theories under test. Since there was only the Shifting Context theory to test in Study 2, the Appropriateness measure must be redefined as the mean proportion of experiments involving Source Context and Test Context minus the mean proportion of experiments involving Repetitions, Spacing, Test Task, and Delay. Under this measure, students in the Prediction group had a significantly higher Appropriateness scores than the No Prediction students, with means of -0.12 and -0.35 respectively, $F(1,47)=3.9$, $MSE=.17$, $p<.05$.

Types of Conclusions Made As in Study 1, we coded whether the students responded to the final "what did you find?" question with a discussion of the factors that could be manipulated or a discussion of the theory under test. Students in the Prediction group mention the theory 11% of the time, whereas students in the No Prediction group never mentioned the theory on their own ($\chi^2(1)=2.9$, $p<.1$). Thus, the manipulation did have the same trend of an effect as in Study 1. This time, however, all students were quite unlikely to mention the theory on their own. It is possible that the students did not feel that the theory should be part of their final report since there was only one theory to test and they could not come up with an alternative theory.

As in Study 1, if the students did not volunteer information about the factors at the end of the task, then the students were explicitly asked about each factor. This time, however, there was a significant effect of condition on the number of factors for which the students had correct statements about their effects, with means of 3.2 and 3.8 in the Prediction and No Prediction conditions, respectively, $F(1,46)=4.77$, $MSE=0.98$, $p<.05$. That the No Prediction students had more correct responses establishes that the difference in propensity to make conclusions about the theory was not due to differences in what was learned about the factors.

Why did students in the No Prediction group produce a larger number of correct responses? It is likely that this effect occurred because the participants in the No Prediction task designed more experiments and explored more of the factors (especially the irrelevant factors). There were no differences between groups on the two most important factors. For source context, the Prediction group had a non-significantly higher proportion of correct responses (.31 versus .18, $F(1,46)<1$). For the test context, the less relevant factor of the two, the Prediction group had a non-significantly lower proportion of correct responses (.54 versus .68, $F(1,46)=1.0$, $p>.3$).

General Discussion

The two studies found generally quite consistent results: forcing students to making numerical predictions improves their scientific reasoning performance because it leads them to focus on the theories being tested and design more appropriate experiments.

The effects found in these studies were not large. However, the task given to the students is a very realistic scientific discovery task and was quite difficult for the students¾ in other words, there may have been relatively small improvements because the task was so difficult and there were possible floor effects in performance. Moreover, designing experiments which actually address the theories under test is such a central and important aspect of science. Any improvement from such a simple manipulation is important. Finally, previous research (Schunn & Anderson, In press) with this exact task has shown that even an entire course in research methods has relatively little impact on these same measures. Thus, that we found any improvement with such a simple manipulation is impressive.

While students were found to have a difficult time overall focusing on theories, we do not want to claim that most students could not focus on theories if the situation were made simple enough. However, that caveat is of little use to the educational setting in which students must learn to deal with experiments in real content domains. This consideration is what led us to use an authentic problem.

Our manipulations involve forcing students to make quantitative predictions for each cell in the design. What about other methods of generating predictions (e.g., generating more qualitative predictions)? Lehrer et al. (in press) argue that focusing on quantitative aspects of science is fundamentally important to scientific reasoning generally. However, one might imagine other methods for generating quantitative predictions. For example, what if one used graphical tools for generating predictions, or only forced predictions for each factor being manipulated and simple interactions among factors (rather than each individual cell)?

Our manipulation also focused on college students working on a problem in psychology. What about students working on problems in the physical sciences? One might imagine students in physics also losing track of the larger theories under test and focusing on the roles of particular concrete factors instead. Along those lines, Chabay and Sherwood (1999) have argued that giving physics students simulators that allow them to see the precise predictions of different theoretical assumptions improves students' understanding of the theories.

What about students in high school or elementary school? If university students lose track of the theories that are supposed to be tested, one can only imagine that this problem would be compounded in younger children. Indeed Deanna Kuhn's (1991) work suggests that children generally have a lack of differentiation between theory and evidence in scientific reasoning situations. However, whether making predictions actually improves performance for younger students (who may have other reasoning difficulties as well), is an open question.

Acknowledgments

Thanks to Elizabeth Mazzanti for her help in running and analyzing Study 1. Work on this paper was supported by funds to the first author from the Department of Psychology and the College of Arts and Sciences at George Mason.

References

Chabay, R. W., & Sherwood, B. A. (1999). Bringing atoms into first-year physics. *American Journal of Physics*, *67*, 1045-1050.

Chinn, C., & Malholtra. (in press). Epistemologically authentic scientific reasoning. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from everyday, classroom, and professional settings*. Erlbaum.

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

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*(2), 461-469.

Kuhn, D. (1989). Children and Adult as Intuitive Scientists. *Psychological Review, 96(4)*, 674-689.

Kuhn, D. (1991). *The skills of argument*. Cambridge, MA: Cambridge Press.

Lehrer, R., Schauble, L., & Petrosino, A. (in press). Reconsidering the role of experiment in science education. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from Professional, Instructional, and Everyday Science.* Erlbaum.

Loh, B., Reiser, B. J., Radinsky, J., Edelson, D. C., Gomez, L. M., & Marshall, S. (in press). Developing reflective inquiry practices: A case study of software, the teacher, and students. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from everyday, classroom, and professional settings*. Erlbaum.

Okada, T., & Shimokido, T. (in press). The role of hypothesis formation in a community of psychology. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from everyday, classroom, and professional settings*. Mahwah, NJ: Erlbaum.

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., 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(9)*, 859-882.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science, 23*(3), 337-370.

Schunn, C. D., & Anderson, J. R. (In press). Science education in universities: Explorations of what, when, and how. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from Professional, Instructional, and Everyday Science.* Erlbaum.

Schunn, C. D., & Klahr, D. (1993). Self- vs. other-generated hypotheses in scientific discovery. In W. Kintsch (Ed.), *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

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

White, B. Y. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition & Instruction, 10*(1), 100