Helping Students Make Controlled Experiments More Informative

Kevin W. McElhaney, Marcia C. Linn
University of California, Berkeley, 4523 Tolman Hall MC1670, Berkeley, CA 94720, USA
kevin777@berkeley.edu, mclinn@berkeley.edu

Abstract: We examine how encouraging students to *compare* rather than *isolate* variables affects their experimentation strategies and insights. We designed a week-long, technology-enhanced inquiry module on car collisions that logs students' interactions with a visualization. Physics students (N=166) were assigned to conditions that prompted them either to isolate or compare variables. Students responded to pretests, posttests, and embedded prompts that assessed students' understanding of motion graphs and collisions. Both groups made significant pretest to posttest gains. Students in the *compare* treatment used more diverse experimentation strategies than students in the *isolate* treatment. *Compare* students made nuanced interpretations of collision events based on threshold values. Case studies illustrate how comparing rather than isolating helped students use wide-ranging strategies to reach complex insights. The findings illustrate the value of encouraging multiple approaches to experimentation and connecting experimentation to real-life contexts.

Introduction

This study examines how two different types of experimentation goals led students toward different investigation strategies and insights about a complex science problem. We designed *Airbags: Too Fast, Too Furious?*, a week-long inquiry module for high school physics classes that guides students through an investigation about car collisions. In *Airbags*, students conduct experiments using a visualization to investigate what factors increase drivers' risk for injury from an airbag. The software logs students' experimentation choices and explanations to provide detailed information about their inquiry activities and inferences. This study examines two main research questions: (1) What is the overall impact of *Airbags* on students' ability to interpret and construct motion graphs? and (2) How does prompting students to *compare*, rather than *isolate,* variables in their experiments guide students toward different experimentation strategies and insights?

Rationale

Early research on scientific reasoning addressed children's ability to isolate variables in knowledge-lean experimentation contexts. For instance, Inhelder and Piaget (1958) designed a task [later adapted by Kuhn and Phelps (1982)] that asked subjects to determine what combination of colorless fluids would yield a specific reaction outcome. Siegler and Liebert (1975) examined the ways subjects determined how an electric train runs on the basis of four binary switches (though in actuality, a researcher operated the train using a secret switch to ensure that subjects would test all 16 combinations). These studies examined experimentation as domain-general logical inference, as subjects had no information on which to base testable hypotheses. In these situations, subjects could make valid inferences only by isolating variables to logically eliminate possibilities.

Over time, research has increasingly examined knowledge-rich contexts and revealed the important role of context-specific knowledge in experimentation. For example, studies show that children are more likely to test plausible rather than implausible hypotheses (Klahr, Fay, and Dunbar, 1993), focus on variables they believe to be causal (Kanari & Millar, 2004), and use experiments to achieve specific outcomes rather than test hypotheses (Schauble, 1996). Though learners' ideas about the investigation context may lead them toward invalid experimental designs or inferences, students may also use ideas productively, such as by narrowing the range of testable values or eliminating implausible explanations. Tschirgi (1980) argued that children's tendency to use "invalid" strategies when determining the ingredients needed to bake a good cake is reasonable, given real-life goals of reproducing positive results (good cakes) and eliminating negative ones (bad cakes). Koslowski (1996) also argued that using prior knowledge to generate and interpret evidence is a good strategy, particularly when understanding mechanisms informs the interpretation of outcomes.

In *Airbags*, students rely on many types of context-specific ideas to draw conclusions. Students' everyday understanding of motion, their physics domain knowledge about motion graphs, and evidence from the World-Wide Web all contribute to the way students design and interpret their experiments. Our previous work on *Airbags* (McElhaney & Linn, 2008) compared students who were constrained to conduct a specific number of trials to unconstrained students. During planning, constrained students attended more to the logistics of isolating variables, resulting in more controlled trials than unconstrained students. Unconstrained students, who attended more to the relationships among collision factors, variables, and outcomes, demonstrated a superior understanding of the situation. This result suggests that incorporating a context-specific understanding into experimentation was more important than employing "valid" strategies.

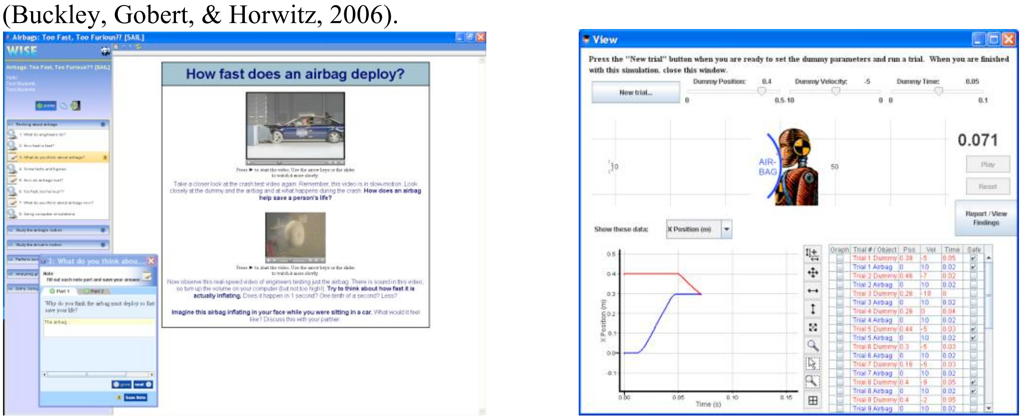
The current study extends our prior work by examining how experimentation goals can highlight the nature of the variables and make students' experimentation more informative. Research shows that goals for experimentation can influence learners' strategies and inferences, such as when children use different strategies to achieve positive outcomes than negative ones (Tschirgi, 1980) or when learners use different approaches for optimization than for rule-generation (Schauble, Klopfer, & Raghavan, 1991). This study draws from research in science instruction showing that comparisons can highlight key features of variables (Clement, 1983; Linn, 2005; Schwartz & Bransford, 1998) and investigates whether goals that encourage students to *compare* variables promote different insights about *Airbags* than goals that encourage students to *isolate* variables.

This study also explores how students come to understand *thresholds* (values of a particular variable above or below which the outcome is independent of the other variables). The presence of thresholds requires students to consider factors other than covariation to achieve a complete understanding of *Airbags*. Though students may readily observe the effects of thresholds by isolating variables, we hypothesized that comparing variables would highlight the distinct characteristics of each variable, helping sophisticated learners achieve a more nuanced understanding of *Airbags* based on thresholds as well as covariation. We also expected that the more familiar and tractable task of isolating variables would be more beneficial for less sophisticated learners.

Methods

Module and visualization design

We designed *Airbags* (Figure 1a), using the Web-based Inquiry Science Environment (Linn, Davis, & Bell, 2004). *Airbags*' learning goals are (1) how motion graphs represent one-dimensional motion and (2) what factors make airbags likely to injure drivers. Students experiment using a dynamic visualization (Figure 1b) designed by the Concord Consortium (www.concord.org). Students investigate the role of three *collision factors* on the driver's safety by experimenting with three *motion variables*. Table 1 summarizes factors, variables and questions for each treatment condition. To conduct a trial, students select an investigation question (or indicate that they are *Just exploring*), specify the variable values, and run the simulation. For each trial, students judge whether the driver was "safe" (encountered a completely inflated airbag) or "unsafe" (encountered the airbag within its deployment zone). The visualization presents the motion of the airbag and driver using graphs that coordinate with an animation. The visualization also helps students manage their trial history, sort trial outcomes, and compare multiple trials. The software logs students' experimentation choices for subsequent analysis (Buckley, Gobert, & Horwitz, 2006).



(a) (b)

Figure 1. (a) The first activity of *Airbags*. (b) The experimentation visualization.

Students may use two types of inferences to explain their findings: covariation and threshold values. First, over a particular range of values, each of the three variables covaries with the time that elapses before the driver and airbag collide. Tall drivers, low speed collisions, and a large crumple zone therefore make a driver more likely to encounter a fully inflated airbag than short drivers, high speed collisions, and a small crumple zone. Second, two threshold values (for position and time) determine situations where the likelihood of injury is invariant: (1) short drivers who sit within an airbag's zone of deployment will *never* encounter a fully inflated airbag, and (2) for sufficiently tall drivers, if the duration of the crumple zone exceeds the deployment time for the airbag, drivers will *always* encounter a fully inflated airbag. Responses to embedded assessments indicate whether students attribute their findings to covariation and/or thresholds.

Study design

Groups in the *isolate* and *compare* conditions received investigation questions that encouraged them to isolate or compare variables, respectively. Except for the investigation questions, the modules used for each condition were identical. We used a pretest/posttest experimental design with embedded assessments and two comparison conditions. We randomly assigned student groups to one of the conditions using a stratified approach to distribute their ability equally across the two conditions. To keep instruction similar across conditions, teachers and researchers focused class discussions on graph interpretations rather than on investigation questions and refrained from guiding individual students on designing experiments.

Table 1: Collision factors, variables, and investigation questions for the *isolate* and *compare* conditions.

Collision factor	Motion variable	Investigation question	
		Isolate condition	*Compare* condition
Driver height	Position	Are TALL or SHORT drivers more likely to be injured by a deploying airbag?	Does the DRIVER'S HEIGHT make the biggest difference in whether the driver is injured?
Collision speed	Velocity	Do HIGH or LOW SPEED collisions make drivers more likely to be injured by a deploying airbag?	Does the COLLISION SPEED make the biggest difference in whether the driver is injured?
Car crumpling	Time	Does MORE CRUMPLING or LESS CRUMPLING make drivers more likely to be injured by a deploying airbag?	Does HOW MUCH THE CAR CRUMPLES make the biggest difference in whether the driver is injured?

Participants

Physics students (N=166) at five socially diverse high schools in the United States studied *Airbags*. Most students worked in dyads on the module (unpaired students worked alone). Students were usually grouped with other students of nearly equal ability. Most teachers had taught the topic using previous versions of *Airbags*.

Data sources and scoring

Pretests and posttests

Ten constructed response pretest and posttest items assessed students' ability to interpret and construct graphs of one-dimensional motion in a context outside of *Airbags*. We administered pretests individually to students the day before the beginning of *Airbags* and posttests within a few days of completion. We scored pretest and posttest responses from zero to four using knowledge integration rubrics (Linn, Lee, Tinker, Husic, & Chiu, 2006) that measured how well students connected characteristics of motion to the features of motion graphs.

Embedded assessments

Use of the Control-of-Variables Strategy (CVS). We used students' experimentation sequences (as captured online in log files) to compute a CVS score, the percentage of each group's trials that were part of a controlled comparison between successive trials using the variable appropriate to the chosen investigation question. We used only trials that specified one of the three investigation questions (i.e. not *Just exploring*) for this score. This proportion score indicates the extent to which students mindfully used CVS to investigate the three questions.

Interpreting and constructing Airbags graphs. Twelve items asked students to interpret motion graphs from the visualization or generate graphs that represented a collision situation. Each group received a COLLISION GRAPHS score that captured how well they connected characteristics of the graphs to the events in a collision. We scored these explanations from zero to five using knowledge integration rubrics.

Understanding of covariation and thresholds. Three items prompted students to explain their answers to the three investigation questions listed in Table 1. We used these responses to determine whether each group attained a covariation-based and/or thresholds-based understanding of the *Airbags* situation. To examine how frequently students attributed their findings to covariation, we counted the number of responses by each group that described covariation between the factor or variable in question and elapsed time before the driver encounters the airbag. To examine how frequently students attributed their findings to thresholds, we counted the number of responses by each group that used at least one of the thresholds in support of their finding.

Videorecords

We videorecorded 12 dyads as they engaged in the experimentation activity, so that we could closely examine the discussions that occurred during experimentation within specific dyads. We chose two of these dyads for a case comparison to illustrate how the investigation questions they received influenced their experimentation strategies and insights. A more detailed description of these two dyads and the reasons we chose them for analysis follows the presentation of the results of the comparison study.

Results

Impact of *Airbags* on students' understanding of motion graphs

Students made moderate, significant pretest-posttest gains [M = 29.99, SD = 6.66 (pre); M = 32.76, SD = 5.96 (post), $t(128) = 5.17$, $p < .001$ (two-tailed), $d = 0.44$]. Gains were positive for all five schools, and significant for four of the schools. Low prior knowledge learners made the greatest gains. The gains indicate that *Airbags* was successful in helping diverse learners interpret and construct motion graphs. The gains are impressive considering that all students had completed a kinematics unit shortly before studying *Airbags*, and they represent value added to a traditionally taught kinematics unit.

The difference in posttest scores between the two conditions was not significant. We did not expect differences because the experimentation activity generally comprised just 10-20% of the total time spent on *Airbags*. Students in both conditions therefore had many opportunities to improve their graphing knowledge other than by experimentation. Furthermore, there were no significant differences between the conditions on the COLLISION GRAPHS scores, suggesting that the *isolate* and *compare* prompts were equally effective in helping students connect the characteristics of the *Airbags* graphs to the collision events.

To examine subtle impacts of experimentation goals on students' strategies and insights, we sorted student groups into low, middle, and high prior knowledge tertiles using the mean pretest score for each group. Our exploratory analyses show variation in students' strategies and inferences by tertile.

Impacts of prior knowledge on students' strategies and inferences

CVS scores showed that the high tertile groups conducted a significantly higher proportion of controlled trials than the low and middle tertile groups [M = .79, SD = .30 (high), M = .45, SD = .35 (low/middle), t(79) = 4.39, $p < .001$ (two-tailed)]. This difference suggests that high prior knowledge students as a whole were more focused on controlling variables in their investigations.

Overall, about 30% of the students generated covariation-based explanations of the collision events. The percentage was somewhat higher for the high (38%) and middle (36%) tertile groups than for the low (19%) tertile groups, but this difference was not significant, indicating that a covariation-based understanding of *Airbags* was accessible to students at all prior knowledge levels. However, just 2% of low and middle tertile students generated thresholds-based explanations, compared to 31% of the high tertile students. A Wilcoxon rank-sum test showed that the difference in the average number of thresholds-based explanations was significant [M = .59, SD = .98 (high), M = .038, SD = .28 (low/middle), U = 3.78, $p < .001$]. This result suggests that only sophisticated learners were able to achieve a thresholds-based understanding of *Airbags*.

Impacts of experimentation goals on students' strategies and inferences

CVS scores showed significant differences between the conditions only for the high tertile students [M = .94, SD = .06 (*isolate*), M = .70, SD = .35 (*compare*), t(27) = 2.22, $p = .035$ (two-tailed)]. A close examination of these data revealed that while virtually all the *isolate* groups devoted nearly all their trials to controlled comparisons between successive trials, one-third of the *compare* groups used at least half their trials for other strategies. (The case comparison that follows will illustrate some of these strategies.) This result suggests that though most students used similar approaches to investigate the *isolate* and *compare* questions, the *compare* questions led some high prior knowledge students to use alternative strategies to CVS in their investigations.

There were no differences in the number of covariation-based explanations between the two conditions, indicating both conditions were equally effective in leading students toward a covariation-based understanding of *Airbags*. However, we observed significant differences in the number of thresholds-based explanations for the high tertile students. Just 8% of the high tertile *isolate* groups generated thresholds-based explanations, compared to 44% of the *compare* groups. A Wilcoxon rank-sum test showed that the difference in the average number of thresholds-based explanations for high tertile students was significant [M = .091, SD = .30 (*isolate*), M = .89, SD = 1.13 (*compare*), U = 2.09, $p = .037$]. This finding suggests that the *compare* questions promoted a thresholds-based understanding of *Airbags*, though this difference occurred only with sophisticated learners.

Case comparison

We use a case comparison to illustrate some ways the *isolate* and *compare* tasks might have led sophisticated students toward different experiences with the *Airbags* visualization. We chose two similar high-tertile dyads for a case comparison. The students in these dyads were enrolled in the same honors physics curriculum with the same teacher and had similar pretest and posttest scores. Furthermore, all four students were concurrently enrolled in calculus and thus had strong mathematics skills. Because the student population and array of schools used for this study was highly diverse, no single case can represent the "typical" experience students have with *Airbags*. These cases rather aim to illustrate how the *compare* questions might have prompted sophisticated students to consider aspects of the *Airbags* situation that the more traditional *isolate* question might not have.

Case 1: Brett and Eric (*isolate* condition)

Overview

Brett and Eric (pseudonyms) studied the *isolate* version of Airbags. They conducted 10 total trials. For their first four trials, they chose *Just exploring* as their goal, during which time they explored outcomes for default and extreme values of all three variables. They devoted the last six trials to conducting a pair of controlled trials for each of the three investigation questions. Their experimentation session lasted about 10 minutes. This analysis focuses on the three controlled trial pairs.

Trials 5 & 6: "He's going to be safe, obviously"

In trials 5 and 6, Eric and Brett investigated whether tall or short drivers are more at risk. In trial 5, they examined the outcome for a short person. The following excerpt illustrates their variable choices for trial 6 and their interpretation of the two trials:

(84) E: …Short or tall. And now we have to move the guy back, cause he's taller. So we gotta keep everything except position. So move him back some. Like right there.
(85) B: He's going to be safe, obviously.
(86) E: He might not, let's just check. [They run trial 6.] Yeah. So mark that as safe. OK, put the graphs for the previous two. [They compare the graphs of trials 5 and 6]
(87) B: They're both safe.
(88) E: Yeah. So let's go to the next question.

Two things are apparent from this exchange. First, because trial 5 produced a "safe" outcome, Brett knew that the outcome of the trial 6 would also be "safe" before conducting it. However, rather than choosing a set of values that would provide them with more information, they simply ran the test. Second, in these two trials they achieved the same "safe" outcome, failing to provide strong evidence for the effect of the position variable on the risk to the driver. Further tests aiming to illustrate conditions that led to an "unsafe" outcome would have better informed their understanding. However, their variable choices and the brevity of their discussion about the results suggest they are focused more on isolating variables than on gaining insight about the situation.

Trials 7 & 8: "We can only change one of them, we can't change multiple ones"

The following exchange occurred immediately after isolating the velocity variable in trials 7 and 8:

(93) E: …since we're doing, like, experiments, we can only change one of them, we can't change multiple ones.
(94) B: Yeah.
(95) E: Cause like in real life, there would be a combination of all three.

Eric's comments shed light on their commitment to using CVS throughout the activity. Eric believed that they were prohibited from using other strategies (though at no point does *Airbags* suggest how students should conduct their trials). Furthermore, the distinction he made between their experiments and "real life" indicates he believed these other strategies would be permissible in other contexts. Though Eric did not elaborate on what he meant by "a combination of all three", his words suggest that by isolating variables they aim to fulfill expectations imposed on them by the culture of classroom science.

Trials 9 & 10: "Just go low first"

Brett and Eric used their final two trials to isolate the time variable. This exchange occurred as they decided how to conduct these two trials:

(97) E: So--more or less crumpling. So keep the velocity constant. So then, if there's less delay, that means more crumpling. So then, yeah let's do less crumpling.
(98) B: That's more crumpling.
(99) E: That's…yeah. So just go low first. Like there's low crumpling, almost no crumpling.
…
(103) E: OK, so then, that was low speed, so this is high speed. All right.
(104) B: More crumpling.
(105) E: Yeah, more crumpling.

To begin, Eric incorrectly stated the relationship between the crumpling factor and the time variable. Though this error could be conceptual, Eric's subsequent request to "just go low first" suggests that the correct relationship was unimportant to him. Because the precise nature of the relationship between the factor and the variable (direct or inverse) would not change how they employed CVS, the only decision Eric believed they need to make is whether to test the "low" or the "high" value first. Next, even after being reminded of the correct relationship, Eric attributed the time variable to the wrong factor entirely (speed, rather than crumpling). By this time, Eric was no longer attending to the nature of the variables and appears to have sequestered their experimentation strategy from their understanding of the *Airbags* context. At this point, the variables might as well have been X, Y, and Z rather than position, velocity, and time.

Case 2: Joann and Linda (*compare* condition)

Overview

Joann and Linda (pseudonyms) studied the *compare* version of *Airbags*. They began their experimentation by carrying out an initial plan to isolate each variable. They quickly abandoned that approach and employed other strategies such as testing extreme values and incrementally varying individual variables. Their experimentation session lasted about 30 minutes. This analysis will focus on three excerpts that illustrate the evolution of their investigation strategies and the insights they achieved by using these diverse approaches.

Trials 1 - 4: Abandoning CVS

Joann and Linda used their first three trials to isolate the position variable and test its full range. As they decided on the values for the fourth trial, Linda began to reconsider their approach.

(41) L: I don't know. Maybe we just test ummm, like, test the position at, like, three different points. That's just so—that's just so many tests, never mind….

At first they were discouraged by the sheer length of their proposed approach, but after choosing some intermediate values for trial 4 and discussing the outcome, their discussion about the "effect" of the car crumpling empowered them to abandon their initial strategy:

(82) J: Then it all falls back to what we said originally, the crash, the speed of the crash dictates if position and dummy time, you know, the crumpling of the car would have an effect.

(83) L: Yeah. I don't understand why we have to do different tests for each three different sections [investigation questions]. You know? You click on them and be like whatever trials for this, kind of. 'Cause it looks like we're kind of figuring it out as we're looking at this.

Here, the *compare* questions promoted a discussion about the relative "effects" of the variables on the outcomes that the *isolate* questions, by their very nature, were unlikely to promote. The discussion about effects appears at least in part to have led Linda and Joann to abandon their initial approach. Unlike Brett and Eric, they had prioritized "figuring it out" over conducting "whatever trials" that were expected of them for each investigation question. At this point, they took a different approach toward making sense of the *Airbags* situation.

Trials 5 - 7: Exploring extreme values

In trials 5, 6, and 7, Joann and Linda explored extreme values of the velocity and time variables. In trial 5 they simulated a "high impact crash", setting the velocity to the "fastest possible." In trial 6, they "try it with dummy time if we put it at, like, zero." In trial 7, they tested the maximum crumple zone ("All the way?"). This sequence culminated in an important observation Joann made as they examined trials 6 and 7:

(137) J: Well, I guess, dummy time can also have an effect with position, cause like I'm saying, if you have no dummy time, then how close you are to the steering wheel matters a lot.

Again, only the *compare* questions would prompt Joann make this insight about how much the proximity to the steering wheel "matters". In this case, the extreme value approach illustrated to Joann that the effect of time on the outcome may depend on the value of position. Though this initial understanding of the interaction between position and time was at this point incomplete, it became increasingly sophisticated through continued experimentation and discussion, eventually leading to a highly nuanced understanding of the threshold values.

Trials 8 - 11: Incremental variation

Beginning with trial 8, Joann and Linda changed approaches once more to examine the effects of minute changes to the variables. From trial 8 to trial 9, they decreased the time by the smallest possible increment. After not finding anything conclusive, they incrementally adjusted the velocity in trial 10. Comparing trials 9 and 10 revealed to them nearly identical graphs and led to the following exchange:

(175) L: …So then that one doesn't make much of a difference.
(176) J: No. OK.
(177) L: So the speed didn't make a difference in that—
(178) J: Oh. Doesn't that counteract what we first said? ...I mean, contradict?
…
(190) J: I'm thinking that, he doesn't start moving until, I don't know, I can't put it in words. Even though he's moving faster, by even a little bit, it doesn't really have much of an effect because then, the amount of time….

Observing the similarity between the trial 9 and trial 10 graphs made them realize for the first time that their initial hypothesis (that speed would have the largest effect) was incorrect. Joann was almost able to explain why, but could not quite articulate the precise reason for the insensitivity of the outcome to the velocity. However, using trial 11 to retest the situation once more with the highest allowable velocity value crystallized their understanding of the time threshold:

(207) L: So it's the same, the speed doesn't change anything, it's the same graph we just had.
(208) J: Oh, in that case, cause he won't start moving until—he won't start moving *period* [verbal emphasis] until—this has already been inflated no matter how fast he's going. Yeah. Because he won't start moving until .06 seconds has gone by.
(209) L: Right.
(210) J: So he could be moving at 100,000 miles an hour and he won't hit it until the airbag's already inflated, according to how we set this up.

One revealing aspect of Joann's and Linda's last four trials is that they do indeed constitute use of CVS. However, they conducted these controlled trials in a different way (using incremental variation) and for a much different purpose (to compare the magnitude of effects) than for their earlier controlled trials, or for Brett's and Eric's controlled trials. Not surprisingly, they were also much more informative than the other trials. The design of these trials made use of their previous insights, and their use of CVS emerged spontaneously as the best way to further their understanding of the situation, rather than as a strategy they learned in science class.

After conducting trial 11, Joann's and Linda's understanding of the situation was so complete, they did not require any additional trials to succinctly characterize the relative effects of position, velocity and time on the outcome, on the basis of threshold values:

(227) J: …if we move it closer to the steering wheel, then dummy velocity and dummy time wouldn't matter, because as the airbag starts inflating, he'd be in the way…
…
(238) J: And, he waits that certain amount of time for the airbag to inflate, then velocity doesn't matter.

Discussion

Airbags benefited learners across the distribution of prior knowledge, albeit in different ways. Learners with initially poor understanding of motion graphs made large gains in their abilities to interpret and construct motion graphs, and they were able to generalize their knowledge from *Airbags* to other motion contexts. However, these students' experimentation choices and responses to embedded prompts show that many of them struggled to investigate the questions systematically and reach meaningful insights. On the other hand, learners with initially strong understanding of motion graphs had less room for improvement in that area, but conducted experiments that led to more valid inferences. In particular, the *compare* questions helped the most sophisticated learners examine key distinctions between the variables, such as the magnitudes of the effects of each variable on the outcome.

The stark differences between the two cases illustrate how the *isolate* and *compare* questions led students to reason about the *Airbags* situation. For Brett and Eric, the *isolate* questions appeared to provoke a "schoolish" interpretation of the task. They viewed the task as a simple covariation problem (a common task in school science), and as a result they limited themselves to a predetermined pattern of using CVS rather than spontaneously employing diverse strategies to investigate new questions. They prioritized validly implementing CVS over gaining insight. They sequestered their understanding of the *Airbags* situation from their investigation strategy and, more generally, from their conceptions of the *Airbags* task and the real life practice of science. Though their strategy was "valid" (as judged by criteria often imposed by classroom science), it was not especially informative. Their analysis did not go beyond a superficial characterization of the variables.

In contrast, the *compare* questions prompted Joann and Linda to incorporate a wider range of strategies to elucidate variation patterns. They conducted trials intending to understand the relationships between variables and outcomes and the mechanisms that governed these relationships. Though their initial efforts to use CVS did not yield useful ideas, in the end they spontaneously used CVS to achieve a nuanced understanding of *Airbags* by building on ideas they refined using other strategies. In this way, Joann's and Linda's use of CVS during their final four trials evolved naturalistically as a way to investigate questions that arose from their previous trials. The *compare* questions encouraged Joann and Linda to deeply consider the nature of the variables and to make important and meaningful distinctions between them.

The complexity of Joann's and Linda's analysis illustrates why we observed differences between the conditions only for high prior knowledge students. Joann and Linda needed highly sophisticated knowledge of experimentation strategies and graph interpretation in order to reach their advanced level of understanding. Less sophisticated students likely lacked sufficient knowledge either to distinguish the *compare* questions from the *isolate* questions or to adequately investigate either type of question. Future research will examine how to refine the guidance to help more typical students conduct informative experiments.

Conclusions and Implications

This study investigated the effect of encouraging students to compare rather than isolate variables. In *Airbags*, students who were guided to *isolate* variables often missed important insights by limiting their analysis to one variable at a time. Prompting students to *compare* variables led to a more nuanced understanding of the *Airbags* situation particularly for high prior knowledge students, achieved at least partly by using diverse experimentation strategies.

Our findings point to the value of providing opportunities for multiple approaches to experimentation rather than guiding students only to isolate variables. The findings suggest that instructional designers should balance guidance designed to promote CVS with opportunities to explore the nature and meaning of the variables. Designers of instruction should select problems where subtle distinctions such as thresholds are necessary for complete understanding. Many everyday problems such as decisions about drug dosage require an understanding of thresholds. Our study demonstrates the value of connecting experimentation to real-life contexts such as airbags, where students can appreciate connections between science instruction and everyday life.

This study also illustrates the value of providing students with graphical representations to help them make valid inferences from their experiments. Our detailed analysis of the ways students reasoned with the *Airbags* visualization revealed that the graphs were essential for comparing multiple trials, revisiting previous ideas, and interpreting collision events. How graphing tools should be designed to support scientific reasoning merits further study.

References

Buckley, B., Gobert, J. & Horwitz, P. (2006). Using Log Files To Track Students' Model-based Inquiry. In S. A., Barab, K. E Hay & D. T. Hickey (Eds.) *Making a difference: Proceedings of the Seventh International Conference of the Learning Sciences* (pp. 57–63). Mahwah, NJ: Erlbaum.

Clement, J. (1983). Using Bridging Analogies and Anchoring Intuitions to Deal with Students' Preconceptions in Physics. *Journal of Research in Science Teaching, 30*, 1241–1257.

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

Kanari, Z. & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching, 41*, 748–769.

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

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

Kuhn, D. & Phelps, E. (1982). The development of problem-solving strategies. In H. Reese (Ed.), *Advances in child development and behavior* (vol. 17, pp. 1-44). New York: Academic Press.

Linn, M. C. (2005). WISE Design for Lifelong Learning—Pivotal Cases. In P. Gärdenfors & P. Johannsson (Eds.), *Cognition, Education and Communication Technology*. Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, M. C., Davis, E. A., & Bell, P. (Eds.). (2004). *Internet Environments for Science Education*. Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, M. C., Lee, H.-S., Tinker R., Husic, F., & Chiu J. L. (2006). Teaching and assessing knowledge integration in science. *Science, 313*, 1049–1050.

McElhaney, K.W. & Linn, M.C. (2008). Impacts of students' experimentation using a dynamic visualization on their understanding of motion. *International Perspectives in the Learning Sciences: Cre8ting a Learning World. Proceedings of the 8th International Conference of the Learning Sciences.* International Society of the Learning Sciences Inc., Utrecht, The Netherlands.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology. 32*(1), 102–119.

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.

Schwartz, D. L. & Bransford, J. D. (1998). A Time for Telling. *Cognition and Instruction, 16*(4), 475–522.

Siegler, R.S. & Liebert, R.M. (1975). Acquisition of formal experiment. *Developmental Psychology*, *11*, 401–412.

Tschirgi, J. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Development*, 51, 1–10.

Acknowledgments

We thank Carol Lee, Brian Reiser, and Bruce Sherin for helpful discussions and three anonymous reviewers for suggestions. We acknowledge the support of National Science Foundation Grant No. ESI-0334199.