Investigations of a Complex, Realistic Task: Intentional, Unsystematic, and Exhaustive Experimenters

Kevin W. McElhaney and Marcia C. Linn

Graduate School of Education, University of California, 4523 Tolman Hall, MC 1670, Berkeley, California 94720-1670

Received 16 January 2010; Accepted 17 May 2011

Abstract: This study examines how students' experimentation with a virtual environment contributes to their understanding of a complex, realistic inquiry problem. We designed a week-long, technology-enhanced inquiry unit on car collisions. The unit uses new technologies to log students' experimentation choices. Physics students ($n = 148$) in six diverse high schools studied the unit and responded to pretests, posttests, and embedded assessments. We scored students' experimentation using four methods: total number of trials, variability of variable choices, propensity to vary one variable at a time, and coherence between investigation goals and experimentation methods. Students made moderate, significant overall pretest to posttest gains on physics understanding. Coherence was a strong predictor of learning, controlling for pretest scores and the other experimentation measures. We identify three categories of experimenters (intentional, unsystematic, and exhaustive) and illustrate these categories with examples. The findings suggest that students must combine disciplinary knowledge of the investigation with intentional investigation of the inquiry questions in order to understand the nature of the variables. Mechanically executing well-established experimentation procedures (such as varying one variable at a time or comprehensively exploring the experimentation space) is less likely to lead students to valuable insights about complex tasks. Our proposed categories extend and refine previous efforts to categorize experimenters by linking scientific procedures with understanding of the science discipline. © 2011 Wiley Periodicals, Inc. J Res Sci Teach 48: 745–770, 2011

Keywords: curriculum development; inquiry; science education; technology education/software design

Introduction

Research on students' views of the nature of science show that they hold impoverished views of scientific experimentation (Carey, Evans, Honda, Jay, & Unger, 1989). Recent reports such as America's Lab Report (National Research Council, 2006) and Taking Science to School (National Research Council, 2007) point to the superficial nature of students' laboratory activities as one source for students' limited views of authentic science. These reports characterize typical laboratory activities as having a strong emphasis on procedures instead of concepts. Experimentation is often isolated from domain knowledge. The reports call for engaging students in complex, multivariate investigations to promote sophisticated views of

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: National Science Foundation; Contract grant numbers: REC-0334199, DRL-0918743.

Correspondence to: K.W. McElhaney; E-mail: kevin777@berkeley.edu

DOI 10.1002/tea.20423

Published online 20 June 2011 in Wiley Online Library (wileyonlinelibrary.com).

professional science inquiry. Designing such investigations and ensuring that students can learn from them has proven difficult (Fishman, Marx, Blumenfeld, Krajcik, & Soloway, 2004; Fortus, Dershimer, Krajcik, Marx, & Mamlok Naaman, 2004; Krajcik, Blumenfeld, Marx, & Soloway, 2000; McDermott, 1991; Polman, 2000).

This article examines how students investigate a realistic inquiry topic: car collisions. Students studied *Airbags: Too Fast, Too Furious*? (henceforth *Airbags*) in high school physics. The week-long unit enables students to conduct experiments using a dynamic visualization that illustrates the motion of an airbag and the driver of a car during a collision. The unit, implemented in the Web-based Inquiry Science Environment (WISE; Linn, Davis, & Bell, 2004) is designed to help physics students gain an understanding of kinematics in one dimension. *Airbags* guides pairs of students to conduct experiments using the visualizations, discuss their findings with peers, reflect on their progress, and refine their ideas. *Airbags* has undergone several refinements based on classroom trials to make the topic of airbag safety accessible to students without the need for expensive or dangerous equipment.

Airbags takes advantage of the benefits of virtual experimentation environments (Klahr, Triona, & Williams, 2007; Zacharia, Olympiou, & Papaevripidou, 2008). To study how students conduct virtual experiments to investigate a realistic inquiry problem about airbag safety, we designed *Airbags* and used logging software combined with embedded assessment to capture experimentation practices. In *Airbags*, variables exhibit complex relationships such as threshold values and interactions, therefore not all controlled comparisons are equally informative. The software can log students' explanations, drawings, and interactions with the visualization to provide detailed information about student inquiry activities during regular classroom instruction.

We asked students not only to identify variable relationships but also to explain the mechanisms that govern these relationships. Students could not rely solely on prescribed procedures for controlling variables and needed to use disciplinary knowledge and everyday understanding of the investigation context to design informative experiments.

The log files and embedded assessments track each student group's experimentation goals and choices. These detailed accounts of students' experimentation document how students plan experiments and conduct realistic investigations. Pretest, posttests, and embedded assessments capture what students learn from this experience. These records allow us to address the following research questions: (1) What is the overall impact of *Airbags* on physics students' understanding of motion graphs? (2) What roles do controlled experiments and knowledge about the variables play as students investigate a realistic problem exhibiting complex variable relationships? and (3) How do students' experimentation practices relate to their learning outcomes and their understanding of the mechanisms governing airbag safety?

Rationale

This research builds on prior research and seeks to extend understanding of the role of experimentation methods in student learning. The National Research Council (2006) has identified comprehensive learning goals for lab experiences including "enhancing mastery of subject matter; developing scientific reasoning; understanding the complexity and ambiguity of empirical work; developing practical skills; understanding the nature of science; cultivating interest in science and interest in learning science; developing teamwork abilities." (p. 3). Typical experimentation tasks that require students to isolate linearly related variables may develop practical skill but neglect the complexity of science. Realistic experimentation activities pose problems with variables that interact, connect to everyday or personally relevant dilemmas, require students to use their disciplinary knowledge to interpret the results, and engage students in weighing alternatives. Realistic tasks can make knowledge of subject matter consequential to the design and interpretations of experiments, provide faithful representations of the complexity of real-life science investigations, present science as relevant to everyday life, and require students to work together in order to succeed.

Research on everyday experimentation raises issues that are not addressed in tasks that minimize the influence of prior knowledge on reasoning about the system. Inhelder and Piaget (1958) designed a task [later adapted by Kuhn and Phelps (1982) and others] 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 participants had no information on which to base testable hypotheses. In these situations, participants could make valid inferences only by studying all the variables to logically eliminate possibilities.

Many classroom studies that promote mastery of the ‘‘control-of-variables strategy’’ in science classrooms (Chen & Klahr, 1999; Dean & Kuhn, 2007; Klahr & Nigam, 2004) make domain knowledge essentially irrelevant to the investigations. These classroom tasks present experimentation to students as a procedure to follow or a skill to practice rather than as a learner-initiated component of authentic science inquiry.

This study examines how students use experiments to investigate everyday dilemmas. Students in science classes need to learn how to interpret everyday dilemmas, not just apply procedures. In realistic problems variables exhibit complex relationships to each other and to outcomes. In everyday situations, variables interact, are often non-linear, and have critical or threshold values that determine relationships to outcomes. Learners must bring together multiple sources of evidence such as science principles, personal observations, understanding of the investigation context, and knowledge of experimentation practices. Using all the available evidence to reach valid insights from experiments requires learners to integrate knowledge. Learners must understand the nature of the variables and apply ideas in the discipline in order to design informative experiments and draw conclusions that have meaning in their lives.

Adding everyday or realistic tasks to science instruction comes with challenges. Students use their everyday knowledge to inform the design and interpret the results in both helpful and potentially confounding ways. Linn, Clement, and Pulos (1983) compared students’ reasoning in laboratory tasks and naturalistic tasks involving the effects of system variables on an outcome. The study found that a significant part of the variance in performance on these tasks was associated with domain knowledge of the experimental variables.

When students try to use prior knowledge to make their experiments more informative they may narrow the range of testable values or eliminate implausible explanations. Tschirgi (1980) argued that children’s tendency to use ‘‘invalid’’ approaches 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 approach, particularly in situations where understanding mechanisms informs the interpretation of outcomes. These studies indicate that learners’ experimentation approaches sometimes stem from practices that are essential in science such as focusing on consequential investigation questions or exploring the nature of key variables.

Everyday knowledge about the variables can distract learners from testing crucial assumptions. For example, studies show that children are more likely to test plausible rather than implausible hypotheses (Klahr, Fay, & Dunbar, 1993; Tschirgi, 1980), focus on variables they believe to be causal (Kanari & Millar, 2004), and use experiments to achieve specific outcomes rather than test hypotheses (Schauble, 1996). Schauble (1996) examined experimentation by children and adults in two science domains and found that subjects who conducted valid experiments often reached invalid conclusions informed by their prior knowledge of the system.

In realistic investigations students must incorporate many sources of evidence to design and interpret experiments, isolate variables judiciously, and interpret the results by weighing tradeoffs and often identifying unanswered questions. In these situations, the ability to incorporate and evaluate a broad range of evidence is essential. Yet, many classroom experiments neglect the authentic practice of science inquiry. They may make the control of variables strategy the only informative strategy. They could reward the belief that all controlled comparisons are equally valid. They might treat the ability to control variables as the learning goal, rather than combining this goal with understanding of the investigation context.

Experimenting in realistic contexts requires learners to consider a wide range of ideas to design informative experiments. Learners need to integrate everyday ideas they have about the topic, formal knowledge about the science domain, and knowledge about experimentation practices in order to make decisions about how to investigate complex questions. They need to focus their inquiry on the most salient issues. To make sensible decisions about experimental designs that test the multitude of ideas they hold, learners need to combine their knowledge of combinatorial reasoning and controlling variables with methods for sorting out their disciplinary knowledge and identifying compelling questions. Learners must weigh multiple sources of knowledge to conduct informative experiments.

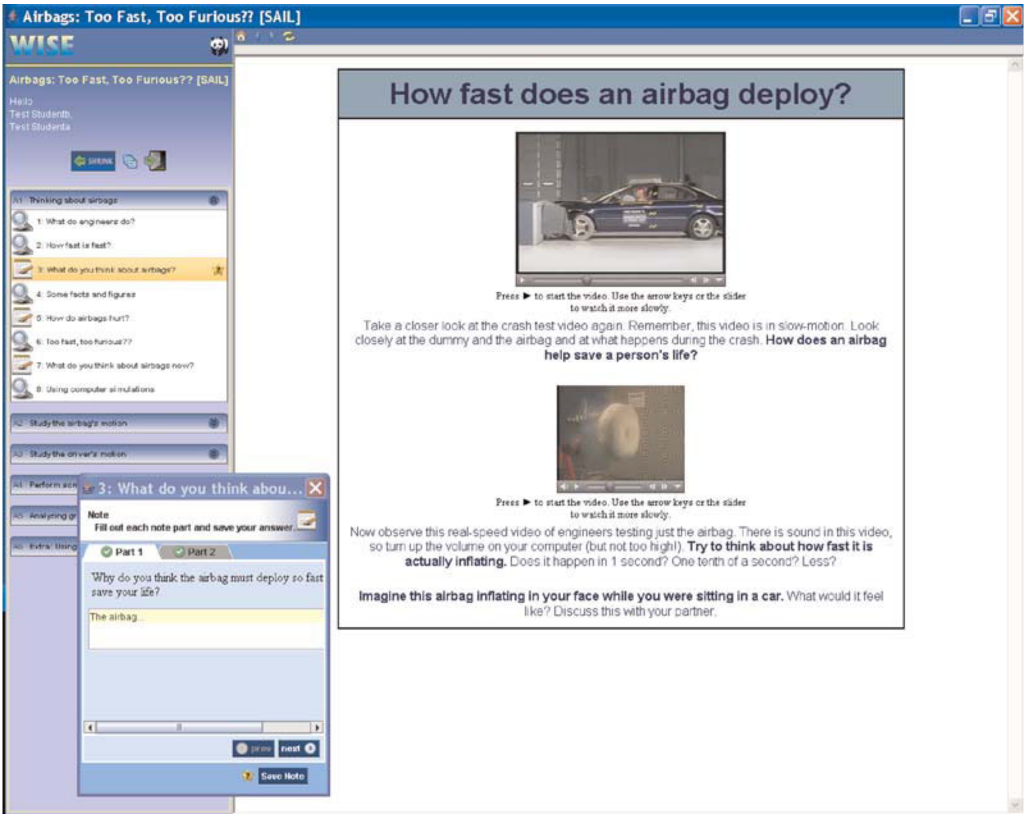
Several research programs integrate computer-supported, student-initiated experiments with broader investigations. The Computer as Learning Partner (Linn & Songer, 1991) project gave students opportunities to conduct hands-on experiments as part of a larger inquiry unit on heat and temperature. In Thinkertools (White, 1993; White & Frederiksen, 1998) students conduct experiments both in the real world and within a microworld to develop and refine conceptual models relating force and motion. In the Galapagos Finches environment (Reiser et al., 2001), students explore a large dataset by conducting comparisons in a manner similar to experiments, with the goal of relating environmental characteristics of an island habitat and characteristics of the island's finch population. These studies draw on students' explanations as well as the validity of students' experimentation methods. They illustrate the importance of helping students integrate experimentation methods and understanding of the relevant disciplinary ideas.

Researchers studying realistic tasks have often categorized both the processes and goals of learners. Many distinguish whether students control or confound variables (e.g., Tschirgi, 1980; Vollmeyer, Burns, & Holyoak, 1996). Some categorize the comprehensiveness (whether students explore all the variables) of the investigation (Klahr & Dunbar, 1988; Schauble, 1996; Tabak & Reiser, 2008). In categorizing outcomes researchers have distinguished scientific and engineering approaches (Schauble, 1996). This study extends these findings to identify how students in high school physics—generally the last science class taken in high school—combine processes and goals to explore a complex science task.

Curriculum Design

The National Research Council (2007) suggests developing ‘‘integrated instructional units’’ (p. 4) that connect laboratory experiences with other types of instruction to provide authentic classroom experiences. These units incorporate experimentation activities within inquiry investigations, provide learners with opportunities to test their own ideas about the domain, and use the outcomes of experimentation to generalize knowledge to new contexts. They blur the line between methods of investigation and scientific ideas. They lead to authentic and normative views of science. Well-designed inquiry projects can scaffold learners to successfully navigate complex tasks by helping students pull together multiple sources of evidence (Quintana et al., 2004) and make science accessible to diverse learners (Wilson, Taylor, Kowalski, & Carlson, 2010).

We created *Airbags* to achieve the goals of an integrated instructional unit. *Airbags* is a 1-week curriculum unit for high school physics classes designed using WISE to scaffold inquiry. A screenshot of the first activity of *Airbags* appears in Figure 1. WISE allows instructional designers to build inquiry units using steps that promote scientific understanding. In WISE units, students view evidence, compose reflection notes, engage in online discussions or debates, interact with visualizations, and use drawing tools to illustrate their ideas. *Airbags* has two primary learning goals. The main disciplinary learning goal of *Airbags* is the relationship between the nature of one-dimensional motion and the characteristics of position and


Airbags: Too Fast, Too Furious?? [SAIL]
WISE
Airbags: Too Fast, Too Furious?? [SAIL]
Hello
Test Studentb,
Test Studenta
SHRINK
A1 Thinking about airbags
1: What do engineers do?
2: How fast is fast?
3: What do you think about airbags?
4: Some facts and figures
5: How do airbags hurt?
6: Too fast, too furious??
7: What do you think about airbags now?
8: Using computer simulations
A2 Study the airbag's motion
A3 Study the driver's motion
A4 Perform sc
A5 Analyzing gr
A6 Extra: Using
3: What do you think abou...
Note
Part 1
Part 2
prev
next
Save Note

Fill out each note part and save your answer.

Why do you think the airbag must deploy so fast save your life?

The airbag...

How fast does an airbag deploy?

Press ▶ to start the video. Use the arrow keys or the slider to watch it more slowly.

Take a closer look at the crash test video again. Remember, this video is in slow-motion. Look closely at the dummy and the airbag and at what happens during the crash. **How does an airbag help save a person's life?**

Press ▶ to start the video. Use the arrow keys or the slider to watch it more slowly.

Now observe this real-speed video of engineers testing just the airbag. There is sound in this video, so turn up the volume on your computer (but not too high!). **Try to think about how fast it is actually inflating.** Does it happen in 1 second? One tenth of a second? Less?

Imagine this airbag inflating in your face while you were sitting in a car. What would it feel like? Discuss this with your partner.

Figure 1. The web-based inquiry science environment, showing the first activity of *Airbags*.

velocity graphs. *Airbags* addresses students' difficulties with connecting graphs and physics (McDermott, Rosenquist, & Zee, 1987), differentiating between the height and slope of a graph (Leinhardt, Zaslavsky, & Stein, 1990), and distinguishing position, velocity, and acceleration (Trowbridge & McDermott, 1980, 1981). These topics constitute part of national science and mathematics standards (American Association for the Advancement of Science, 1993; National Council of Teachers of Mathematics, 2000).

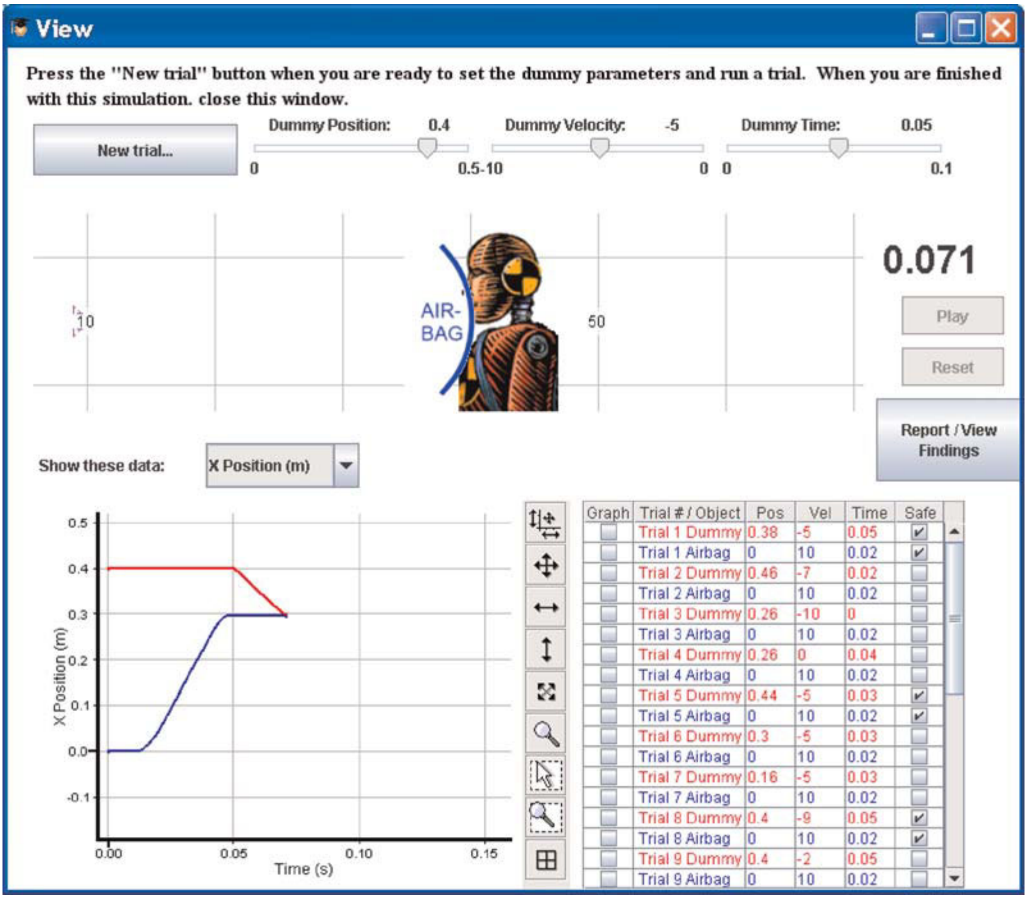
The second learning goal is the goal of the inquiry investigation, which is to understand the dynamics of airbag deployment and how they govern the risks for injury from an airbag in a head-on collision. In *Airbags*, students investigate factors that lead to a high risk for injury to the driver from an airbag. We discuss the specific variables students investigate in more detail below. The design of *Airbags* integrates the inquiry goal with the disciplinary learning goal by requiring students to use motion graphs to further their understanding of the collision dynamics.

We designed *Airbags* using the knowledge integration framework to promote coherent understanding (Kali, 2006; Linn & Eylon, 2006). The knowledge integration framework describes learners as simultaneously holding multiple, sometimes conflicting, and often isolated ideas about science (Linn & Hsi, 2000). The instructional sequence of *Airbags* is guided by the knowledge integration pattern (Linn & Eylon, in press). The scaffolds help students organize and link their diverse ideas into a more coherent understanding of motion graphs. The pattern involves *eliciting students' current ideas* to ensure that new ideas connect to existing knowledge. Following the pattern, instruction *adds new, normative ideas* in ways learners can easily integrate into their understanding. Students then conduct experiments to *distinguish* among the alternatives in their repertoire of ideas including new ideas and their prior ideas. Finally, students *sort out and refine their ideas* by reflecting on their understanding, identifying gaps in their understanding, and repairing these gaps.

Airbags activities support use of experimentation to understand a complex phenomenon. The first activity orients students to the investigation context of airbags and elicit their ideas from crash test videos about what factors affect a driver's risk for injury in a collision. The second and third activities use dynamic visualizations to add and distinguish ideas about the relationship between motion and graphs. The fourth activity incorporates the experimentation environment (described below), which students use to investigate the factors leading to driver injuries from airbags. The fifth and sixth activities prompt students to construct arguments that bring together multiple sources of evidence from the unit, helping students assess the quality of their understanding.

Design of the Virtual Experimentation Environment

This paper focuses on the experimentation activity of *Airbags*. This activity uses a virtual experimentation environment (Figure 2) that illustrates of the interaction between the airbag and driver during a head-on collision, using the steering wheel as a point of reference. The visualization was designed using the modeling environment Dynamica by The Concord Consortium (http://www.concord.org). Students use the experimentation environment to investigate three questions concerning airbag safety: (1) Why are shorter drivers at greater risk for injury than taller drivers? (2) Are drivers at greater risk for injury in high speed or low speed collisions? and (3) How does the car's propensity to "crumple" during the collision affect the driver's risk for injury? These three questions map onto the three motion variables students can manipulate in the visualization: (1) the initial position of the driver (ranging from zero to 0.5 m from the steering wheel), (2) the velocity of the driver toward the airbag after impact (ranging from zero to 10 m/second toward the steering wheel), and (3) the time over which



Press the "New trial" button when you are ready to set the dummy parameters and run a trial. When you are finished with this simulation. close this window.

Dummy Position: 0.4

Dummy Velocity: -5

Dummy Time: 0.05

Show these data: X Position (m)

Graph	Trial # / Object	Pos	Vel	Time	Safe
□	Trial 1 Dummy	0.38	-5	0.05	☑
□	Trial 1 Airbag	0	10	0.02	☑
□	Trial 2 Dummy	0.46	-7	0.02	□
□	Trial 2 Airbag	0	10	0.02	□
□	Trial 3 Dummy	0.26	-10	0	□
□	Trial 3 Airbag	0	10	0.02	□
□	Trial 4 Dummy	0.26	0	0.04	□
□	Trial 4 Airbag	0	10	0.02	□
□	Trial 5 Dummy	0.44	-5	0.03	☑
□	Trial 5 Airbag	0	10	0.02	☑
□	Trial 6 Dummy	0.3	-5	0.03	□
□	Trial 6 Airbag	0	10	0.02	□
□	Trial 7 Dummy	0.16	-5	0.03	□
□	Trial 7 Airbag	0	10	0.02	□
□	Trial 8 Dummy	0.4	-9	0.05	☑
□	Trial 8 Airbag	0	10	0.02	☑
□	Trial 9 Dummy	0.4	-2	0.05	□
□	Trial 9 Airbag	0	10	0.02	□

Figure 2. The experimentation environment presents an animation, a graph, and students' experimentation history.

crumpling occurs, modeled as the time between the car's impact and the driver's initial motion relative to the steering wheel (ranging from zero to 0.1 seconds). The visualization assumes that the drivers' motion prior to the car's crumpling is negligible. This assumption is supported by observations of crash test videos.

Students conduct experiments to answer each investigation question. To conduct an experimental trial, students first select an investigation question from a drop down menu (or choose an alternative such as *Just exploring*). After specifying their goal, students use sliders to specify the values of the position, velocity, and time variables and run the crash simulation. When the simulation runs to completion, students can judge whether the trial was "safe" or "unsafe" and record this outcome in the experimentation history. In previous activities, students determine that a driver must encounter an airbag after it has finished inflating in order to be "safe." While students conduct their experiments, the software logs the investigation question and variable values students select for each trial. In this way, requiring students to choose an investigation question for each trial provides information about students' intentions for each trial.

In *Airbags*, two types of relationships govern the risk for injury to the driver from an inflating airbag. 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.

The threshold values lead to complex relationships between collision outcomes and variable values. Some controlled comparisons illustrate more important aspects of the system than others. For instance, if a student keeps the position of the driver below the threshold value, a pair of controlled trials at different velocity values will yield identical "unsafe" outcomes. In order to make relevant insights about the system, students must consider not only whether they have varied only one variable, but also whether the outcomes of their trials provide evidence for the variable relationships they intend to illustrate.

Methods

Participants and Implementation

Six high school physics teachers used *Airbags* in the classrooms, encompassing 148 students across the United States. Table 1 illustrates the social diversity of school settings where students studied *Airbags*. Three of the teachers were experienced and had taught previous versions of *Airbags*, though none of the students had previously used WISE. All teachers participated in targeted professional development (Varma, Husic, & Linn, 2008). Most students worked in dyads on the activities. Students were assigned in groups of similar ability as judged by their teacher. Unpaired students used their own computers while working jointly with another group. At all six schools, every student taking physics at the school participated in this study (some schools had low enrollment in physics). As a result, our results apply to students who take physics at schools with these characteristics. At each school the unit occurred shortly after the students' regular classroom instruction on kinematics.

Table 1
Summary of Airbags classroom implementations

School	Description of School Setting	Ability	*n*	# Classes
1	Wide geographical area, magnet program[a]	Honors	38	2
2	Suburban, 18% reduced lunch, 41% African-American, Asian, and Hispanic	Mixed	15	1
3	Urban, 67% reduced lunch, 89% African-American, Asian, and Hispanic	Mixed	28	1
4	Suburban, 31% reduced lunch, 61% African-American	Mixed	12	1
5	Urban, 54% reduced lunch, 95% African-American	Mixed	9	1
6	Suburban, 52% reduced lunch, 81% African-American, Asian, and Hispanic	Mixed	46	3

[a] School 1 is comprised of students who are officially enrolled at other schools. Data for School 1 were therefore unavailable.

At all schools except for School 3, a researcher (generally the first author) was present in the classroom as a co-teacher alongside the students' regular classroom teacher. The researcher played an active role in the implementation of *Airbags*, engaging students in individual and whole class discussion, responding to students' questions about the curriculum and content, asking students for verbal explanations for their inquiry choices, and supporting the teacher in using the technology.

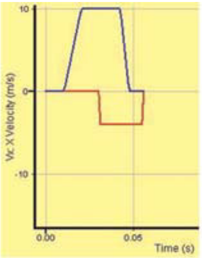
Classroom Observations

To document the fidelity of implementation, the researcher recorded and summarized observations. The researcher wrote brief notes of students' activities during class. Each day, the researcher summarized the notes, recorded any technical difficulties, and described overall progress. At the end of the unit, the researcher discussed students' experience with *Airbags* with the teacher and noted things that should be improved for the next classroom implementation.

Content Knowledge Assessments

To investigate the impact of *Airbags* on student learning, we used both pretest/posttest and embedded assessments. We aligned the instruction, assessments, and scoring rubrics using the knowledge integration framework (Linn & Eylon, in press). The assessments measured progress in developing coherent, integrated understanding of motion, graphs, and airbag safety. Examples of pretest and embedded assessment items and the rubrics we used to score them appear in Figure 3 and in Tables S1–S4. All pretest items are available as supplementary material accompanying the online article.

Describe what happened between the driver and airbag in this crash.
Was the driver injured by the airbag? Explain based on the graph.



Score	KI level	Description
0	Blank	No response
1	None	Off task
2	Isolated/ Invalid/ Irrelevant	*One of the following incomplete interpretations* - description of collision outcome without justification - description of airbag deployment without connection to outcome
3	Partial	Connection between safe outcome and complete airbag deployment but without reference to motion attributes of airbag or driver
4	Full	Connection between safe outcome and driver-airbag motion attributes (the airbag stops moving before the driver stops moving)

Figure 3. An embedded interpretation item and the rubric we used to score it. The item asks students to interpret a velocity versus time graph of the driver-airbag collision.

Pretest/Posttest Assessments. We used pretests and posttests to measure how well students could generalize their understanding of motion graphs from *Airbags* to new motion contexts and articulate criteria for injury risk to a driver by a deploying airbag. Pretests and posttests were administered to individual students the day before the start of implementation and the day after completion. Posttests covered the same issues as the pretests but were changed slightly to reduce possible gains due to retesting, except in School 3 where the pretest was mistakenly administered as the posttest. Due to absences, some students did not take either the pretest or the posttest. Pretests and posttests consisted of 11 constructed response items. Ten of the items addressed motion graphs (four on interpretation and six on construction) and were scored from zero to four. We were able to use only one item that addressed airbag safety due to class time constraints. This item captured students' improvement in understanding the broad investigation context of airbag deployment. We scored the airbag safety item from zero to five. Students in School 3 did not receive the airbag safety item because of additional time constraints.

We scored pretest and posttest items using a knowledge integration rubric (Linn, Lee, Tinker, Husic, & Chiu, 2006) that rewards valid scientific connections between concepts. The total pretest and posttest scores were the sum of the scores of the individual items, justified by Item Response Theory studies on knowledge integration items (Liu, Lee, Hofstetter, & Linn, 2008). Where appropriate, we examined the individual contributions of students' performance on the motion graphs items and the airbag safety item, as illustrated in subsequent sections of the paper.

Interpretation Score. We used six constructed response embedded assessment items following the experimentation activity to measure students' understanding of the airbags situation. These items (Figure 3) presented position and velocity graphs of the airbag-driver collision (as coming from a hypothetical "black box" data collector) and asked students to explain why the driver was safe or unsafe. Students responded to these prompts in their working groups. These items measure whether students are able to link their understanding of the motion variables to the collision outcomes. The interpretation items therefore capture more than just whether students observed covariation between variables and outcomes. Students are rewarded for their ability to interpret the sequence of collision events and to apply their understanding of motion graphs to assessing the safety of the driver. We scored these items on a scale from zero to four using knowledge integration rubrics (Figure 3). The total interpretation score was the sum of the scores on the six items.

To determine inter-rater reliability, two independent coders coded a random subset of the students' responses on the pretests and embedded prompts. (Posttest rubrics were structurally identical to pretest rubrics.) Inter-rater reliability was 93%.

Experimentation Scores

Using logging software Pedagogica (Buckley, Gobert, & Horwitz, 2006) we recorded the investigation question and variable values students chose for each trial. We used the reports of students' trials to characterize each student groups' experimentation approach in four ways.

Total Trials. We computed the total number of trials conducted by each group. Students occasionally conducted identical trials so we also computed the number of unique trials each group conducted. Unique trials correlated highly ($r = 0.95$) with total trials, so we used total trials in the analysis.

Trial Variability. To measure how widely students changed the variable values, we computed a variability score. We computed for each of the three investigation variables (1) the number of unique values, (2) the range of values tested, and (3) the number of boundary values tested by each student group. The three subscores exhibited an internal consistency (Cronbach's α) of 0.91, suggesting that the mean of the subscores provides a reliable overall measure of the variability of students' experimentation. Therefore we expressed each of these values as a fraction of the maximum attainable value for each investigation variable, computed the mean of these three fractions to generate a subscore for each investigation variable, then computed the mean of these three subscores to generate the overall variability score scaled from zero to 100.

VOTAT Score. To measure students' propensity to Vary One Thing At a Time (VOTAT) we computed the fraction of consecutive trial pairs where students changed the value of exactly one of the three variables. This proportion reflects the extent to which students could observe the effect of changing just one variable on the outcome—but does not tell us whether this was the goal of the comparison.

Coherence Score. To estimate the coherence of students' goals and experimentation choices, we measured the alignment between the students' investigation goals and the experiments they conducted. Only trials where students selected one of the three investigation questions were used for this score.

We scored each series of consecutive trials for the same investigation question using a knowledge integration rubric from zero to five (Table 2). We designed this rubric to capture the strength of the link between students' investigation goals and their variable choices in several ways. First, the rubric rewards conducting at least two unique trials for a particular investigation question, as comparisons between multiple trials are essential for illustrating variable relationships. Second, the rubric rewards varying the variable that corresponds to the chosen investigation question for that comparison. Third, the rubric rewards controlled comparisons that produce evidence for a variable effect, as measured by achieving opposite outcomes (safe or unsafe).

To compute the coherence score, we used the highest score achieved for each investigation question as a coherence subscore. The three subscores exhibited an α consistency of 0.75. The overall coherence score was the mean of the three subscores.

Table 2
Knowledge integration rubric for the experimentation coherence score

KI Level	Score	Description
Blank	0	No trials conducted
None	1	One trial conducted, or no variables changed between two consecutive trials
Isolated	2	At least two trials conducted with all three variables changed between trials, or with the investigation variable left unchanged
Partial	3	At least two trials conducted with the investigation variable and one other variable changed between trials
Full	4	At least two trials conducted with only the investigation variable changed between trials, achieving the same outcome (safe or unsafe)
Complex	5	At least two trials conducted with only the investigation variable changed between trials, achieving opposite outcomes (safe and unsafe)

Rubric is applied to each set of consecutive trials for the same investigation question. If no trials exist for an investigation question, the coherence subscore for that question is zero.

Analysis

In School 1, a subset of students' records failed to upload to the servers on 1 day of the unit. In order to confirm the completeness of their experimentation logs, students at this school reported how many experimentation trials they conducted on a survey. Six student workgroups (12 students) at this school whose self-reports differed obviously from the incomplete uploaded information were removed from analysis. Eleven student workgroups (19 students) at all the schools who failed to respond to at least 75% of the unit's prompts due to class absences were also removed. To measure pretest to posttest learning gains, we used a two-tailed, paired *t*-test for the motion graph items and a Wilcoxon signed-rank test for the airbag safety item.

Because of the small class sizes in our study, we pooled the students from all schools and controlled for prior knowledge using scores on the motion graph pretest items. To justify this decision, we examined school effects on student learning from *Airbags* by conducting a one-way analysis of variance (ANOVA) on the motion graph items (which were common to all students), using school as the between-groups factor. These analyses showed significant school effects on both pretest [$F(5, 108) = 17.35$, $p < 0.001$] and posttest [$F(5, 107) = 6.38$, $p < 0.001$]. When School 1 (the only school with selective admission criteria) was removed from the analysis the effect for school on the pretest was significant [$F(4, 87) = 3.70$, $p < 0.01$], but for the posttest was not significant [$F(4, 86) = 1.39$, $p = 0.25$]. This analysis of school effects provides evidence that the students' experience with *Airbags* and what they learned was similar across typical school settings.

Though some of the class sizes are quite small, taken together, the students from these six schools capture much of the diversity of physics classes across the United States. This diverse sample of students allows us to examine the effectiveness of *Airbags* across a range of classroom settings and to characterize the different types of experimentation approaches students use to investigate a realistic problem such as *Airbags*.

We used standard multiple linear regression models (Agresti & Finlay, 1997) to analyze students' performance on embedded assessments, experimentation measures, and students' pretest and posttest scores. Multiple regression models estimate the average value of the explanatory variable while the other variables are held fixed. The regression coefficient of each variable is, therefore, the partial effect of that variable, controlling for the other variables in the model.

Because most students responded to embedded assessments and the experimentation activity in dyads, we used the mean pretest/posttest score for each dyad in these regression analyses. Student groups of either one or (usually) two students comprise our unit of analysis for these regression models. To confirm that the embedded assessment responses and experimentation measures reflect contributions from both students within dyads (rather than just the stronger student), we compared the regression analyses using the highest pretest/posttest score from each dyad to the analysis using the mean score and found that the results were virtually identical.

Results

Fidelity of Implementation

Overall, 75% of the students responded to at least 85% of the embedded prompts in *Airbags*, illustrating that the majority of students engaged with the *Airbags* unit as intended. Two schools (3 and 6) typically had high rates of absenteeism. Student absences and the uploading error described above account for nearly all of the missing responses. The researcher in the classroom observed that students were generally on task and that dyads discussed their responses prior to entering them. Students were generally motivated by their teachers to complete the project. Teachers used WISE teacher tools to grade students' responses to selected questions and to provide regular feedback in the form of notes to the students. Except for the uploading error, there were no major technical glitches during the six classroom implementations.

Overall Impact of the Unit

Students made significant pretest to posttest gains on the motion graph items [$M = 22.61$, SD $= 10.57$ (pre); $M = 28.08$, SD $= 7.39$ (post), $t(108) = 6.40$, $p < 0.001$, $d = 0.60$] and the airbag safety item [$M = 2.06$, SD $= 0.96$ (pre); $M = 3.76$, SD $= 1.17$ (post), $Z = 7.58$, $p < 0.001$, $d = 1.58$]. Using Cohen's criteria (1988), these gains are considered moderate and large, respectively. For all pretest/posttest items taken together, gains were positive at all schools and significant ($p < 0.05$) for every school except for School 1 (which had very high pretest scores) and School 5 (which had just nine students). Table 3 shows the total pretest scores, posttest scores, and effect sizes for students at each school. Considering that *Airbags* typically requires just 4–5 hours of class time, these gains in understanding of motion graphs across the six schools illustrate the success of *Airbags* in promoting understanding among diverse student populations. To determine whether dyads learned from each other we tested for a change in magnitude of the difference between individual scores within dyads from pretest to posttest and found a significant decrease ($t = 2.11$, $p = 0.04$). Consistent with our classroom observations, this suggests that both students within dyads participated actively and that their understanding converged during the unit.

Relationship Between Students' Experimentation Scores and Posttest Performance

To investigate the relationships among scores and posttest performance, we used multiple linear regression with the four experimentation scores (total trials, trial variability, VOTAT score, coherence score) and the group pretest scores as explanatory variables and the group posttest scores as the response variable. The regression model revealed that none of the relationships between the experimentation scores and the posttest scores were significant. Pretest scores were a much stronger predictor of posttest scores than any of the experimentation scores. The weak relationships between experimentation scores and posttest scores show that

Table 3
Pretest and posttest means, standard deviations, and effect sizes for each school

		Pretest		Posttest		
School	*n*	*M*	SD	*M*	SD	Cohen's *d*
1	22	37.73	3.01	39.23	4.85	0.37
2	12	27.67	8.17	33.75	9.40	0.69*
3[a]	19	18.16	9.55	27.78	7.27	1.13***
4	9	25.78	8.32	30.11	7.32	0.55*
5	9	25.44	11.26	29.89	6.56	0.48
6	38	17.95	8.79	27.97	6.77	1.28***

[a] The scores of School 3 include just 10 of the 11 total items administered to the rest of the schools.
*$p < 0.05$.
***$p < 0.001$.

the mechanics of the investigation each group carried out is not a primary determinant of learning outcomes as measured by the graph interpretation items on the posttest. This is consistent with the many opportunities students had to improve their understanding in addition to the experimentation activity (which lasted between 20 and 30 minutes of the 4–5 hours unit). Students had an opportunity to learn from numerous graph interpretation and construction activities and from reflection prompts. In addition, much of the focus of the experimentation activity was on understanding the nature of the variables in the *Airbags* investigation which was tapped by only one of the posttest items.

Relationship Between Experimentation Scores and Embedded Assessment Performance

We investigated the relationships between the four experimentation scores and the students' responses to the embedded prompts as represented in the interpretation score, controlling for students' prior knowledge using group pretest scores. We used multiple linear regression with the four experimentation scores (total trials, trial variability, VOTAT score, coherence score) and the group pretest scores as explanatory variables and the interpretation score as the response variable. Table 4 lists the regression coefficients. The coherence score was the only significant predictor in the regression model. The standardized coefficients (β) indicate that the coherence score was an even stronger predictor of the interpretation score than students' prior content knowledge.

The strong relationship between the coherence score and the interpretation score, even after controlling for students' pretest scores, reflects the success of students who were able to plan their investigations in advance and use appropriate methods to test their ideas. The weak relationship between pretest scores and the interpretation score shows that successful students benefited from their investigations and that their success does not necessarily reflect their prior knowledge about physics. It is also revealing that the coherence score is a powerful predictor of the interpretation score after controlling for students propensity to vary one variable at a time. This result suggests that understanding of the investigation context, and not

Table 4
Summary of regression analysis for predicting students' interpretation scores

Predictor	B	SE B	β
All groups ($N = 48$)			
Pretest	0.08	0.06	0.19
Total trials	–0.07	0.05	–0.28
Variability	1.50	2.19	0.12
VOTAT proportion	–0.99	2.25	–0.07
Coherence score	1.22	0.49	0.53*
Low trials (≤ 12, $N = 22$)			
Pretest	0.01	0.08	0.03
Total trials	0.38	0.34	0.29
Coherence score	0.54	0.67	0.22
High trials (>13, $N = 26$)			
Pretest	0.16	0.08	0.40*
Total trials	–0.09	0.04	–0.41*
Coherence score	0.93	0.51	0.39†

$^\dagger p < 0.1$.
$^* p < 0.05$.

merely a procedural understanding of controlling variables, distinguished more successful students from less successful ones.

The weak relationship between students' trial variability and the interpretation score suggests that trial design alone is not sufficient to determine whether students gained insight from their investigation. Students can gain insight into the *Airbags* situation by varying variables widely and exploring the full range of values as well as by making minute changes or testing just a few unique values.

The weak relationship between the VOTAT score and the interpretation score also is consistent with the possibility that students follow the procedure of controlling variables without making sense of the findings. Conducting informative controlled tests involves more than mechanically varying one variable at a time. Students need to align their experimentation methods with investigation goals with the intention of testing their ideas.

We were initially somewhat surprised by the negative relationship we observed between total trials and the interpretation score. We expected that, all other things being equal, more opportunities to examine the variable relationships would benefit learners. We conjectured that this negative relationship was a result of unplanned or haphazard experimentation approaches used by students who conducted a very large number of trials. To test this conjecture, we divided students into two groups according to the number of trials they conducted and generated separate regression models for the high trials and low trials group. We dropped the variability and VOTAT scores from these models, as they were weak predictors in the previous model. We observed a positive relationship between total trials and the interpretation score for the low trials students, and a significant negative relationship for total trials and the interpretation score for the high trials students. We observed that the strongest relationships (as measured by the standardized regression coefficient β) occurred when we used 12.5 trials as the dividing point between the high and low groups. This result supports our conjecture and suggests that conducting a very large number of trials may reflect students' failure to intentionally investigate the variables and interpret trial outcomes. Table 4 includes the results of all three regression models relating experimentation and the interpretation score.

Overall, the regression analysis indicates that students who linked their investigation goals and experimentation methods using a moderate number of trials were the most successful on the embedded assessments. A high coherence score reflects several dimensions of students' knowledge other than just being able to isolate variables. First, students must map the investigation questions onto the appropriate variables. Second, students must correctly interpret the outcomes of their trials (as safe or unsafe). Third, because students must articulate their investigation goal before conducting each trial, the coherence score measures students' ability to plan investigations in advance. The findings thus highlight several aspects of informative experimentation. In addition to isolating variables, connecting experimentation practices to investigation goals, building on everyday knowledge of the situation, and planning in advance can all contribute to valid inferences.

Variation in Students' Experimentation Methods

To illustrate the variation in students' experimentation methods, we first plot the two significant predictors of learning on the embedded assessments (coherence score and total trials) against each other (Figure 4a) and make a few general observations. For methods involving relatively few trials, coherence scores are necessarily low. Students must conduct at least two trials for each of the three investigation question in order to achieve a coherence score of more than 3. The smallest number of trials a student group conducted while fully investigating each question was 8.

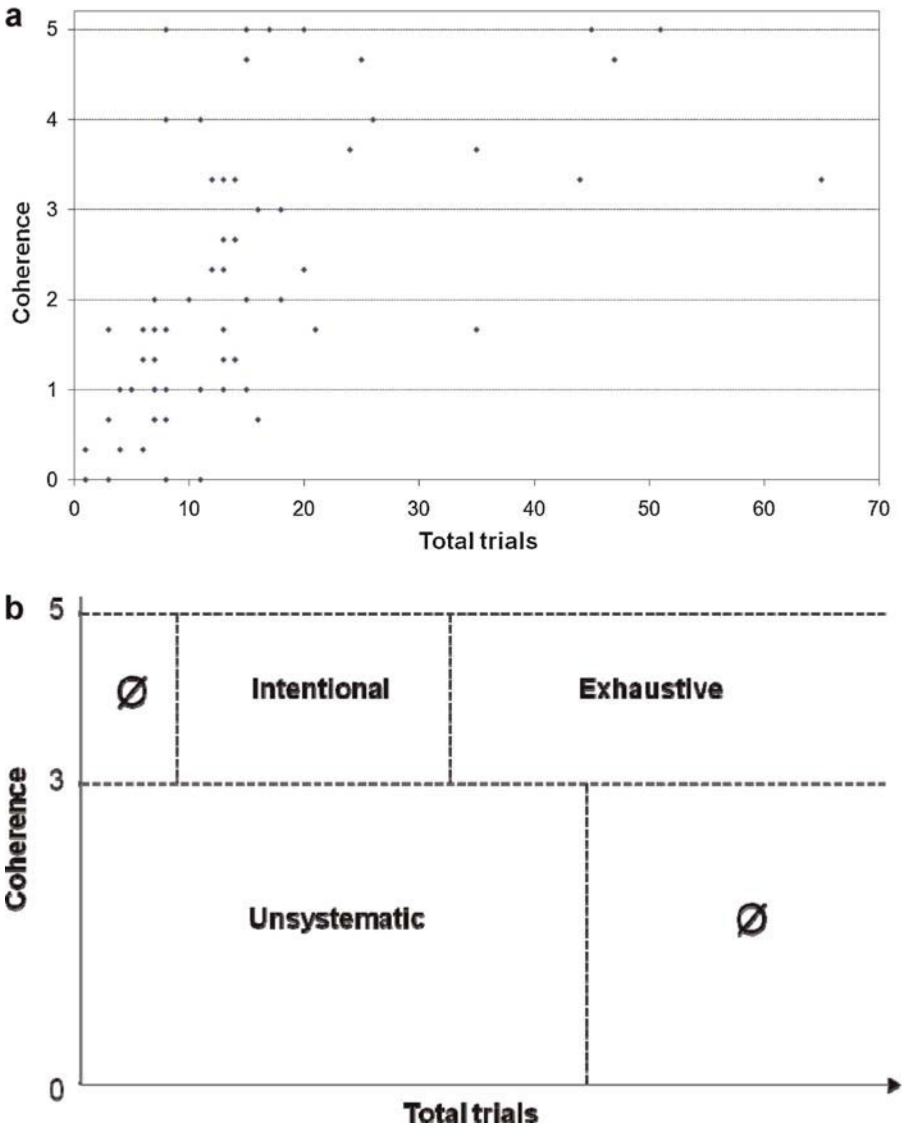


Figure 4. **a**: Scatter plot of all groups' experimentation coherence score versus the total number of trials they conducted. **b**: Schematic representation of three experimenter categories on the coherence-total trials space. (Ø) Represents areas of the space where experimentation approaches are unlikely to occur.

Similarly, for methods involving many trials, coherence scores are likely to be high. Students can achieve high coherence scores by exhaustively sampling the experimentation space if they sometimes choose to vary just one variable between trials. Except for one outlying group (who specified an investigation question for just 11 of their 35 total trials), every group that conducted at least 22 trials achieved a coherence score of greater than 3.

Based on the results of the regression analysis above we identified three broad categories for students' experimentation approaches. We refer to these categories as *intentional*, *exhaustive*, and *unsystematic*. We illustrate in Figure 4b where these categories lie in the coherence-total trials space. Although the boundaries between categories are not sharp, they capture the primary focus of student experimentation for that group. We selected a coherence score of 3 as the lower boundary for intentional and exhaustive experimenters, as this represents on average a partial link between students' investigation goals and experimentation methods. For the students in this study, intentional experimentation was achieved within 25–30 trials. A small number of students who conducted over 30 trials attempted to exhaustively test the full range of each variable but were less systematic in linking their goals to their experiments than students who conducted fewer trials. Most students scoring below coherence level 3 conducted seemingly haphazard or incomplete sequences of trials.

Using these criteria, we find that 67% of student groups fell into the unsystematic category, 24% of groups fell into the intentional category, and 9% of groups fell into the exhaustive category. This distribution is consistent with other research that shows many students and adults do not conduct experiments systematically (Schauble, 1996). Although some students may have been able to gain relevant insights about the airbags situation using diverse methods not captured by our experimentation measures, these measures are helpful indicators of experimentation practices. They demonstrate the need for guidance and support in complex investigations.

To illustrate the value of these categories, we discuss the experimentation sequences of three typical dyads, one for each experimentation category. To select typical students, we chose dyads with similar pretest scores, all coming from the middle 50% of our student population.

To succinctly capture the experimentation practices of each group, we created a visual representation showing the variable values each dyad chose for each trial (Figures 5–7). The representation makes apparent when the group made changes to one or more of the variables. The dashed boxes indicate consecutive trials where students varied exactly one variable that did not correspond to their investigation goal. The solid boxes indicated where students varied only the variable corresponding to their investigation goal.

Intentional Experimenters

The intentional experimenters use a moderate number of trials to intentionally investigate each variable. These students are able to articulate their investigation goals in advance and conduct trials that illustrate the effects of specific variables. For example, a group in the 32nd percentile of all groups with a pretest score of 24.5 (out of a possible 40) conducted eight trials and investigated the variable they planned to study in two out of three sequences (Figure 5). They earned a coherence score of 4.00.

This group first tested the default variable values. They used their next two trials to conduct a controlled comparison for the position variable, though they specified their intention to do so only for trial 2. They then conducted two controlled trials for the velocity variable and three controlled trials for the time variable, fully specifying their intentions for both sets of trials. Furthermore, each of these sets of trials demonstrated both safe and unsafe outcomes, providing evidence that these students appreciated the importance of distinguishing between the conditions that led to each outcome. Finally, the sophistication of this groups' controlled tests appears to have improved, progressing from an unplanned pair of controlled trials, to a planned pair of controlled trials, to a controlled set of three trials. This progress suggests that these students were able to refine their approach over time. These students scored 18 (out of a

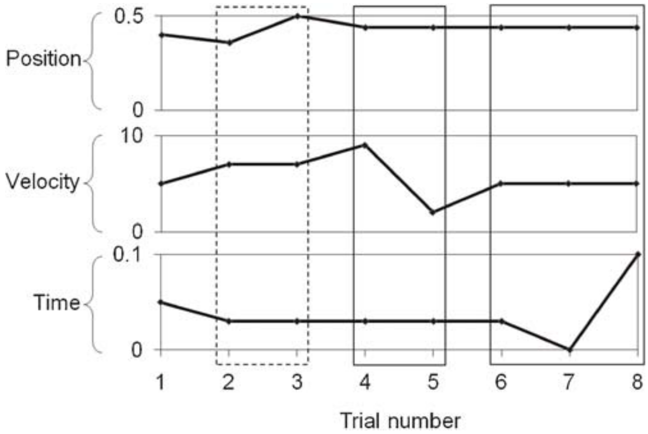


Figure 5. Experimentation sequence illustrating the approach of an intentional experimenter.

possible 24) on the interpretation items, a fairly high score relative to their pretest performance, and near the median score of 18.5.

Unsystematic Experimenters

Unsystematic experimenters do not link their experimentation methods to the investigation goals. Unsystematic approaches are characterized by lack of alignment between goals and experimental trials and by failing to vary the appropriate variable between trials. For example, a group with a pretest score of 32 (65th percentile of all groups) conducted 8 trials with a coherence score of 1.67 (Figure 6).

These students' experimentation approach illustrates several characteristics of unsystematic experimentation. First, when the students varied one variable at a time (Trials 1–2 and 7–8), this variable did not reflect the investigation questions they specified. Second, they did not adequately investigate two of the three collision factors (height and crumple time), conducting just one trial corresponding to each of these questions. Third, when they did conduct two trials for the velocity investigation question (Trials 4–5) students varied the crumple time in a

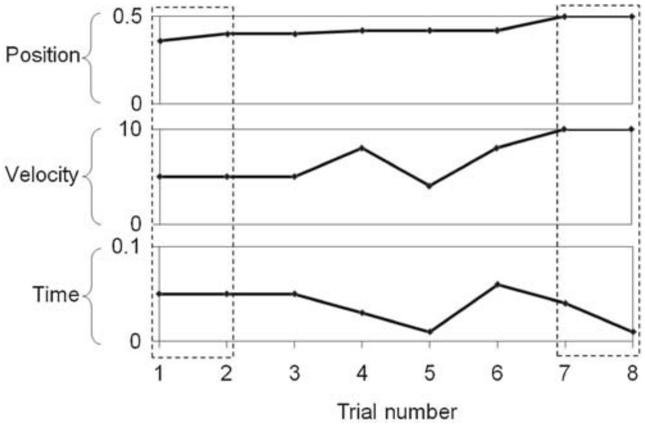


Figure 6. Experimentation sequence illustrating the approach of an unsystematic experimenter.

way that negated the effect of the velocity. As a result they produced similar outcomes for the two trials and could not detect the effects of individual variables on the outcome. In the other comparison investigating velocity (Trials 7–8), the students did not vary the velocity variable at all. Despite attaining an above-average pretest score and conducting the same number of trials as the intentional group, these students' experimentation method was unsystematic and yielded little information, as evidenced by their low interpretation score of 15 (24th percentile of all groups).

Exhaustive Experimenters

Like intentional experimenters, exhaustive experimenters generate enough evidence to draw valid conclusions about the situation. However, they do not intentionally link their experimentation methods to their investigation goals and do not keep track of their findings. They conduct a few controlled trials among many additional trials that appear to have the goal of considering every alternative. These students neglect planning and do not take advantage of the evidence they generate with controlled comparisons. As an example, a group with a pretest score of 27 (44th percentile of all groups) conducted 47 trials with a coherence score of 4.67 (Figure 7).

These students changed the investigation question 18 times during their 47 trials and duplicated trials 13 times. Their high coherence score reflects the large number of trials they conducted for each investigation variable. Because they did not carefully link their experiments to their goals, these students also achieved a low interpretation score of 15. They were unable to produce sophisticated insights despite having enough evidence to do so.

These three examples illustrate the diverse approaches students used to investigate *Airbags*. The lack of correspondence between these students' experimentation approaches and their pretest scores supports the finding that students' prior knowledge of physics does not strongly determine sophistication of their investigation methods. Some students with high pretest scores conducted unsystematic, uninformative experiments, while some students with low pretest scores conducted well-planned experiments. Well-planned experiments tested the relationships among the variables. They also illustrated how minor variations in conditions can lead to dramatic differences in outcomes (safe versus unsafe for the driver).

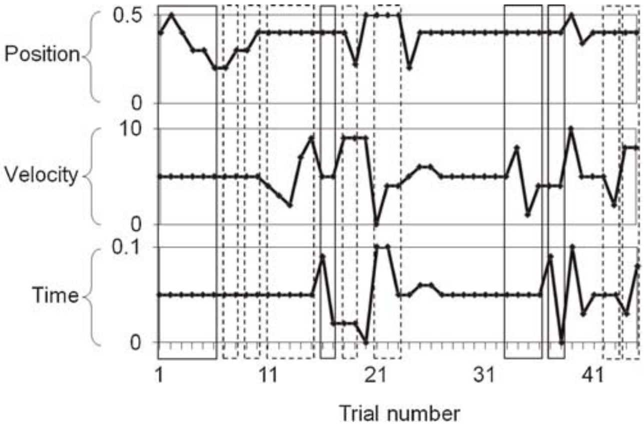


Figure 7. Experimentation sequence illustrating the approach of an exhaustive experimenter.

Discussion

Our findings support the idea that students' propensity to vary one variable at a time is not sufficient to enable them to make sense of a complex problem. Investigations can break down at many points. Linking investigation questions to trials, interpreting results, and determining how variables contribute to the outcome are each potentially problematic. Our analysis suggests that students may use well-established procedures but fail to interpret their findings. Learning from complex experimentation activities often requires planning investigations, recognizing when the investigation is uninformative, revising the investigation, and making sense of the evidence.

Categories of Experimentation

Our results extend and refine previous efforts to categorize experimentation strategies. To capture the aspects of experimentation that are important for understanding the *Airbags* investigation we created categories that represent the linking of procedures such as VOTAT with understanding of the science discipline. Our categories combine the processes used and the goals of the experimentation.

Zimmerman's (2007) review of studies on scientific thinking differentiated categories for the conduct of experiments and for the goals of experimentation. Zimmerman identified two primary ways to characterize how students conduct experiments. Some studies (e.g., Tschirgi, 1980; Vollmeyer et al., 1996) classify students' experiments according to whether they control variables (by varying exactly one variable at a time) or confound them (by varying more than one variable at a time). Other studies (e.g., Klahr & Dunbar, 1988; Schauble, 1996) use ‘‘comprehensiveness’’ (expressed as the fraction of the total experimentation space that learners explore) as a measure of experimentation. We explored controlling variables (as measured by the VOTAT score) and found it was not strongly related to learning about *Airbags*. We also explored comprehensiveness (similar to our variability score) and found that it was also not associated with learning.

Our results show that for a complex task like *Airbags* controlling variables and comprehensiveness do not fully capture important features of experimentation such as planning, attention to outcomes, and alignment with investigation questions. As our exhaustive and unsystematic experimenters demonstrate, mechanically varying one variable at a time or comprehensively exploring all the variable values do not necessarily yield useful insights. Our results show that neither the variability of students' experimentation nor the propensity for students to vary one variable at a time predicts learning. In *Airbags*, where students must explain the mechanisms behind the relationships they observe, they need experimentation practices that reveal these mechanisms. Procedures such as controlling one variable at a time or comprehensive exploration can actually occlude important relationships if students do not consider why they achieved their observed outcomes.

Zimmerman (2007) also summarizes two ways that studies characterize learners' experimentation goals. One method classifies learners as either ‘‘theorists’’ or ‘‘experimenters’’ (e.g., Klahr & Dunbar, 1988; Schauble, 1990). Theorists conduct experiments with the goal of testing established predictions, while experimenters conduct experiments with the goal of generating explanations based on the data. Experimenters resemble students whom we categorized as exhaustive in *Airbags* because they explore the possibilities—if they search for variable relationships, they do so post hoc. In *Airbags* these students could not fully explain the situation because they were not able to sort out the data they generated. Theorists resemble students we classified as intentional in the sense that they use trial comparisons to explore particular inquiry questions, but they do not necessarily need to make specific predictions about the outcomes in order to achieve their insights.

Another method classifies learners as ‘‘scientists’’ or ‘‘engineers’’ (e.g., Schauble, Klopfer, & Raghavan, 1991; Vollmeyer et al., 1996). These studies present learners with the goal of either identifying variable relationships (science context) or optimizing variables to achieve a specific outcome (engineering context). In *Airbags* successful experimenters need to integrate the engineering goal of achieving either a safe or unsafe outcome with the scientific goal of explaining the mechanism that leads to that outcome. Engineering goals may help students identify critical values (such as thresholds in *Airbags*) or to rapidly prune the experimentation space (Klahr et al., 1993). Learners who are successful may combine these outcome-oriented approaches with more scientific approaches to reveal underlying mechanisms that describe a complex system. The coherence score reflects the benefits of both approaches by measuring the conceptual alignment among goals, variable choices, and trial outcomes.

The lack of alignment of category systems like engineers and scientists with the findings from complex situations like *Airbags* stems from differences in the experimental tasks. Learners need to adjust or make ‘‘midcourse corrections’’ to their experimentation approaches in light of new evidence in these tasks (Dunbar, 1993). Students repeatedly revise their hypotheses and investigation goals when exploring a complex task. Whether they begin the investigation with specific hypotheses matters less than whether their methods evolve as they gather evidence. Thus, for *Airbags* scientific and engineering goals are complementary investigation methods rather than alternatives. The intentional experimenters can be described as combining these two approaches. Unsystematic and exhaustive approaches are not consistent enough to be captured by the distinction between scientists and engineers or theorists and experimenters.

Guidance Needs

These results illustrate the need to guide students investigating *Airbags* toward more informative and intentional experimentation. It is not sufficient to emphasize controlling variables alone. Students who can control variables in isolation may have difficulty controlling variables in complex settings as shown for unsystematic experimenters. Students who can control variables may also have difficulty keeping track of their results and combining them as shown for exhaustive experimenters. Intentional experimenters were able to conduct appropriate controlled comparisons, interpret outcomes, combine their findings, and reach valid conclusions, but they constituted only one quarter of our student groups in high school physics. The small number of students who conducted sophisticated experiments in *Airbags* likely reflects the dearth of opportunities students have to explore complex problems during their regular classroom instruction and the tendency for students to focus on following directions rather than engaging in authentic inquiry activities (Waight & Abd-El-Khalick, 2011).

A related study underscores the need for instruction to highlight conceptual, rather than logistical, aspects of experimentation. Reid, Zhang, and Chen (2003) compared *interpretive* and *experimental* support for computer-based experimentation activities about buoyancy. Interpretive support was designed to activate scientific concepts (such as balanced forces) students needed to make valid inferences about the system. Experimental support was designed to help students conduct valid experiments systematically. Reid et al. found a significant effect for interpretive support but no effect for experimental support. To help students make connections between new and prior knowledge about weight, balanced forces, and buoyancy students needed interpretive support.

The results for *Airbags* and the findings of Reid et al. show that students need guidance to combine their experimental procedures with conceptual understanding of a complex topic. This is especially challenging for a topic like *Airbags* where it is possible to control variables but not learn whether the condition under investigation leads to a safe or unsafe outcome. Such guidance could draw from research that suggests ways to make learning more intentional rather than incidental (Bereiter & Scardamalia, 1989; Linn et al., 2004). Other related studies also illustrate the difficulties students have investigating complex situations and highlight the need to focus on factors beyond experimentation (Krajcik, Blumenfeld, Marx, & Soloway, 1994; Lehrer, Schauble, & Petrosino, 2001).

Investigating Realistic Situations

These findings raise further questions about student investigations of realistic situations. Even students in high school physics investigating variables that are important in physics have difficulty adjusting experimentation methods to interpret a realistic situation. Realistic experimentation tasks, where outcomes have consequences for everyday life, provide students with opportunities to make connections between experimentation methods and inquiry goals. These tasks increase the resemblance between classroom experimentation and authentic science. In these complex tasks it is not sufficient to apply a general procedure such as controlling variables. Students need to refine their experimentation approaches based on the results of previous trials. Experimentation as part of inquiry investigations encourages students to apply their understanding of the variables to the design and interpretation of their experiments.

These realistic tasks present the authentic practice of science more faithfully than typical experimentation tasks. They also make science inquiry more compelling for diverse populations of students. Students commonly report that they think about their experience with *Airbags* as they climb into a car to get to school. Anecdotes such as this one reinforce the importance of using compelling driving questions in inquiry units (Krajcik, Blumenfeld, Marx, & Soloway, 1999) to improve classroom learning and promote lifelong appreciation of science.

Conclusions and Implications

Results from *Airbags* suggest the importance of designing instruction to promote informative experimentation (National Research Council, 2007). Our findings suggest ways to respond to the need for more effective science laboratory activities. Focusing on complex, realistic tasks allows students to combine their everyday experiences, knowledge of the discipline, and experimental practices. Students in current high school physics classes have considerable difficulty with such investigative contexts as illustrated in this research. Only about one-quarter of our students were able to conduct experiments that systematically investigated the inquiry goals, even though physics students are arguably among the most sophisticated of high school students.

Courses need to provide more practice as well as effective guidance to help students focus on interpreting sequences of activities and collections of evidence rather than on specific procedures. In addition, students need guidance that prepares them to act intentionally and check their understanding using their knowledge of the situation rather than applying procedures but not making sense of the outcomes.

These results also underscore the importance of outcome measures that capture learners' ability to interpret experiments about complex problems. Researchers need ways to assess students' understanding of the scientific concepts and mechanisms that govern the impact of these concepts in realistic settings. In our study, experimentation scores were not as useful as interpretation scores for determining whether students gained insights into the *Airbags* situation.

These results demonstrate that knowing how to control variables is important but not sufficient for exploring the *Airbags* visualization and reaching valid conclusions. In *Airbags*, students must also incorporate conceptual knowledge (about airbags, motion, and graphs) into designing and interpreting their experiments in order to be successful. The relationship between students' conceptual knowledge and their experimentation can be explored and taught by embedding experimentation in inquiry about variables that have complex relationships to each other.

Students' ability to connect virtual experiments to real-life situations appears valuable and merits further study. Students' everyday experiences provide an important source of ideas to build from when investigating complex topics. Instructional designers need ways to ensure that students interpret their findings by integrating them with their experiences. Ultimately students need to use their knowledge from physics classes to interpret everyday questions such as whether airbags are safe for everyone.

Finally, our study illustrates the advantages of using logging technologies to understand students' science reasoning during the course of their classroom instruction. Even detailed observations of selected groups working with *Airbags* would not have provided us with the information available in the logs of student activities. We needed the log data to classify students' experimentation approaches. Logging technologies offer a valuable resource for designing and researching instruction around complex science topics.

We thank Sophia Rabe-Hesketh and three anonymous reviewers for their helpful comments on this manuscript. We are also grateful to the classroom teachers who took part in this study. We acknowledge the support of National Science Foundation Grants REC-0334199 and DRL-0918743 in conducting this research.

References

Agresti, A., & Finlay, B. (1997). Statistical methods for the social sciences (3rd ed.). Upper Saddle River, New Jersey: Prentice Hall.

American Association for the Advancement of Science. (1993). Benchmarks for science literacy: Project 2061. New York: Oxford University Press.

Bereiter, C., & Scardamalia, M. (1989). Intentional learning as a goal of instruction. In L. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 361–392). Hillsdale, NJ: Lawrence Erlbaum Associates.

Buckley, B., Gobert, J., & Horwitz, P. (2006). Using log files to track students' model-based inquiry. In S. Barab, K. Hay, & D. Hickey (Eds.), Making a difference: Proceedings of the Seventh International Conference of the Learning Sciences (pp. 57–63). Mahwah, NJ: Lawrence Erlbaum Associates.

Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). 'An experiment is when you try it and see if it works': A study of grade 7 students' understanding of the construction of scientific knowledge. International Journal of Science Education, 11(5), 514–529.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. Child Development, 70(5), 1098–1120.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.

Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. Science Education, 91(3), 384–397.

Dunbar, K. (1993). Concept discovery in a scientific domain. Cognitive Science: A Multidisciplinary Journal, 17(3), 397–434.

Fishman, B., Marx, R., Blumenfeld, P., Krajcik, J., & Soloway, E. (2004). Creating a framework for research on systemic technology innovations. Journal of the Learning Sciences, 13(1), 43–76.

Fortus, D., Dershimer, R., Krajcik, J., Marx, R., & Mamlok Naaman, R. (2004). Design-based science and student learning. Journal of Research in Science Teaching, 41(10), 1081–1110.

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

Kali, Y. (2006). Collaborative knowledge building using the Design Principles Database. International Journal of Computer-Supported Collaborative Learning, 1(2), 187–201.

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. Journal of Research in Science Teaching, 41(7), 748–769.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science: A Multidisciplinary Journal, 12(1), 1–48.

Klahr, D., Fay, A., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. Cognitive Psychology, 25(1), 111–146.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction. Psychological Science, 15(10), 661–667.

Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. Journal of Research in Science Teaching, 44(1), 183–203.

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

Krajcik, J., Blumenfeld, P., Marx, R., & Soloway, E. (1994). A collaborative model for helping middle grade science teachers learn project-based instruction. The Elementary School Journal, 94(5), 483–497.

Krajcik, J., Blumenfeld, P., Marx, R., & Soloway, E. (1999). Instructional, curricular, and technological supports for inquiry in science classrooms. In J. Minstrell & E. H. van Zee (Eds.), Inquiry into inquiry: Science learning and teaching (pp. 283–315). Washington, DC: AAAS Press.

Krajcik, J., Blumenfeld, P., Marx, R., & Soloway, E. (2000). Instructional, curricular, and technological supports for inquiry in science classrooms. In J. Minstrell & E. van Zee (Eds.), Inquiring into inquiry learning and teaching in science. (pp. 283–315). Washington, DC: American Association for the Advancement of Science.

Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. Advances in Child Development and Behavior, 17, 1–44.

Lehrer, R., Schauble, L., & Petrosino, A. J. (2001). Reconsidering the role of experiment in science education. In K. Crowley, C. D. Schunn & T. Okada (Eds.), Designing for science: Implications from everyday, classroom, and professional settings. (pp. 251–278). Mahwah, NJ: Lawrence Erlbaum Associates.

Leinhardt, G., Zaslavsky, O., & Stein, M. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. Review of Educational Research, 60(1), 1–64.

Linn, M., Clement, C., & Pulos, S. (1983). Is it formal if it's not physics? (The influence of laboratory and naturalistic content on formal reasoning). Journal of Research in Science Teaching, 20(8), 755–770.

Linn, M., Davis, E., & Bell, P. (2004). Internet environments for science education. Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, M., & Eylon, B. (2006). Science education: Integrating views of learning and instruction. In P. A. Alexander & P. H. Winne (Eds.), Handbook of educational psychology (2nd ed., pp. 511–544). Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, M., & Eylon, B. (2011). Science learning and instruction: Taking advantage of technology to promote knowledge integration. New York: Routledge.

Linn, M., & Hsi, S. (2000). Computers, teachers, peers: Science learning partners. Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, M., Lee, H., Tinker, R., Husic, F., & Chiu, J. (2006). Teaching and assessing knowledge integration in science. Science, 313(5790), 1049–1050.

Linn, M., & Songer, N. (1991). Teaching thermodynamics to middle school students: What are appropriate cognitive demands? Journal of Research in Science Teaching, 28(10), 885–918.

Liu, O., Lee, H., Hofstetter, C., & Linn, M. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. Educational Assessment, 13(1), 33–55.

McDermott, L. (1991). Millikan Lecture 1990: What we teach and what is learned—Closing the gap. American Journal of Physics, 59(4), 301–315.

McDermott, L., Rosenquist, M., & Zee, E. (1987). Student difficulties in connecting graphs and physics: Examples from kinematics. American Journal of Physics, 55, 505–513.

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, VA: NCTM.

National Research Council. (2006). America's Lab Report: Investigations in High School Science. Washington, DC: The National Academies Press.

National Research Council. (2007). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: The National Academies Press.

Polman, J. (2000). Designing project-based science: Connecting learners through guided inquiry. New York: Teachers College Press.

Quintana, C., Reiser, B., Davis, E., Krajcik, J., Fretz, E., Duncan, R., . . . Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. Scaffolding: A Special Issue of the Journal of the Learning Sciences, 13(3), 337–386.

Reid, D., Zhang, J., & Chen, Q. (2003). Supporting scientific discovery learning in a simulation environment. Journal of Computer Assisted Learning, 19(1), 9–20.

Reiser, B., Tabak, I., Sandoval, W., Smith, B., Steinmuller, F., & Leone, A. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In S. M. Carver & D. Klahr (Eds.), Cognition and instruction: Twenty-five years of progress (pp. 263–305). Mahwah, NJ: Lawrence Erlbaum Associates.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. Journal of Experimental Child Psychology, 49(1), 31–57.

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

Schauble, L., Klopfer, L., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. Journal of Research in Science Teaching, 28(9), 859–882.

Siegler, R., & Liebert, R. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. Developmental Psychology, 11(3), 401–402.

Tabak, I., & Reiser, B. J. (2008). Software-realized inquiry support for cultivating a disciplinary stance. Pragmatics & Cognition, 16(2), 307–355.

Trowbridge, D., & McDermott, L. (1980). Investigation of student understanding of the concept of velocity in one dimension. American Journal of Physics, 48(12), 1020–1028.

Trowbridge, D., & McDermott, L. (1981). Investigation of student understanding of the concept of acceleration in one dimension. American Journal of Physics, 49(3), 242–253.

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

Varma, K., Husic, F., & Linn, M. (2008). Targeted support for using technology-enhanced science inquiry modules. Journal of Science Education and Technology, 17(4), 341–356.

Vollmeyer, R., Burns, B., & Holyoak, K. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. Cognitive Science: A Multidisciplinary Journal, 20(1), 75–100.

Waight, N., & Abd-El-Khalick, F. (2011). From scientific practice to high school science classrooms: Transfer of scientific technologies and realizations of authentic inquiry. Journal of Research in Science Teaching, 48(1), 37–70.

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

White, B., & Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. Cognition and Instruction, 16(1), 3–118.

Wilson, C. D., Taylor, J. A., Kowalski, S. M., & Carlson, J. (2010). The relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge, reasoning, and argumentation. Journal of Research in Science Teaching, 47(3), 276–301.

Zacharia, Z. C., Olympiou, G., & Papaevripidou, M. (2008). Effects of experimenting with physical and virtual manipulatives on students' conceptual understanding in heat and temperature. Journal of Research in Science Teaching, 45(9), 1021–1035.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. Developmental Review, 27(2), 172–223.