Effects of Terminological Concreteness on Middle-School Students’ Learning of Experimental Design

Stephanie Ann Siler and David Klahr

CITATION
Effects of Terminological Concreteness on Middle-School Students’ Learning of Experimental Design

Stephanie Ann Siler and David Klahr
Carnegie Mellon University

One obstacle to understanding abstract concepts such as the “control of variables” strategy (CVS) is the tendency for learners to focus on surface rather than deep features in instructional materials. However, in tasks such as learning CVS, these same surface features may also support understanding, provided learners realize the underlying task goal. In this study, we explored the effect of surface features in textually described experiments on middle-school students’ understanding of CVS. We investigated whether the amount of surface detail—or surface-level concreteness—of experiments interacts with student tendency to focus on deep or surface features. As predicted, deep focusers showed better posttest performance when given all concrete examples (concrete-only condition) than when subsequent examples became more concrete (abstract-fading condition) or less concrete (concrete-fading condition). Concrete representations helped deep focusers understand the rationale for controlling variables. Although surface focusers who were given only concrete examples showed better understanding on some measures, they generally failed to develop complete explicit understanding of CVS, including its rationale. Consequently, surface focusers showed similarly poor transfer across conditions. Although students generally benefited from concrete representations, surface focusers may need more support to develop sufficiently coherent understandings that facilitate transfer.

**Keywords:** concreteness, transfer, experimental design, control of variables strategy, middle-school students

The design of controlled experiments using the control of variables strategy (CVS) is a core and recurring topic in science education at state and national levels. For example, one objective of the National Science Education Standards (NSES, 1995) is that “students must become familiar with modes of scientific inquiry, rules of evidence, ways of formulating questions.” According to Pennsylvania State Standards for STEM Education (Pennsylvania Department of Education, 2009), students should be able to design controlled experiments by Grade 7. There is no question that a thorough understanding of CVS is essential to doing and understanding experimental science in a wide range of contexts, from consumers assessing reports of drug-efficacy studies.

CVS is a domain-general procedure for satisfying a science goal of finding out whether a particular variable (i.e., the focal variable, F) affects an outcome. The procedure involves contrasting values of the focal variable (e.g., F₁, F₂) while holding values of all other variables constant. If the experimental outcome differs across conditions (i.e., O₁ ≠ O₂), one can conclude F caused the difference.

Despite its seeming simplicity, many students have difficulty learning CVS (e.g., Lorch et al., 2010). Several studies have found that rather than understanding CVS-related tasks as about determining effects of variables, students frequently focus on surface features such as problem-specific variables (Kuhn & Phelps, 1982; Schauble, 1990; Schauble, Klopfner, & Raghavan, 1991; Siler, Klahr, & Matlen, 2013; Tschirgi, 1980). Consequently, students misinterpret the task goal as discussing beliefs about the effects of problem-specific variables or applying such beliefs in designs to “engineer” particular outcomes (Schauble et al., 1991; Siler & Klahr, 2012). For example, when asked to design experiments that would allow them to discover which ramp variables are causal, students often simply assert their beliefs about variable effects, or whether and how particular variables would affect an outcome (e.g., “Steeper ramps make balls roll farther”). Similarly, students may indicate they set up ramps to satisfy engineering goals such as making the balls roll farthest, fastest, or different distances (Siler & Klahr, 2012).

Alternatively, surface features such as problem-specific variables may help students understand the underlying concept (CVS). For example, a variable-specific property such as ramp steepness, which students generally know affects ball speed, may help students understand why they must control it. However, such surface features may only benefit students who understand the underlying goal of the task. More generally, whether surface features support learning may depend on students’ focus during the task.

Stephanie Ann Siler and David Klahr, Department of Psychology, Carnegie Mellon University.

This research was supported in part by a grant from the U.S. Department of Education, Institute of Education Sciences (R305A100404), and in part by a grant from the National Science Foundation (SBE-0836012). We thank Kevin Willows and Cressida Magaro for their invaluable contributions on many aspects of this project. We also thank the participating school and especially the teachers who allowed us to conduct research in their classrooms.

Correspondence concerning this article should be addressed to Stephanie Ann Siler, Psychology Department, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213. E-mail: siler@cmu.edu
The main question of the present study was whether individual differences in initial focus on a task interact with the amount of surface-level detail to affect learning. Specifically, we varied the amount of surface-level detail in textually described experiments given to middle-school students during CVS instruction. To motivate this question, we first discuss learner-related factors likely to affect focus in CVS tasks. Afterward, we discuss the widely studied construct of “concreteness,” or the amount of realistic detail, of instructional materials and mechanisms by which concreteness may affect learning in general and CVS learning in particular. Finally, we discuss our predictions for the present study.

**Learner-Related Factors Affecting Focus in CVS Tasks**

Surface focus and resulting misinterpretations may be triggered by learner-related factors. In text-based tasks, these factors include domain-specific knowledge (Chi, Feltovich, & Glaser, 1981; Stanovich, 2008), reading comprehension (Siler, Klahr, & Matlen, 2013), self-monitoring (West, Toplak & Stanovich, 2008), and general cognitive ability (Evans, Handley, Neillens, & Over, 2010; Newstead et al., 2004; West et al., 2008). The influence of domain knowledge on focus is shown in the seminal work of Chi et al. (1981). Physics novices tended to sort physics problems by their surface features (e.g., as “ramp” or “spring” problems), whereas physics experts sorted based on deep features (underlying physics principles). Similarly, students’ knowledge of experimentation likely affects whether they realize the science-goal nature of the task: Fifth graders’ science-achievement scores predicted correctly evaluating experiments, independently of science achievement (Siler, Klahr, & Matlen, 2013). Factors correlated with reading comprehension predicted students’ focus when evaluating experiments, independently of science achievement (Siler, Klahr, & Matlen, 2013). Factors correlated with reading comprehension such that metacognitive monitoring of understanding (e.g., Kolic-Vehovec & Bajsanski, 2007) may also affect focus. Students may go on “autopilot” until they experience cognitive conflicts that trigger monitoring processes (cf. Kahneman, 2003), as did fifth and sixth graders given an experimentation task, who often only “shifted” from pursuing engineering to science goals when they met unexpected results (Schauble et al., 1991). Further, students with better monitoring skills may be more aware of their confusion and more likely to exert the mental effort needed to resolve conflicts. If students realize the science goal underlying the task but cannot produce a solution satisfying that goal, they may revert to other (e.g., engineering) goals. Thus, as found for general intelligence (Evans et al., 2010; Newstead et al., 2004), problem-solving skills too may affect focus.

**Concreteness of CVS Instruction**

In addition to learner-related factors, students’ focus on surface-level features in CVS-related tasks may be affected by the amount of realistic detail included in—or concreteness of—instructional material and its sequencing (e.g., De Bock et al., 2011; Goldstone & Son, 2005; Kaminski, Sloutsky, & Heckler, 2006, 2008, 2013; Koedinger, Alibali, & Nathan, 2008; McNeil, Uttal, Jarvin, & Sternberg, 2009; Siler & Willows, 2014). Our analysis of several elementary- and middle-school science textbooks revealed that descriptions of controlled experiments varied in the range of detail they included. Consistent with findings that the “vast majority” of science textbooks “proceed from rules to examples” (Bagchi & Wells, 1998; Nathan et al., 2002) (Goldstone & Son, 2005, p. 488), the most common approach (e.g., DiSpezio, 1999; Hsu, 2005; Miller & Levine, 2006; Prentice Hall, 2007) is to first define terms abstractly then reify them in a concrete example. In contrast, some textbooks (e.g., Lawrence Hall of Science, 2000) first present example scenarios then define the concepts (e.g., variables, controlling) by generalizing from the examples; similarly, others (e.g., Foresman, 2010) introduce terms in the context of example experiments before defining them abstractly. Others (e.g., Pottenger & Young, 1992) define terms only in the context of specific experiments and never abstractly. Thus, how much concrete detail to include in instruction and how to best sequence it remain important—yet unresolved—questions in science education.

**Variable Concreteness**

That concreteness of instructional content may vary widely adds to the complexity of addressing these questions. The concreteness of a representation is proportional to how much information it communicates, via (a) more unique features—both surface and deep—and (b) more detail in the individual features. In the computerized instruction used in the present study, students were taught CVS using virtual ramps in which surface was one variable. Figure 1 shows how representations of ramp surface may vary in concreteness. An actual physical surface such as a piece of sandpaper is the most concrete representation, with more features (e.g., texture, size, color) and feature-specific detail than even a realistic...
representation such as a photograph of the sandpaper. In addition, more detail about texture can be conveyed by the sandpaper than a photograph. Visual representations are generally more concrete than even the most concrete textual representations (e.g., the notion that the photo of sandpaper conveys more detail about texture than does real-life sandpaper).

Textual representations, too, may vary in concreteness (those used in instruction of the present study are circled in Figure 1). "Rough surface" conveys information about texture, though not as much detail as "sandpaper." Further, "rough surface" does not possess features other than texture. The most abstract representation, "Variable x," does not convey any information about texture or any other features. However, it may represent many more entities, including surface (cf. Kaminski et al., 2008).

Mechanisms Underlying Effects of Concreteness on Learning and Transfer

We now discuss four mechanisms addressed in the literature on concreteness (summarized in Table 1) through which surface features in instructional materials might affect learning and transfer, both in general and in the context of learning CVS. As discussed first, whether concrete representations support or impair learning is partly determined by whether additional information conveyed by the features is relevant to the underlying concepts, principles, or procedures.

Distract from deep features. Concrete representations may impair learning if they include features that distract learners from relevant aspects of the task by eliciting conceptions that are irrelevant or even contradictory to the task (e.g., Butcher, 2006; DeLoache, 1991; Moreno, Ozogul, & Reisslein, 2011; Moreno, Reisslein, & Ozogul, 2009; McNeil et al., 2009). As previously discussed, surface features in CVS tasks—including the experimental context and problem-specific variables—may contribute to misinterpretations of the task goal. For example, students may think the instructional goal is to discuss their beliefs about the effect of ramp steepness (e.g., that steeper ramps cause balls to roll farther) may help them understand why steepness must be controlled.

Further, explainatory principles such as the rationale for controlling increase conceptual coherence (Murphy & Medin, 1985). The rationale for controlling variables may increase coherence by linking the procedures “contrast the focal variable” and “control all other variables” so that only the focal variable can affect the outcome. Concepts that are more coherent are more likely to be transferred (Patalano, Chin-Parker, & Ross, 2006). Thus, by supporting students’ understanding of the rationale for controlling variables, surface features may increase conceptual coherence and facilitate transfer.

Cause context-specific learning. Surface features present in the training domain may prompt context-specific reasoning that is tied only to the learning domain, suppressing transfer (Goldstone & Son, 2005). For example, Bassok and Holyoak (1989) found that high-school and college students transferred better from algebra to physics than the reverse, which they accounted for as “... students learn that the physical concepts involved in word problems are critical to the applicability of the relevant equations. Accordingly, they ... fail to recognize any direct relation between physics problem-solving procedures and isomorphic problems drawn from nonphysics domains” (p. 165). However, undergraduates were

<table>
<thead>
<tr>
<th>General Mechanism</th>
<th>Evidence of?</th>
<th>Initial learning</th>
<th>Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concrete features distract focus from deep features</td>
<td>Yes (general &amp; CVS)</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>Concrete features support learning:</td>
<td>Yes (general)</td>
<td>Positive (rationale for CVS)</td>
<td>Positive (via more coherent conception)</td>
</tr>
<tr>
<td>Concrete features prompt context-specific learning</td>
<td>Weak or none</td>
<td>(n/a)</td>
<td>Negative</td>
</tr>
<tr>
<td>Concrete features help recognize applicability</td>
<td>(n/a)</td>
<td>(n/a)</td>
<td>Positive</td>
</tr>
</tbody>
</table>

a (Butcher, 2006; DeLoache, 1991; Moreno, Ozogul, & Reisslein, 2011; Moreno, Reisslein, & Ozogul, 2009; McNeil et al., 2009)  
 b (Cheng & Holyoak, 1985, 1989; Goldstone & Sakamoto, 2003; Goldstone & Son, 2005; Siler & Willows, 2014)  

c (Bassok & Holyoak, 1989; Nisbett & Ross, 1989; Gick & Holyoak, 1983)
equally likely to apply context-specific and abstract versions of a
principle in a transfer domain with novel surface features (Gick &
Holyoak, 1983). Thus, when learning CVS, it is unclear whether the
additional features in concrete representations would induce
formation of domain-specific knowledge that inhibits transfer. For
example, it is unclear whether students would come to understand
the rationale for controlling variables only for ramps—not appreciat-
ing the need to avoid confounds in other domains.

Help recognize applicability. If students can recognize the
abstract concepts underlying concrete representations, concrete
representations may facilitate transfer to novel situations by help-
ing students recognize when abstract concepts are applicable
(Goldstone & Wilensky, 2008). That is, concrete representations
may provide cues that help students realize applicable contexts for
CVS and support transfer. Earlier studies have found that super-
ficial similarities between target and transfer contexts facilitate the
application of abstract principles in the transfer context (e.g., Ross,
1989). However, the relevant question of the present study is
whether concrete representations facilitate application of abstract
concepts in contexts that do not share any surface features, but
involve similar scenarios (i.e., experimenting). That is, does con-
creteness, per se, facilitate recognition of CVS’s applicability in
other concrete contexts? To our knowledge, this mechanism has
not been empirically tested. Evidence of this mechanism includes
better transfer from concrete than abstract examples by learners
who show similar CVS understanding during training.

Effects of Mechanisms for Deep-Focus and
Surface-Focus Students

These mechanisms may differentially affect students who focus
on deep versus surface features of a task. Deep focusers, who by
definition are unlikely to be distracted by surface features, may
primarily be supported by what (to them) are relevant features.
However, among surface focusers, there is a distract–support
trade-off, in which surface features may either primarily distract
students and prevent learning or support their understanding of
CVS. This trade-off is not possible only across students but within
individual students. For example, a surface focuser may initially be
distracted by surface features of representations. But if she is able
to later realize the task goal (i.e., adopt a deep focus), the surface
features may then help her understand the need to control vari-
ables.

Evidence that surface features have both distracting and sup-
portive effects is stronger than evidence for causing context-
specific learning or helping students recognize when abstract con-
cepts apply in future contexts. Thus, we expected that the distract
and support mechanisms would primarily determine learning out-
comes for surface focusers and the support mechanism would be
the primary determiner for deep focusers.

Present Study

In this study, middle-school students evaluated three exper-
iments in one of three training conditions. (a) Concrete-only:
All experiments were concrete and involved ramps (see Figure
2a); (b) concrete-fading: The first experiment was concrete, the
second intermediate (see Figure 2b); and the final one, abstract
(see Figure 2c); and (c) abstract-fading: The first experiment
was abstract, the second intermediate, and the final concrete.
All experiments were represented as text in tables. Even in the
most concrete form, we chose only enough variable-related
detail to convey critical information involving potential causal-
ity. This choice may give the concrete representations the best
chance to “win” the distract–support trade-off.

The range of concreteness of the present study is more abstract
than ranges in many studies of concreteness, which often include
pictures (Goldstone & Son, 2005; Kaminski, Sloutsky, & Heckler,
2006, 2008, 2013; Moreno, Ozogul, & Reisslein, 2011; Mason,
Pluchino, Tornatora, & Ariasi, 2013; McNell, Uttal, Jarvis, &
Sternberg, 2009; Reisslein, Moreno, & Ozogul, 2010; Siler &
Willows, 2014). However, like the present study, some have
compared text to text-based symbolic representations (e.g., alge-
bra: Koedinger, Alibali, & Nathan, 2008; logic: Cheng & Holyoak,
1985, 1989; Evans, Handley, & Harper, 2001; general problem
solving: Gick & Holyoak, 1983).

Predictions: Deep focusers. We generally expected that con-
tinuous reflecting on concrete representations in the concrete-only
condition would best support deep focusers’ understanding of the
rationale for controlling variables. This understanding was found

![Figure 2. Concrete (top), intermediate (middle), and abstract (bottom)
representations of experiments given during instruction in the present
study. In each representation, the focal variable is highlighted, and there is
at least one confounded variable. See the online article for the color version
of this figure.](image-url)
to predict transfer (Siler, Klahr, Magaro, Willows, & Mowery, 2010). Thus, we predicted that deep focusers would show better posttest performance in the concrete-only than the abstract-fading condition (Prediction 1a). We expected surface features in the initial experiment (Experiment 1) to support deep focusers’ CVS understanding (Prediction 1b). Because they reflect on experiments with (what to them are) relevant surface features throughout instruction, we also expected deep focusers to show better final CVS understanding (on Experiment 3)—particularly, of the rationale for controlling—in the concrete-only than the abstract-fading condition (Prediction 1c).

Similarly, because concrete-only deep focusers reflect on experiments with relevant surface features throughout instruction, we predicted that they would show better posttest performance than concrete-fading deep focusers (Prediction 2a). We also expected deep focusers in the concrete-only condition to show better understanding of CVS—especially its rationale—by the end of instruction (Prediction 2b). However, because students in the concrete-only condition only evaluate ramp experiments during instruction, their CVS understanding may become tied to ramp-specific features, hindering transfer. If so, the concrete-fading condition, in which the final experiment is completely abstract (devoid of surface features), may best promote transfer. However, as discussed earlier, evidence that surface features hinder transfer this way is weak, and deep focusers by definition are likely to recognize the abstract, deep features of instructional content. In addition, because deep focusers attend to deep features, we did not expect them to need concrete representations to recognize applicable contexts. Thus, we expected deep focusers’ transfer of their final understanding to the posttest to be similar in the concrete-only and concrete-fading conditions (Prediction 2c).

Tests of effects: Surface focusers. Effects of surface features for surface focusers are unclear due to the distract–support trade-off. Thus, we turn to research comparing the effects of abstract representations to those with irrelevant features. Some studies have found advantages of abstract representations (e.g., Butcher, 2006; DeLoache, 1991; Goldstone & Son, 2005; Moreno, Ozogul, & Reisslein, 2011; Moreno, Reisslein, & Ozogul, 2009; McNeil et al., 2009). However, the “abstract” representations included some surface features. For example, in Goldstone and Son (2005), the abstract representation of the complex system simulation included visual components labeled “ants” and “food.” When, as in the present study, purely abstract representations are used (e.g., Cheng & Holyoak, 1985, 1989; Siler & Willows, 2014), concrete versions typically include relevant features. One research area that has compared complete abstract representations to those with irrelevant content is the study of belief bias in deductive reasoning. However, even this research has mixed results. Some studies (Evans, Handley, & Harper, 2001) found better performance on abstract syllogisms (e.g., if A, then B) than syllogisms with believable conclusions. However, others (e.g., Goel & Dolan, 2003) found no differences.

Thus, it was unclear whether surface focusers in the present study would learn more from abstract or concrete representations. Comparing the concrete-only and abstract-fading conditions, we examined whether surface focusers in the abstract-fading condition were more likely to show surface focus (by giving engineering and variable-effects responses; Test 1b). We also examined whether surface features in the initial experiment supported surface focusers’ CVS understanding (Test 1c). We expected the “winner” of the distract–support trade-off to determine whether surface focusers in the concrete-only or abstract-fading condition showed better posttest performance (Test 1a). We also examined whether reflecting on concrete experiments throughout instruction would support surface focusers’ final understanding (Test 1d).

Likewise, we tested for whether surface focusers in the concrete-only or concrete-fading condition showed better posttest performance (Test 2a). We expected that posttest performance would depend on whether reflecting on only concrete representations or fading from concrete to abstract representations led to better final instructional CVS understanding (Test 2b) and how surface features in the final representation would affect transfer (Test 2c) for surface focusers.

Method

Participants

Participants were sixth- and seventh grade-students from a suburban public middle school serving mostly middle-class students (about one quarter were eligible for free or reduced-price lunch); 94% of students were White, 2% African American, 2% mixed race, 1% Asian, and 1% Hispanic. The sample included a similar number of boys and girls. On standardized state assessments of math and science, students at this school were near state and national averages. All students at this school have science class daily.

A total of 261 students (146 sixth graders and 117 seventh graders) completed the pretest. Of these students, 80 (31%) showed above-chance levels of CVS understanding (i.e., set up at least two controlled experiments out of six on the pretest) and were excluded from further analyses. Among the 181 students who did not show above-chance CVS understanding, six (3%) did not complete the study (three only completed the pretest; three completed the instruction but not the posttest). Log-file records indicated that the remaining 175 students completed all steps of the instruction (in their respective condition) and thus received the treatment as intended. However, four students completed the posttest but some posttest responses were missing due to logging errors; thus, their data were excluded from analyses. The final sample included 171 students (106 sixth graders and 65 seventh graders) who did not show above-chance CVS understanding on the pretest and completed all phases of the study. Because 31% of students showed above-

1 The sixth and seventh graders in the present study had not recently taken the science portion of the Pennsylvania System of School Assessment (PSSA; Pennsylvania Department of Education, 2012), given at Grades 4 and 8. However, the proficiency rate of eighth graders at this school who had taken the science PSSA the previous year was 63%, near the state average (60%). Similarly, proficiency rates of students in the present study were similar to state averages for math (Grade 6: 70%; Grade 7: 69%; Pennsylvania averages: 73% and 77%, respectively) and reading (Grade 6: 70%; Grade 7: 76%; Pennsylvania averages: 65% and 69%, respectively). Because Pennsylvania state proficiencies are similar to national averages for fourth and eighth graders in science, math, and reading (NCES, 2011), the sixth and seventh graders who took the pretest in the present study are likely near national averages in science, reading, and math knowledge and skills.
Materials and Procedure

This study occurred during students’ regular 40-min science classes. Each of the students worked individually at a computer in a school computer lab. Each study phase—the pretest, instruction, and posttest—was completed during one class period on three consecutive days. In each phase, student responses were saved in log files and later analyzed. Students first completed one of two isomorphic versions of a pretest, consisting of six “story problems”; the other version was given as the posttest. In each of three domains (e.g., building rockets), students first designed an experiment then evaluated (and, if necessary, fixed) a given experiment. Students explained each design and evaluation response. Figure 3 shows a design item (top) and evaluate item (bottom).

Students were randomly assigned to instructional conditions within each class. The computer-delivered instruction followed a Socratic questioning method. During instruction, students evaluated three confounded experiments in their respective condition: (a) concrete-only condition: All experiments (Experiment 1–3) were concrete; (b) concrete-fading condition: Experiment 1 was concrete, Experiment 2 intermediate, and Experiment 3 abstract; (c) abstract-fading condition: Experiment 1 was abstract, Experiment 2 intermediate, and Experiment 3 concrete. (Figure 2 shows all three representations.)
Concrete experiments had concrete variables and values (e.g., the ramp surface could be smooth or rough). All concrete variables and values involved ramps.

Intermediate experiments had concrete variables, but specific values for variables were not given. Rather, variables were described as having “the same” or “different” values across ramps.

For abstract experiments, neither variables nor their values corresponded to concrete entities. Variables were labeled "Variable A/B/C," and values were described only as the same or different across conditions.

All experiments were described textually without pictures. For each example experiment (Experiments 1–3), designed by a hypothetical student, “Amal,” students answered a series of multiple-choice (MC) and open-ended (OE) questions, described below.

Q1-MC: “Do you think this [experiment] is a good way for Amal to find out whether the [focal variable, i.e., Variable x or ramp surface/type of ball/starting position] makes a difference?” After students responded, they received feedback (e.g., “That’s right. It’s not a good way to find out if the [focal variable] makes a difference”).

Q1-OE: “Why is it (actually) a bad way?” Students then received feedback on why it was a bad way (i.e., “Only the [focal variable] should be different.”). Students were asked to imagine that Amal ran the experiment and found the that the balls rolled different distances.

Q2-MC: “Could you know for sure that the [focal variable] caused this difference?” Again, students received feedback on their response (e.g., “That’s right. You could not know for sure that the [focal variable] caused the difference in [outcome]”).

Q2-OE: “Why could you [actually] not know for sure?” No feedback was given.

Q3-MC: “Could the way he has set each of these other variables cause a [different outcome across conditions]?” (See Figure 4.) Students indicated whether each of the three nonfocal variables could cause a hypothetical difference in outcomes. Because Q3-MCs directly assess students’ understanding of the link between the experimental setup and determinacy of the outcome’s cause, they most directly assess the rationale for controlling variables. For each response, students received corrective feedback (i.e., right/wrong) and explanatory feedback (e.g., “The [variable] is different, so it could cause a difference in [outcome]”).

Students corrected the setup by clicking in the table and changing values.

Q3-OE: “Explain to Amal why this is now a good way to find out whether or not the [focal variable] makes a difference.” Students received explanatory feedback (e.g., “Everything is the same except for the [focal variable] . . . If there is a difference in [outcome], it has to be because of the [focal variable] . . .”).

The second and third experiments were introduced with: “Now, we’ll look at Amal’s next experiment” (concrete-only condition), and “Now, we’ll [remove/add] some details to describe Amal’s next experiment” (concrete-fading/abstract-fading condition). The next day, students completed the posttest (isomorphic to the pretest), which assessed students’ transfer of their learning to novel domains.

Scoring. Students earned one point for each experiment they correctly designed (out of three) and one point for each experiment they evaluated as “bad” and converted to a controlled experiment (out of three). The maximum possible score on the pretest and posttest was six.

 Codings

Focus level. We categorized students as likely to focus on surface versus deep features initially in instruction based on their responses to the last pretest item (see Figure 3, bottom) in part because students sometimes show learning during the pretest; thus, their final response may be most indicative of their initial instructional understanding. Second, recency effects for delayed recall (e.g., Howard & Kahana, 1999; Tzeng, 1973) suggest that students are most likely to remember the last pretest item by the start of
instruction. Third, both the last pretest item and instruction involved evaluating experiments.

Students’ responses fell into four categories: (a) engineering responses in which students indicated that the setup was “good” (or “bad”) because it would (or would not) result in a desired outcome (e.g., the noncontrastive setup was “good” because “pictures are easier to remember” or “you have quiet and something healthy to drink plus more resources [at the library]”), (b) nonresponses, or inability to explain evaluation responses (e.g., “I guessed” or “I don’t know”), (c) contrastive responses that generally indicated that something (one or more variables) should be contrasted (e.g., “that’s a bad way to find out because there are no differences”) and that specifically the focal variable should be contrasted (e.g., “They are using the same person”), and (d) “Science-goal-only” responses indicating that the goal was to find out about something, but without explicitly expressing the need to contrast. Students who expressed engineering goals (50% overall) or were unable to explain their responses (20%) were classified as “surface focus” (SF). Students who gave contrastive responses (29%) or science-goal-only responses (1%) were categorized as “deep focus” (DF).

Reliability and validity. A second coder, blind to condition, coded a randomly selected 10% of responses as one of the four categories. Interrater reliability was high (κ = 1). Classification as deep versus surface focus based on last pretest responses was highly related to the total number of deep-focus responses students expressed on the pretest, F(1,169) = 203.30, p < .001, ν2 = .55 (MDF = 3.05, SD = 1.52; MSD = 0.44, SD = 0.85). Further, results using last pretest responses students gave as the basis for categorization and results using the total number of deep-focus responses students gave on the pretest were analogous, including the same significance patterns for the main effect and interactions with condition (discussed in the Results section).

Deep focusers. Overall, 51 students (30%) were classified as deep focusers. Of these students, 49 gave contrastive and two gave science-goal-only responses. Among contrastive responders, (a) 31% explicitly expressed a science goal for contrasting (e.g., “It is bad because it is the same thing and you wouldn’t really get to test anything”), (b) 29% gave contrastive responses without indicating a science goal, but had previously expressed a science goal on the pretest, and (c) 41% did not previously show science-goal understanding. These contrastive subgroups (a–c) did not differ on pretest (Ma = 0.40, SD = 0.51; Mb = 0.43, SD = 0.51; Mc = 0.55, SD = 0.51, p = .65) or posttest (Ma = 3.07, SD = 2.76; Mb = 3.71, SD = 2.20; Mc = 3.25, SD = 2.38, p = .77).

Surface focusers. Most students (70%) were classified as surface focusers. The decision to include nonresponders with engineering responders as surface focusers was partly based on students’ earlier pretest responses. Most of the nonresponders (69%) had previously expressed engineering—but no science—goals on the pretest. Of the remaining nonresponders, 20% indicated that they guessed or did not know throughout the pretest; the remaining 11% previously gave a single science-only or contrastive response (along with engineering or “don’t know” responses). In addition, there were no pre- or posttest differences between engineering and nonresponders, indicating between-groups similarity. Further, when we excluded nonresponders from the ANCOVA for the primary analyses, we found the same patterns of results (i.e., the condition by pretest and condition by focus interactions were both still significant and in the same directions as when nonresponders were included).

Surface focusers versus deep focusers. Surface focusers scored significantly lower than deep focusers on both the pretest (MDF = 1.85, SD = 2.23; MDF = 3.29, SD = 2.43, p < .001). Thus, overall, there were no within-group focus differences for either surface or deep focusers, but there were between-groups (SF vs. DF) differences in pre- and posttest performance. This suggests within-focus group homogeneity and gives validity to the category groupings used.

Student explanations given during instruction. To assess students’ explicit CVS understanding during instruction, we also coded their responses to Q2-OE, “Why could you [actually] not know for sure [that the focal variable caused a difference in outcomes]?” as one of the following:

Rationale–complete. Responses explicitly expressing the rationale for controlling nontested variables by explicitly linking the variable setups (i.e., as same or different across conditions) to the experimental outcome (e.g., “Everything else is different too, not just the surface so anything else could cause it;” “The surface, the steepness, and the starting position all could have affected the distance the ball rolled because they were all different.”)

Rationale–vague/incomplete. Responses indicating that variables other than the focal variable may affect the outcome but did not refer to variable setups (e.g., “it could be the surface or the steepness”); these responses may also have failed to reference specific alternative causal variables (e.g., “Something else can change the results”).

Procedural. Responses expressing procedural knowledge of controlling variables (e.g., “[you could not know for sure] because Variables A and C need to be the same instead of different because the steepness is different.”)

Responses not clearly indicating procedural or rationale understanding:

Engineering/variable effect. Responses expressing outcome goals or beliefs about specific variable effects (e.g., “The balls don’t roll too good on rough ramps;” “If it’s rough then it would be hard to roll, but if it’s smooth it rolls a lot easier”).

Do not know. Inability to explain (e.g., “I don’t know” “I guessed”).

Vague. Unclear responses (e.g., “the variables need to be changed” or “it doesn’t talk about the surface”).

All different. Responses that all variables should be contrasted (e.g., “[It’s bad because] both balls are heavy”).

Too abstract. Responses that one could not tell if the focal variable caused the difference in outcome because information about the setup was missing (e.g., “[There is] not enough info;” “It doesn’t give you what each setup really is;” “You wouldn’t know how steep the ramp was, what type of ball you used, and the starting position”).

Other. Responses indicating the effect of the focal variable is unknown because the experiment was not run (e.g., “We couldn’t know for sure because we didn’t test it out”), incomplete responses (e.g., “steepness an”), and random entries (e.g., “ti ujnv6ot3hin3”).

Reliability. A second coder, blind to condition, coded a randomly selected 10% of responses for each question. Agree-
ment in categorizing responses to one of the nine categories was high ($\kappa = .80$).

**Results**

**Preliminary Analyses**

Including only students who did not show above-chance CVS understanding on the pretest and were in final analyses, conditions (concrete-only, concrete-fading, abstract-fading) were similar in pretest score ($M_{CO} = 0.22, SD = 0.42$; $M_{CF} = 0.29, SD = 0.46$; $M_{AF} = 0.21, SD = 0.41$, $p = .56$) and focused on the last pretest item (32%, 31%, and 27% of students were classified as deep focusers, respectively, $p = .83$). Sixth and seventh graders did not differ on pretest ($M_{GR6} = 0.21, SD = 0.41$; $M_{GR7} = 0.30$, $SD = 0.46$, $p = .18$). However, consistent with younger students showing more goal misconceptions (Siler, Klahr, & Matlen, 2013), sixth graders were more likely to be classified as surface focusers than seventh graders (79% vs. 56%, respectively), $\chi^2(1, N = 171) = 9.48$, $p = .002$.

**Interactions With Condition**

We predicted that deep focusers in the concrete-only condition would show the best posttest performance (Predictions 1a and 2a) but we made no strong predictions for surface focusers. Grade and focus were included with pretest and condition in ANCOVA. All main effects and two-way interaction terms were included in the model (we did not predict—or find—any 3- or 4-way interactions). Condition interacted independently with focus, $F(2, 161) = 5.29, p = .006$, $\eta_p^2 = .06$, and pretest, $F(2, 161) = 4.42, p = .01$, $\eta_p^2 = .05$. This result did not appear because of how we determined focus: When focus based on the last pretest item was replaced with the total number of deep-focus pretest responses in ANCOVA, the condition by pretest and condition by total deep-focus response interactions were again both significant.

**Condition $\times$ Focus interaction.** The Condition $\times$ Focus interaction is shown in Figure 5. For deep focusers, the effect of condition was significant, $F(2, 47) = 5.33, p = .008$, $\eta_p^2 = .19$ (means are in Table 2). Consistent with Predictions 1a and 2a, concrete-only deep focusers had higher posttest scores than deep focusers in the abstract-fading ($p = .004$) and concrete-fading ($p = .02$) conditions. Concrete-fading and abstract-fading conditions were similar ($p = .54$). However, the effect of condition was not significant for surface focusers, $F(2, 115) = 0.66, p = .52$, $\eta_p^2 = .01$ (Tests 1a and 2a were $ns$).

**Condition $\times$ Pretest interaction.** To investigate the unexpected Condition $\times$ Pretest interaction, we examined the effect of Condition $\times$ Pretest score (see Table 2 for means). Of students who designed one controlled experiment on the pretest, there was no effect of condition, $F(2, 36) = 0.31, p = .74$, $\eta_p^2 = .02$. However, for students who designed no controlled experiments, there was a significant effect of condition, $F(2, 126) = 4.84, p = .009, \eta_p^2 = .07$. Post hoc comparisons with Bonferroni correction revealed higher posttest scores in the concrete-only than concrete-fading condition ($p = .01$) and marginally higher scores in the concrete-only than abstract-fading condition ($p = .06$). Concrete-fading and abstract-fading scores were similar ($p = 1$). Because the concrete-only condition was only better for deep focusers (who predominately understood the need to contrast variables) and for students who set up no controlled experiments on the pretest, the concrete-only condition appeared to support learning to control variables in particular.

**Mechanism-Specific Predictions/Tests**

We now investigate whether benefits of the concrete-only condition for deep focusers were from the mechanisms proposed. Then we discuss tests of learning mechanisms for surface focusers. The specific predictions and tests are listed in Table 3 in the order discussed in the results.

**Analyses.** We assessed students’ initial (Experiment 1) and final (Experiment 3) instructional understanding from their responses to these questions:

(a) Q2-OE: “Why would you not know for sure [that the focal variable caused the difference]?” (b) MC questions: Q1-MC: “Do you think this [experiment] is a good way for Amal to find out whether the [focal variable] makes a difference?” Q2-MC: “Could you know for sure that the [focal variable] caused this difference?” Q3-MC: “Could the way he has set each of these other variables cause a [different outcome across conditions]?” Students responded for each of three variables.

For MC responses, we ran ANCOVA with condition as the independent variable, grade and focus as fixed variables, and
pretest covaried. To assess differential transfer of students’ final instructional understanding to the posttest, we included the number of correct MC responses in Experiment 3 in ANCOVA. Evidence of differential transfer included different relationships between MC performance and posttest per condition (i.e., Condition \times MC interactions), or, if this interaction was not significant, significant effects of condition when MC score was covaried.

**Deep Focusers**

**Concrete only versus abstract fading.** As reported, consistent with Prediction 1a, deep focusers showed better posttest performance in the concrete-only than abstract-fading condition. We now examine mechanisms underlying this outcome.

**Initial (Experiment 1) understanding.** To assess whether surface features supported deep focusers’ initial understanding (Prediction 1b), we first examined their Q2-OE responses for Experiment 1. Deep focusers were marginally more likely to give correct (procedural or rationale) responses in the concrete-only than abstract-fading condition (52% vs. 20%, respectively), $\chi^2(1, n = 34) = 3.78, p = .05$

For all students, there was a significant effect of condition for Experiment-1 MC responses, $F(2, 167) = 4.92, p = .008, \eta_p^2 = .06$. Abstract-fading students gave fewer correct responses than concrete-only ($p = .007$) and concrete-fading ($p = .007$) students. As expected, concrete-only and concrete-fading conditions were similar ($p = .97$). For just deep focusers, there was a main effect of condition, $F(2, 46) = 3.35, p = .04, \eta_p^2 = .13$; concrete-only deep focusers gave more correct MC responses than abstract-fading deep focusers ($p = .02$; see Table 4 for means). This was a result of concrete-only deep focusers’ more accurate Q3-MC responses, $F(1, 32) = 6.40, p = .02, \eta_p^2 = .17$.

### Table 3
**Specific Predictions for Deep-Focus Students and Tests for Surface-Focus Students (Actual Outcomes)**

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Deep focus: Specific predictions for mechanism</th>
<th>Surface focus: Specific tests of mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction 1a: Concrete-only (CO) &gt; abstract-fading (AF) on posttest.</td>
<td>Test 1a: Concrete-only vs. abstract-fading on posttest. ($n.s.$)</td>
<td></td>
</tr>
<tr>
<td>(√ - results supported prediction)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction 1b: Surface features support initial CVS understanding</td>
<td>Test 1b: Surface-focus responses for initial experiment</td>
<td></td>
</tr>
<tr>
<td>• More correct Q2-OE in CO than AF ($m.s.$)*</td>
<td>• Surface focus on Q2-OE (CO &gt; AF)</td>
<td></td>
</tr>
<tr>
<td>• More correct MC in CO ($\checkmark$ due to Q3-MC)</td>
<td>• Q2-OE responses ($n.s.$)</td>
<td></td>
</tr>
<tr>
<td>Prediction 1c: Surface features support final CVS understanding (esp. rationale)</td>
<td>Test 1c: Compare initial CVS understanding;</td>
<td></td>
</tr>
<tr>
<td>• More Q2-OE rationale in CO than AF ($\checkmark$)</td>
<td>• MC responses ($m.s.$; CO &gt; AF; due to Q3-MC)</td>
<td></td>
</tr>
<tr>
<td>• More correct MC responses in CO ($\checkmark$)</td>
<td>Test 1d: Compare final CVS understanding (esp. rationale)</td>
<td></td>
</tr>
<tr>
<td>Prediction 2a: Concrete-only (CO) &gt; Concrete-fading (CF) on posttest. ($\checkmark$)</td>
<td>Test 2a: Concrete-only (CO) vs. Concrete-fading (CF) on posttest. ($n.s.$)</td>
<td></td>
</tr>
<tr>
<td>Prediction 2b: Surface features support final understanding (esp. rationale)</td>
<td>Test 2b: Final CVS understanding (esp. rationale)</td>
<td></td>
</tr>
<tr>
<td>• More Q2-OE rationale in CO than CF ($m.s.$)</td>
<td>• Q2-OE rationale (CO &gt; CF)</td>
<td></td>
</tr>
<tr>
<td>• More correct MC (esp. Q3-MC) in CO ($n.s.$)</td>
<td>• MC responses (esp. Q3-MC; $n.s.$)</td>
<td></td>
</tr>
<tr>
<td>Prediction 2c: Exp3 to post: No difference in transfer</td>
<td>Test 2c: Compare transfer from final experiment to posttest</td>
<td></td>
</tr>
<tr>
<td>• No condition by MC response interaction. ($m.s.$ interaction; for higher-MC students: CO &gt; CF)</td>
<td>• Test for condition by MC interaction ($n.s.$)</td>
<td></td>
</tr>
<tr>
<td>• No effect of condition, controlling for MC ($n/a$)</td>
<td>• Test for effect of condition co-varying MC ($n.s.$)</td>
<td></td>
</tr>
</tbody>
</table>

* There was a marginally significant ($m.s.$) effect ($p < .10$) for this outcome. $n.s.$ = not significant.

---

### Table 4
**Mean Instructional Multiple-Choice Score by Focus, Condition, Example Experiment (Exp) Given in Training, and Its Associated Question (Q)**

<table>
<thead>
<tr>
<th>Experiment/Q</th>
<th>Concrete-only</th>
<th>Abstract-fading</th>
<th>Concrete-fading</th>
<th>Concrete-only</th>
<th>Abstract-fading</th>
<th>Concrete-fading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp1: All MC Qs</td>
<td>4.26 (.81)**</td>
<td>3.33 (1.68)</td>
<td>4.18 (.81)**</td>
<td>3.34 (1.33)**</td>
<td>2.83 (1.46)</td>
<td>3.39 (1.03)**</td>
</tr>
<tr>
<td>Q1&amp;2-MC</td>
<td>1.37 (.76)</td>
<td>1.27 (.59)</td>
<td>1.24 (.75)</td>
<td>1.93 (.75)**</td>
<td>1.20 (.68)</td>
<td>.89 (.73)**</td>
</tr>
<tr>
<td>Q3-MC</td>
<td>2.89 (.32)***</td>
<td>2.07 (1.39)</td>
<td>2.94 (.24)***</td>
<td>2.41 (.92)***</td>
<td>1.63 (1.36)</td>
<td>2.50 (.73)***</td>
</tr>
<tr>
<td>Exp2: All MC Qs</td>
<td>4.74 (.56)**</td>
<td>4.00 (1.25)</td>
<td>4.53 (.72)</td>
<td>4.32 (.88)</td>
<td>4.15 (.82)</td>
<td>4.13 (.96)</td>
</tr>
<tr>
<td>Q1&amp;2-MC</td>
<td>1.79 (.42)</td>
<td>1.33 (.82)</td>
<td>1.59 (.71)</td>
<td>1.59 (.67)</td>
<td>1.44 (.67)</td>
<td>1.58 (.60)</td>
</tr>
<tr>
<td>Q3-MC</td>
<td>2.95 (.23)</td>
<td>2.67 (.62)</td>
<td>2.94 (.24)</td>
<td>2.73 (.50)</td>
<td>2.71 (.56)</td>
<td>2.55 (.80)</td>
</tr>
<tr>
<td>Exp3: All MC Qs</td>
<td>4.84 (.37)***</td>
<td>3.53 (.92)</td>
<td>4.41 (1.06)***</td>
<td>4.56 (.74)</td>
<td>4.17 (1.00)</td>
<td>4.16 (1.20)</td>
</tr>
<tr>
<td>Q1&amp;2-MC</td>
<td>1.95 (.23)***</td>
<td>1.00 (.76)</td>
<td>1.94 (.24)***</td>
<td>1.80 (.46)</td>
<td>1.66 (.48)</td>
<td>1.76 (.43)</td>
</tr>
<tr>
<td>Q3-MC</td>
<td>2.89 (.32)</td>
<td>2.53 (.64)</td>
<td>2.47 (1.01)</td>
<td>2.76 (.58)</td>
<td>2.51 (.81)</td>
<td>2.39 (1.17)</td>
</tr>
</tbody>
</table>

**Note.** Mean differs from Abstract-fading mean at:

* $p < .10$.  ** $p < .05$.  *** $p < .005$.  **** $p < .001$. 
Final (Experiment 3) understanding. To assess whether surface features supported deep focusers’ final understanding (Prediction 1c), we first examined responses to Q2-OE for Experiment 3. Deep focusers were more likely to express the rationale for controlling variables in the concrete-only than abstract-fading condition (58% vs. 7%, respectively), \( \chi^2(1, n = 34) = 9.63, p = .002 \).

For all students, there was a significant Condition \( \times \) Focus interaction for MC responses, \( F(2, 147) = 3.10, p = .048, n^2 = .04 \). For deep focusers, there was a main effect of condition, \( F(2, 48) = 10.83, p < .001, n^2 = .31 \). Deep focusers gave more correct MC responses in the concrete-only than abstract-fading condition (\( p < .001 \)). This was due to concrete-only deep focusers’ better performance on: (a) Q1- and Q2-MC, \( F(1, 32) = 26.91, p < .001, n^2 = .46 \), and (b) Q3-MC, \( F(1, 32) = 4.66, p = .04, n^2 = .13 \). Thus, concrete-only deep focusers showed better procedural and conceptual understanding by the end of instruction. When Experiment-3 MC score was covaried, the effect of condition on posttest was no longer significant, \( F(1, 30) = 0.37, p = .55, n^2 = .01 \). Thus, concrete-only deep focusers’ better posttest performance can be explained by their better final understanding.

Concrete only versus concrete fading. As reported, consistent with Prediction 2a, deep focusers showed better posttest performance in the concrete-only than concrete-fading condition. Because students in these conditions received the same instruction for the initial experiment (Experiment 1), we did not predict—or find—any differences in initial understanding.

Final (Experiment 3) understanding. We predicted that deep focusers would develop better final CVS understanding in the concrete-only condition (Prediction 2b). Deep focusers were marginally more likely to express the rationale for controlling variables on Experiment 3 in the concrete-only than concrete-fading condition (58% and 29%, respectively), \( \chi^2(1, n = 36) = 2.95, p = .09 \). As reported, there was a main effect of condition among deep focusers for number of correct MC responses. However, deep focusers in the concrete-only and concrete-fading conditions did not differ (\( p = .12 \)).

Transfer from Experiment 3 to posttest. We did not expect the concreteness of the final experiment to affect deep focusers’ transfer to the posttest (Prediction 2c). To assess differences in transfer, we included students’ MC scores for Experiment 3 in ANCOVA. There was a marginally significant Condition \( \times \) MC interaction, \( F(2, 44) = 23.77, p = .07, n^2 = .11 \). There was a main effect of condition for deep focusers who scored above the mean (4.64), \( F(2, 27) = 3.69, p = .04, n^2 = .22 \). Post hoc comparisons with Bonferroni correction revealed that posttest scores of deep focusers were higher in the concrete-only than concrete-fading condition (\( M_{CO} = 5.00, SD = 1.93; M_{CF} = 2.75, SD = 2.34 \), \( p = .04 \). There were no other significant pairwise differences. For deep focusers who scored below the mean, there was no effect of condition, \( F(2, 17) = 0.58, p = .57, n^2 = .06 (M_{CO} = 2.33; SD = 1.16; M_{CF} = 3.00, SD = 2.55) \). Thus, by the end of instruction, deep focusers in the concrete-only condition only showed (marginally) better understanding in their Q2-OE responses; however, they showed better transfer of their final understanding to the posttest.

Although we did not expect deep focusers to need surface features to realize CVS’s applicability, this result is consistent with this mechanism. If this mechanism were the cause of the better transfer of concrete-only deep focusers, deep focusers in the concrete-only condition should be more likely to recognize the applicability of CVS on the posttest. However, this was not the case. On the first posttest question, deep focusers were similarly likely to express the CVS procedure or rationale in the concrete-only and concrete-fading conditions (58% and 41%, respectively), \( \chi^2(1, n = 36) = 1.00, p = .32 \). There was also no difference between the concrete-only and concrete-fading conditions in likelihood of expressing deep-focus (as previously defined; 74% and 82%, respectively), \( \chi^2(1, n = 36) = 0.39, p = .53 \).

Surface Focusers

Concrete only versus abstract fading. Although surface focusers did not differ in posttest performance, we discuss evidence vis-à-vis the distract–support trade-off. We expected this trade-off to determine whether surface focusers would develop better initial CVS understanding from the concrete or abstract experiment.

Initial (Experiment 1) understanding. Consistent with (removing) distracting effects, abstract-fading surface focusers were less likely to give engineering or variable-effect responses to Q2-OE than concrete-only surface focusers (2% vs. 20%, respectively), \( \chi^2(1, N = 82) = 6.12, p = .01 \) (Test 1b). However, instead of prompting deep focus, the lack of detail in the abstract experiment seemed to confuse surface focusers. Surface focusers in the abstract-fading condition were more likely than concrete-only surface focusers to indicate they could not tell whether the focal variable caused the difference because information was missing (22% vs. 2%, respectively), \( \chi^2(1, n = 82) = 7.29, p = .007 \). Ultimately, surface focusers in the concrete-only and abstract-fading conditions were similarly likely to give correct (rational or procedural) responses (29% vs. 17%, respectively), \( \chi^2(1, n = 82) = 1.71, p = .19 \) (Test 1c).

As previously discussed, students gave more correct Experiment-1 MC responses in the concrete-only than abstract-fading condition. Among just surface focusers, there was a marginal effect of condition, \( F(2, 116) = 2.35, p < .10, n^2 = .04 \). Surface focusers gave more correct responses in the concrete-only than abstract-fading condition (\( p = .07 \)). This was due to the greater number of correct responses concrete-only surface focusers gave to Q3-MC, \( F(1, 80) = 9.30, p = .003, n^2 = .10 \). Thus, although surface focusers in the concrete-only condition did not show better CVS understanding in their OE responses, they showed better understanding of the relationship between variable setups (contrasted or controlled) and their potential to affect results in their MC responses.

Final (Experiment 3) understanding. Comparing final CVS understanding of surface focusers in the concrete-only and abstract-fading conditions (Test 1d), there was no difference in likelihood of expressing the rationale for CVS in their final (Q2-OE) responses, \( \chi^2(1, n = 82) = 1.38, p = .24 \). In addition, although there was a significant Condition \( \times \) Focus interaction for MC for all students, there was no main effect of condition for surface focusers, \( F(2, 115) = 1.95, p = .15, n^2 = .03 \).

Concrete only versus concrete fading. Although surface focusers did not differ on the posttest, we first discuss their final instructional understanding (Test 2b). Then we investigate the
effect of Experiment 3 concreteness on transfer from instruction to the posttest (Test 2c).

Final (Experiment 3) understanding. In responses to Experiment-3 Q2-OE, surface focusers were more likely to express the rationale for controlling variables in the concrete-only than concrete-fading condition (22% vs. 3%, respectively), $\chi^2(1, n = 79) = 6.66, p = .01$. However, this was due to concrete-only students expressing more vague or incomplete rationales that did not explicitly associate variable setups (i.e., whether their values differed) with their potential causality. Although surface focusers’ complete rationale responses were associated with higher posttest scores than incorrect responses ($M_{CR} = 5.25, SD = 0.95; M_{IR} = 1.13, SD = 1.86, p < .001$), vague/incomplete rationale responses ($M_{CI} = 1.82, SD = 2.23$) were not ($p = .37$). Thus, the more frequent (complete) rationale understanding expressed by concrete-only students did not result in better posttest performance than incorrect responses. As previously discussed, the number of correct MC responses did not differ across conditions for surface focusers, $F(2, 115) = 1.95, p = .15; \eta^2_p = .03$.

Transfer from Experiment 3 to posttest. To test for differential transfer of surface focusers in the concrete-only and concrete-fading conditions (Test 2c), we included Experiment 3 MC score with condition and grade in ANCOVA. There was no condition by MC response interaction, $F(2, 113) = 1.12, p = .33, \eta^2_p = .02$. Covarying MC score, the effect of condition was (still) not significant for surface focusers, $F(2, 115) = 0.36, p = .70, \eta^2_p = .006$. Thus, Experiment 3 concreteness did not affect transfer among surface focusers. Outcomes per prediction/test are shown in parentheses in Table 3.

Discussion

One obstacle to learning abstract concepts is a tendency for learners to focus on surface rather than deep features of instructional materials. This tendency arises among some students when engaging in tasks involving experimental design (e.g., Schaubele et al., 1991; Siler & Klahr, 2012). In this study, we investigated the effect of varying the concreteness—specifically, the amount of surface information—of experiments presented during instruction on sixth and seventh grade students’ learning of the control of variables strategy (CVS). When learning CVS, the same surface features (e.g., the specific variables and their values) may impair or support understanding. Surface features may impair learning by eliciting task-irrelevant conceptions, including engineering goals variable-specific beliefs. Conversely, these same surface features may support students’ understanding of why controlling variables is necessary to determine whether a variable is causal, an understanding that has been found to promote CVS transfer (Sao Pedro, Gobert, & Sebuwufum, 2011; Siler et al., 2010, 2011).

We investigated whether students’ focus (on deep vs. surface features) shown on the last pretest item interacts with the concreteness of instructional representations to affect learning and transfer outcomes. We predicted that deep focusers—students likely to focus on deep features—would benefit from information conveyed by the concrete representations, specifically, by helping them understand the rationale for controlling variables. We further predicted that—because this understanding leads to more coherent conceptual understanding that supports transfer to new domains—posttest outcomes would be best among deep focusers given only concrete examples (Predictions 1a and 1a). For surface focusers, we expected the effect of concreteness to depend on the trade-off between the potentially distracting and supportive effects of the surface features. Both deep focusers and surface focusers benefited from concrete representations.

In general, the concreteness manipulation affected posttest performance for deep focusers and students who did not set up any controlled experiments on the pretest. Thus, concrete experiments supported learning to control variables in particular. This is reflected by the better understanding of the link between variable setups and their potential to influence experimental results shown by deep focusers given an initial concrete experiment in instruction. Although there were some indications that concrete representations supported surface focusers’ CVS rationale understanding, the concreteness manipulation did not affect their posttest performance.

Effects of Textual Concreteness for Surface Focusers

At first glance, removing surface features seemed to shift surface focusers’ attention from the surface level; surface focusers initially given an abstract representation expressed fewer surface feature-related misinterpretations, including engineering and variable-effect goals. However, surface focusers did not generally shift their focus to the deep features. Rather, many expressed confusion about the lack of detail in the abstract experiment, suggesting they persisted in surface-feature focus. Thus, removing surface features did not prompt surface focusers to attend to deep features. Although the initial benefit of surface features was somewhat greater than the distraction effect, it was not great enough to produce overall better explicit understanding among surface focusers. However, surface focusers showed better initial understanding that contrasted variables may affect the experimental outcome in their MC responses.

By the end of instruction, surface focusers given only concrete representations showed some benefits compared to those given increasingly abstract representations, including greater understanding that confounded experiments are problematic because other factors may affect the outcomes. However, this did not translate into better posttest performance (possible reasons are discussed later). Finally, there was no effect of the concreteness of the final experiment (Experiment 3) on transfer of their final understanding to the posttest. Thus, reflecting on only concrete representations did not appear to lead to domain-specific conceptualizations that hindered transfer or help surface focusers transfer their final understanding to the posttest.

Effects of Textual Concreteness for Deep Focusers

The concrete representation supported deep focusers’ initial understanding that contrasted variables may affect the experimental outcome. The benefit of reflecting on concrete representations was even more pronounced by the end of instruction, where deep focusers in the concrete-only condition gave more correct MC responses and were more likely to express explicit understanding of the rationale for controlling variables in their OE responses than students in the abstract-fading condition. Further, the better final
instructional understanding of concrete-only deep focusers accounted for their higher posttest scores.

As with surface focusers, the concreteness of the final experiment did not negatively affect transfer among deep focusers. In fact, deep focusers transferred their understanding of the rationale for controlling better when the final experiment was concrete than abstract. Although this is consistent with surface features helping deep focusers recognize the applicability of CVS on the posttest, there are reasons to doubt this. Because they showed deep focus on the last pretest item and throughout instruction, we expected deep focusers would interpret posttest questions within a science-goal framework as well. In fact, there was no effect of condition on deep focusers’ recognition of the science-goal nature of the first posttest question, as expressed in their OE responses. What then was responsible for the better transfer of deep focusers in the concrete-only condition? One possibility is that reflecting on concrete representations led to CVS conceptions that were better integrated with causal understanding and thus more robust.

No Effect of Concreteness on Transfer for Surface Focusers

Given the initial advantages of the concrete-only over the abstract-fading condition for both surface and deep focusers, why did only deep focusers show better transfer on the posttest? Students tend to develop CVS understanding incrementally rather than holistically (Siler, Klahr, & Matlen, 2013) and generally articulate the rationale for controlling only after having developed a complete explicit understanding of the procedure. That is, explicit understanding of the rationale for controlling appears to be the last step of CVS development. Further, as found previously (Sao Pedro et al. 2011; Siler et al., 2010, 2011) and in the present study, this complete CVS/rationale understanding is predictive of transfer. Deep focusers—but not surface focusers—in the concrete-only condition were more likely to express a complete understanding of the rationale for controlling variables in their final instructional responses. Only one surface focuser in the concrete-only condition expressed a full understanding of CVS and its rationale at the end of instruction. Thus, concrete-only surface focusers’ failure to develop complete and coherent CVS understandings may explain why they did no better on the posttest.

Supporting Surface Focusers’ CVS Understanding and Transfer

Even with the initial head start provided by concrete representations, surface focusers in the concrete only condition did not develop complete CVS understandings—including the rationale for controlling—during the 40-min instructional period. Had instructional time been extended, surface focusers may have benefited from concrete representations as well. However, surface focusers are likely to be lower-ability and possess poorer problem-solving skills (Evans, Handley, Neilens, & Over, 2010; Newstead et al., 2004; West, Toplak & Stanovich, 2008); thus, they may need more instructional scaffolding. Such scaffolding may include prompts to explain why potential causal variables may affect the experimental outcomes (i.e., because values differ across conditions). Another option is to require students to explain why experiments are good or bad after they have considered whether each variable may affect the outcome. Both options may help students integrate their procedural and conceptual knowledge of CVS and form more coherent understandings.

Previous Findings: Relevant Surface Features Support Lower Knowledge/Ability Students

Similar advantages of concrete representations for transfer were found in Siler and Willows (2014), where middle-school students who learned mathematical rules with concrete figures conveying task-relevant information showed better transfer than students given abstract versions of the training figures. Students with poorer deductive reasoning skills (i.e., lower-ability students) particularly benefited from concrete representations. However, in the present study, only deep focusers—who were likely also higher-ability than surface focusers (Evans et al., 2010; Newstead et al., 2004; West, Toplak, & Stanovich, 2008)—benefited from concrete representations. One reason different ability students benefited from the surface features across studies may have to do with differences in the relevance of the surface features. In Siler and Willows (2014), there was no evidence that the surface features elicited alternate goals or irrelevant conceptions; however, as shown in students’ OE responses on the pretest in the present study, surface features of experimental representations elicited what are likely quite familiar and well-practiced engineering goals and variable-effect responses. That is, the relevance of the concepts elicited by the surface features may determine whether lower-ability students also benefit from concrete representations. However, consistent with Siler and Willows, if surface features are recognized as relevant, lower-knowledge students (i.e., students who did not design any controlled experiments on the pretest in the present study) and lower-ability students (Siler & Willows) appear to particularly benefit.

The conceptual complexity of CVS may be another factor in why concrete representations did not support surface focusers’ transfer performance in the present study. That is, despite its seeming simplicity, the procedural and its rationale may be too complex for some students this age (sixth and seventh graders) to learn—especially given external distractions inherent to learning in classrooms with other students. As discussed previously, students with weaker problem-solving skills may be more likely to simply give up trying to understand the lesson and revert to interpreting the lesson within the more familiar framework of an engineering or variable-effects goal. However, again, further research is needed to address these issues and determine the best ways to support surface focusers’ learning.

How Did Concreteness Support Deep Focusers’ Transfer?

As discussed earlier, deep focusers showed better transfer in the concrete only than concrete-fading condition. Although deep focusers in the concrete only condition were somewhat more likely to show understanding of the rationale for controlling variables by the end of instruction than those in the concrete-fading condition, deep focusers given all concrete examples were better able to transfer their understanding to the posttest. These results appear discrepant from findings that abstract training domains facilitate transfer (e.g., Bassok & Holyoak, 1989; Goldstone & Son, 2005).
For example, in Goldstone and Son (2005), an “idealized” representation led to better transfer than the “concrete” representation. However, when looking beyond the “concrete” labels, similarities emerge. The idealized version of the training simulation—although visually abstract—was labeled “concretely” (i.e., components were still labeled “ants” and “food”). Their relatively more abstract representation is most similar to concrete representations of the present study, which did not include any pictorial representations but components were described more concretely (e.g., as “smooth/rough”). That is, the “idealized” representation in Goldstone and Son may be more concrete than the concrete representation of the present study. In addition, like the concrete representations of the present study to deep focusers, the idealized representations of the Goldstone and Son study likely provided primarily relevant information that supported transfer. Thus, in both the Goldstone and Son and present studies, representations that included additional, primarily relevant information better supported transfer. It may be that only context-specific information that is not relevant to understanding underlying concepts impedes transfer rather than concreteness, per se. The presence of relevant and absence of irrelevant information may result in more coherent conceptual knowledge, which facilitates transfer (Murphy & Medin, 1985; Patalano, Chin-Parker, & Ross, 2006).

This coherence explanation may also account for why the current results also show more supportive effects of surface features than in Gick and Holyoak (1983), where undergraduates were equally likely to apply context-specific and abstract versions of a principle in a transfer domain with different surface features. A plausible explanation for this difference may have to do with the relative coherence of the conceptions learners derived from the concrete and abstract representations in the current vis-à-vis Gick and Holyoak study. That is, in the present study, concrete representations appeared to help students understand the reason for applying CVS. However, in the Gick and Holyoak study, information regarding the rationale for the procedure (i.e., “attacking” a target in a distributed manner to prevent negative effects of a single direction of attack) was available in both representations. Thus, surface features that provide relevant information that increases conceptual coherence may improve transfer outcomes. However, again, further research—for example, research that assesses the coherence of learners’ knowledge resulting from considering representations that vary in concreteness and investigates evidence of its mediating effect—is necessary to shed light on this issue.

Study Limitations

The benefits of concrete representations found in the present study may be related to the nature of the topic (CVS). That is, prior research indicates that people tend to believe variables are causal rather than noncausal (e.g., Gilovich, Vallone, & Tversky, 1985; Masnick & Klahr, 2003; Schauble et al., 1991; Xu & Harvey, 2014). This bias may actually support cross-domain CVS transfer and amplify positive effects of concreteness. For example, believing a particular nonfocal variable affects an outcome may remind students of the need to control it. Thus, CVS may be a topic in which learner biases and surface-level features act in concert to support transfer. Further, the instructional emphasis on the rationale for controlling variables may have been particularly suitable for concrete representations. Outcomes of the present study may have differed if instruction had focused on only procedural aspects of CVS.

Results of the present study differ from recent studies that have found advantages to fading concrete features (see McNeil & Fyfe, 2012). The present results may have differed had the “fading” been done in a way that made the links between the concrete and abstract representations very explicit and deliberate rather than relying on learners to make the connections themselves. Although there were cues linking subsequent representations (i.e., explicit statements that “Now we’ll [add/remove] some details to describe Amal’s next experiment”), these cues may have been insufficient, especially given the high cognitive load students likely experienced during instruction.

Finally, to convey minimal additional information beyond that of potential causality, the experiments in concreteness shown during instruction tended to include minimal detail. For example, they did not include pictures of the ramp experiments and physical properties were represented relatively abstractly (i.e., “rough/smooth” was used as variable values for type of surface rather than a “sandpaper/glass” surface, for instance). Perceptually richer representations would likely cause poorer learning outcomes, especially among surface focusers (cf. Goldstone & Wilensky, 2008). This is another question for future research. More generally, further research is needed to specify how the concreteness of instructional materials, student characteristics—including domain knowledge and focus proclivities—and the nature of the learning domain interact to affect understanding and transfer.
EFFECTS OF TERMINOLOGICAL CONCRETENESS


