Adapting an Effective Lesson Plan for a Computer-based Tutor[1]

Stephanie Siler[1] (siler@cmu.edu), David Klahr[1], Mari Strand-Cary[2] Cressida Magaro[1],

and Kevin Willows[1]

[1]Carnegie Mellon University

[2]Pacific Institutes for Research

Abstract

We describe the adaptation of an effective method for teaching high-SES middle school children the Control of Variables Strategy (CVS) into instruction delivered by a computer-based tutor (named TED, for "Training in Experimental Design"), which is aimed at a wider range of students. In high-SES schools, our "explicit" CVS instruction has been shown to be very effective (e.g., Chen & Klahr, 1999; Strand-Cary & Klahr, 2008). However, when the same explicit instruction was delivered to low-SES students with limited experience with science inquiry, CVS mastery rates were much lower (Klahr & Li, 2005). Through our one-to-one tutoring of students who failed to develop CVS mastery from the classroom instruction, we discovered that students lacked necessary prerequisite knowledge and frequently held alternative goals and misconceptions about the point of the lesson. To address these problems in TED, we enhanced our previous explicit instructional script by adding an introductory lesson addressing the prerequisite concepts. In addition, we replaced variable levels that have known effects with nonsense variable names to prevent the inadvertent elicitation of engineering goals. We will discuss the results of comparisons of the transfer rates of students given the enhanced human-delivered explicit instruction with the learning rates of both students who were given the original human-delivered explicit instruction and students given TED-delivered enhanced instruction.

Introduction

The goals of this project are to find instructional methods that lead to efficient learning of important concepts and skills and to uncover the mechanisms of learning and transfer of those concepts and skills. The Control of Variables Strategy (CVS) is one such important and domain-general skill that our lab has studied over the past decade (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008). Briefly, CVS involves controlling all variables in an experiment except for the one variable under investigation. In high-SES classrooms, "explicit" CVS instruction that emphasizes students' understanding of the reasons for designing unconfounded experiments, has been shown to be very effective (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008). However, with a more challenging population—students from lower-SES classrooms who have more limited exposure to science inquiry—this

[1] The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H060034 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

success is more limited (Klahr & Li, 2005). In this paper, we describe the process of adapting an effective method for teaching high-SES middle school children the Control of Variables Strategy (CVS) into instruction delivered by a computer-based tutor (named TED, for "Training in Experimental Design") that is adaptable to a wider range of students, including this more challenging low-SES student population.

Perhaps ironically, our current project does not use the CVS method in the iterative process of designing and modifying TED because the limited time and resources available often make one modification at a time in the initial stages of development unfeasible. Rather, TED development uses design-based research, informed by both empirically supported learning theory and prior findings in cognitive science in addition to evidence obtained from our evaluations. Identifying good candidates for underlying predictors of learning from our evaluation data can help to pinpoint which aspect of the instruction are effective and why as well as inform TED development. However, due to the nature of design-based research, we believe it is important to test our hypotheses about causal factors in learning in subsequent evaluations. Our goal in TED development is to increase the adaptivity of TED to allow a wide range of students to develop robust CVS knowledge.

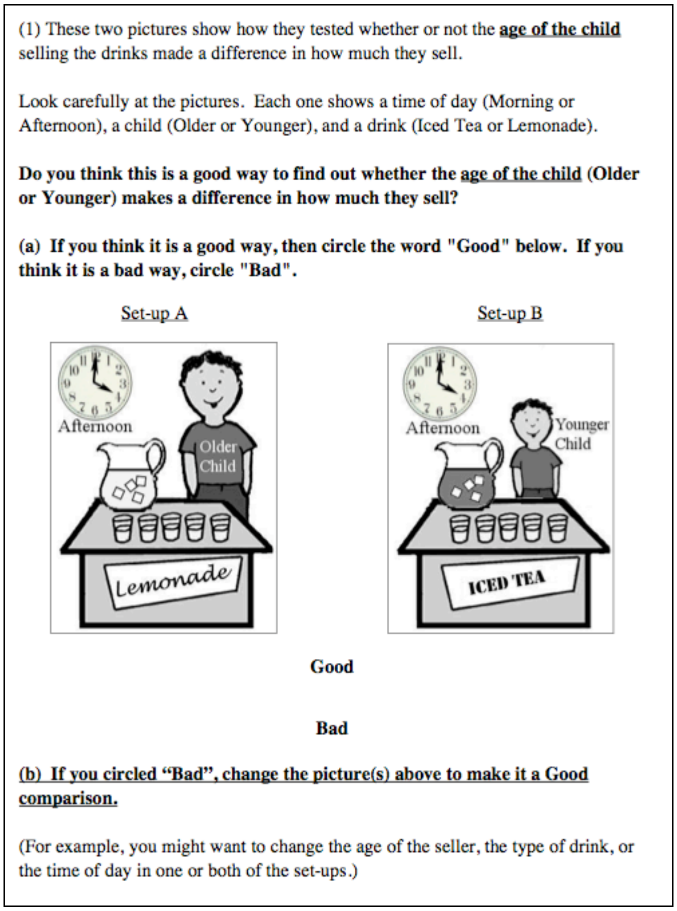
Phase 1: whole-classroom-delivered "original" instruction.

Participants: In the first phase of this project, we worked with U.S. 6th-grade students and three experienced teachers at three different local K-8 private Catholic schools. Two of these schools (classrooms L1 & L2) served low-SES students, and one (classroom M1) served middle-SES students. The low-SES student population in L1 and L2 is the focus of the current study. L1 consisted of 23 students (12 boys and 11 girls) mean age 11 years and 7 months (SD = 5.22), L2 consisted of 17 students (5 boys and 12 girls) mean age 11 years and 9 months (SD = 5.53).

Procedure: Prior to the intervention, participating teachers were trained in a 3-hour workshop on the intervention procedure, including the "Explicit" CVS instruction, which focuses on the rules and rationales for setting up informative (i.e., unconfounded) experiments. This instruction was similar to the one-to-one instruction given to students in Chen and Klahr (1999) and is described in more detail shortly.

During the intervention, students first completed a "Story-evaluation" pretest that required them to evaluate six experiments in three different domains (selling drinks, rockets, and cookies).

Figure 1. Item from the Story-evaluation test of Phase 1.



(1) These two pictures show how they tested whether or not the **age of the child** selling the drinks made a difference in how much they sell.

Look carefully at the pictures. Each one shows a time of day (Morning or Afternoon), a child (Older or Younger), and a drink (Iced Tea or Lemonade).

Do you think this is a good way to find out whether the age of the child (Older or Younger) makes a difference in how much they sell?

(a) If you think it is a good way, then circle the word "Good" below. If you think it is a bad way, circle "Bad".

Set-up A

Set-up B

Good

Bad

(b) If you circled "Bad", change the picture(s) above to make it a Good comparison.

(For example, you might want to change the age of the seller, the type of drink, or the time of day in one or both of the set-ups.)

For example in the "selling drinks" story problem (Figure 1), the "experiment" was to find out whether the age of the seller affected sales, and students were asked to evaluate a comparison between two stands that could differ in time day (early or late), type of drink (lemonade or tea), age of seller (adult or kid). For some problems (as in Figure 1), the story depicted a confounded experiment and in others, the story depicted an unconfounded experiment. The child's task was to decide whether or not the depicted experiment was "good" or not, and if not, to fix it.

The next day, the teacher introduced a ramps apparatus (Figure 2), which consisted of four variables that might determine how far a ball rolled after leaving the ramp, (slope—steep or not steep, starting position of the ball—at the top or middle, surface—smooth or rough, and ball type—yellow or pink).

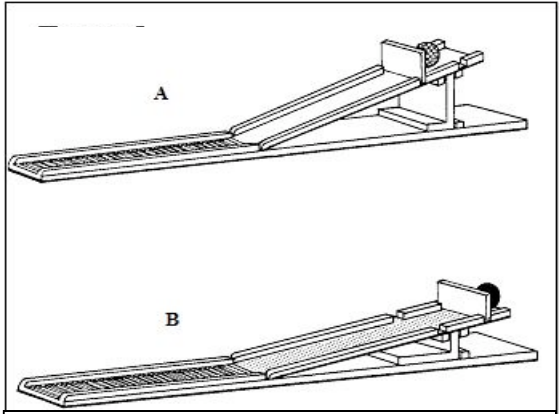


Figure 2: the depicted experiment is completely confounded because it contrasts (A) a “yellow” ball on a steep, smooth, short ramp with (B) a “pink” ball on a shallow, rough, long ramp.

Students designed experiments on paper by filling in settings for each ramp into a table to test each of these four variables on the ramps pretest. Then the teacher presented CVS instruction to the entire class. First, the teacher presented a confounded experiment using two ramps and asked students to individually evaluate the experiment as a “fair” or “unfair” way to find out about the target variable on their worksheets. Next, the teacher led a class-wide discussion about whether or not the experiment under discussion was informative (i.e., enabled one to make inferences about effects of the target variable) and why it was not informative. During this discussion, the teacher repeatedly asked whether the set-up would let them “know for sure” whether the target variable affects how far the balls roll. As students identified the confounded variables, the teacher controlled them until eventually the experiment was unconfounded (i.e., informative for the target variable). Then students individually wrote explanations for why the corrected experiment was now a fair way to find out about the target variable. This sequence was repeated two more times with other ramps set-ups. On the final day of instruction, students completed the ramps posttest in which they again designed experiments on paper to test each of the four ramps variables. This was considered a measure of near transfer performance because students were required to design set-ups in the instructional domain (ramps).

Students who did not show CVS near transfer mastery by designing at least three out of four unconfounded ramps experiments on the posttest were tutored by a member of our research team. For the non-experimenting classroom L1, we repeated the classroom instruction to rule out the possibility that students had simply not paid close attention in class.

Approximately three weeks after the tutoring phase ended, all students who received the classroom instruction—both tutored and untutored students—completed the delayed Story-evaluation posttest, identical to the Story-evaluation pretest. We considered this a measure of far transfer performance because students were required to evaluate and, if necessary, fix experiments in domains different from the instructional domain.

Results: The rate of CVS near transfer mastery–defined as designing at least 3 out of 4 unconfounded ramps experiments—for the two classrooms of low-SES students was low (33% and 36% for L1 and L2, respectively) as expected, but the mastery rate was higher for the middle-SES classroom experienced in science inquiry (77%).

In the remedial tutoring sessions, we discovered that simply repeating the classroom instruction was not successful. Thus, we sought to identify preconceptions that might have affected students' ability to successfully learn CVS in this and subsequent tutoring sessions in L2 and M1, and in all subsequent phases of our project. We found that students often lacked necessary prerequisite knowledge and frequently held alternative goals and misconceptions about the point of the lesson. For example, some students did not understand that an experiment required at least two conditions so that the results from each could be compared. Others were so sure of the non-causal role of some factors that they saw no reason to control them, and thus created confounded experiments. More profoundly, students misunderstood the goal of instruction. Some students adopted "engineering goals" (Schauble, Klopfer, & Raghavan, 1991) in which they attempted to "engineer" a particular outcome, such as maximize the combined effects of all variable settings to maximize outcomes or to produce the same outcome in both conditions. Students also misinterpreted the instruction to be about the substantive domain in which the experiment was being conducted and focused their attention on predicting and discussing the effects of variables (e.g., that rough surfaces make balls roll slower than smooth surfaces) rather than about learning the logic of, and procedures for, designing good experiments. This is not terribly surprising, given the emphasis on factual learning typical in science classes. The underlying problem with all of these difficulties is that students are making use of their beliefs about the causal effects of variables rather than adopting the goal of designing experiments to find out whether such effects indeed exist. These proclivities may have hindered students from understanding the procedural and conceptual content of instruction, including why it is logically necessary to control all variables except the one under investigation.

Through an analysis of students' developing knowledge of CVS rules from available data, and consistent with ACT-R learning theory (e.g., Anderson & Schunn, 2000), we found that students tended to develop CVS knowledge gradually, rather than in "one fell swoop". Understanding the need for two conditions and to vary the target variable, to "compare and contrast" outcomes always preceded understanding the need to control other variables. Furthermore, we found that students who entered the experimental evaluation phase of instruction with some initial CVS rules derived from their ramps pretest designs (e.g., if they consistently varied the target variable, it was assumed that they had some at least implicit understanding of "comparing and contrasting" that variable) were significantly more likely to gain CVS during the experimental evaluation phase than those students with no incoming knowledge.

<u>Instructional enhancements</u>: To address these issues, we enhanced our previous explicit instruction scripts by adding an introduction in which the tutor states the point of the lesson as learning how to design good experiments to answer a question, and discusses prerequisite CVS concepts, including the idea of "comparing and contrasting" different levels of a variable to find out whether that variable has an effect on an outcome. This addition served to highlight the goal of the lesson as well as to provide CVS concepts that were not the focus of the Explicit Instruction. The corresponding computerized form of this instruction was a video of a "live" human presentation of these points, accompanied by Flash-based graphics. These graphics were shown on paper in the human-tutor condition.

In addition, to further help prevent both the inadvertent elicitation of engineering goals and misinterpretations of the instruction as about discussion of the effects of variables rather than designing experiments (to find out about their effects), we replaced variable levels that have known effects (e.g., a rough or smooth surface for a ramp) with nonsense variable level names (e.g., a "SIF" or "FIM" surface). Because these variable levels are not real, we reasoned, students could not resort to applying their knowledge of these variable effects during instruction, thereby decreasing elicitation of engineering goals. Furthermore, because these values are unknown, we thought that the science goal of finding out about these variables would become more salient.

To inform the development of TED, we must confirm that students—and low-knowledge students in particular—learn more from the enhanced than original explicit instruction. Results of the instructional form manipulation (i.e., original versus enhanced instruction), including any student characteristic by instruction interactions, will provide information useful in the development of instructional aspects of TED.

For example, it will help determine the more pedagogically efficient pathway to send a particular student down based on his/her knowledge state and other characteristics. Furthermore, we must ensure that this enhanced instruction—if in fact it is more effective—translates from a human-delivered to computer-delivered modality.

Comparison of human-delivered enhanced instruction with original Phase 1 instruction:

Participants: This “enhanced” instruction was delivered to U.S. 5th-grade students in one middle-SES and two low-SES classrooms (L3 and L4) in different schools. Instruction was administered either by a human tutor or by the TED computer tutor. Though in this section we only discuss results for the Human-tutored condition, we present the procedure for both conditions. Later, we compare the outcomes of Human- versus TED-tutored students. Participant characteristics are shown in Table 1 below.

Table 1. Sample size and ages, by classroom and condition in Phase 2.

	L3	L4	n
Human-tutored	8 (3 girls; 5 boys) M = 11 yrs, 0 months	7 (2 girls; 5 boys) M = 11 yrs, 0.2 months	15
TED-tutored	8 (1 girl; 7 boys) M = 10 yrs, 11 months	7 (3 girls; 4 boys) M = 11 yrs, 0.5 months	15
n	16	14	N = 30

To confirm the advantage of the enhanced explicit instruction for lower-knowledge students in particular, we first compared student learning from the human-delivered enhanced explicit instruction to student learning from the original explicit instruction given in Phase 1. In the human-delivered enhanced instruction condition, students received spoken instruction and gave oral responses, and worked with physical apparatuses (e.g., ramps) when designing and evaluating experiments.

Phase 2 procedure: Students in Phase 2 completed the following:

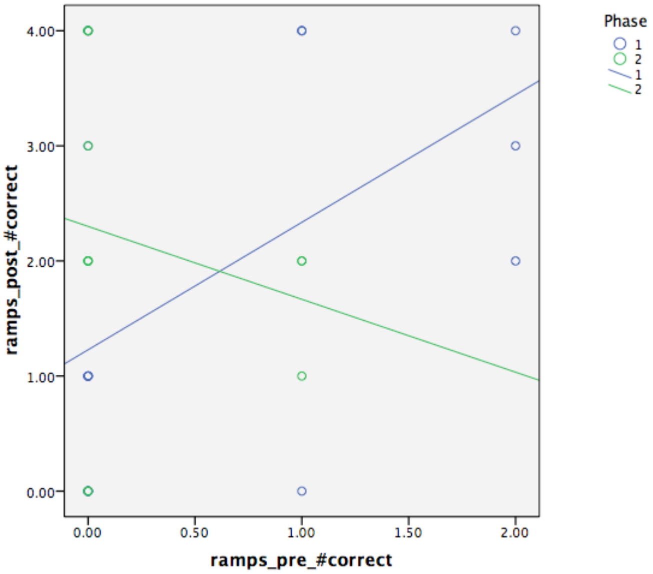
- Story pretest on designing or evaluating experimental design problems presented as "story problems". These story problems included three different contexts—cookie baking, drink sales, and rocket ship design.
- Ramps pretest: Pretest on designing one experiment for each of the four ramps variables. In the Human-tutor condition, physical rather than virtual ramps were used throughout.
- Immediately after completing the ramps pretest, students were given a brief (~ 2 min) introduction to experimental design, its purpose, scope, and the central idea of "comparing and contrasting", either in video form (TED) or presented live by the human tutor (Human condition).
- Afterwards, students underwent the Explicit Instruction with deep questioning. In this portion of the instruction, students were presented with ramps set-ups and probed for why the particular design was or was not "a good way" to find out about the target variable and whether the design would allow them to "know for sure" whether the target variable made a difference. After students responded to these deep questions, they were given feedback and a lengthy explanation for why the design could—or could not—lead to valid inferences about the target variable. Experimental confounds were corrected by the tutor, and students answered the same two deep probes as above, then given feedback and an explanation of why the unconfounded experiment could lead to valid inferences about the target variable.
- Immediately after completing the explicit instruction portion, students completed the ramps posttest in which they designed ramps experiments; this assessed near-transfer performance.
- The following day, students completed the Immediate Story posttest (identical to the Story pretest); this assessed students' immediate far-transfer performance.
- Approximately three weeks after completing the immediate Story posttest, students completed the delayed Story posttest, which was identical to the Story pre/posttests except that the target variables were changed for each item.

Results: First, we wanted to see what effect the "enhanced" CVS instruction had on low-SES students' near transfer performance. To do this, an ANCOVA was run, where the number of unconfounded set-ups on the ramps posttest was the dependent measure, type of instruction (Phase 1 or human-tutored condition in

Phase 2) was the independent variable, and ramps pretest number of CVS set-ups and Grade Equivalent reading comprehension scores from the TerraNova[2] were covariates. Grade Equivalent reading scores were used as a measure of reading comprehension because Phase 1 students were 6th-graders and Phase 2 students were 5th-graders.

There was a significant phase by ramps pretest score interaction, *F*(1, 29) = 10.84, p = .003 (shown in Figure 3). For Phase 1 students, there was a significant positive correlation between ramps pretest and ramps posttest scores (r = +.46, p = .02), but for Phase 2 students given the enhanced instruction, there was no significant relationship (r = -.17, p = .40). For students who did not set up *any* unconfounded comparisons on the ramps pretest, students given the enhanced instruction had significantly higher ramps posttest scores than their counter-parts (M = 2.63, SD = 1.41; M = 1.16, SD = 1.46), *F*(1, 24) = 7.09, *p* = .01. However, for students who set up one or two unconfounded comparisons on the ramps pretest, students given the original instruction had significantly higher posttest scores than students given the enhanced instruction (M = 3.40, SD = 0.89; M = 1.67, SD = 0.58), *F*(1, 24) = 7.09, *p* = .01.

Figure 3. Relationship between ramps pretest and posttest scores by phase.



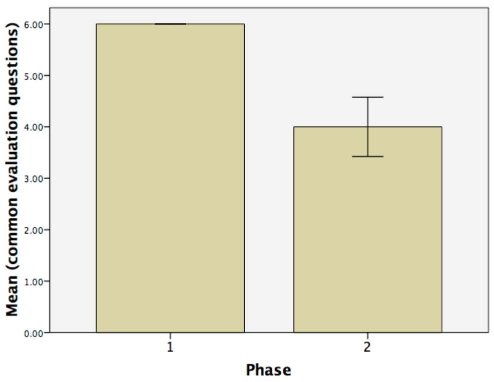
[2] TerraNova is a series of standardized achievement tests used in the United States designed to assess K-12 student achievement in reading, language arts, mathematics, science, social studies, vocabulary, spelling, and other areas. The test series is published by McGraw-Hill. TerraNova scores are vertically equated, allowing one to compare scores across grades.

This result is consistent with our goal in enhancing the instruction—to help the lowest-knowledge students in particular by providing instruction on prerequisite CVS concepts in the introductory portion of instruction. But why did learning for mid-knowledge students decrease when given the enhanced instruction? One possibility is that the prerequisite instruction impaired students with some incoming CVS knowledge by shifting the focus of the instruction away from the core underlying conceptual explanation that is the focal point of the Explicit Instruction experimental evaluation phase of instruction.

Our expectation that the use of nonsense variable names for ramp surfaces would prevent students from hypothesizing about their effects was not supported by the data. Instead, many students assumed that "Sif" was "Stiff" and "Fim" was "Firm" and hypothesized about their effects accordingly. Even if they did not mistake these terms for familiar ones, they still often hypothesized about their effects. Though we believe that the addition of the introduction video likely produced the ramps pretest by instruction interaction, it is important to point out that other factors that differed between Phase 1 and 2 instruction (e.g., whole-class vs. individualized instruction) may have played a causal role in weakening the relationship between prior knowledge and CVS gain; evidence of causality will be investigated in future evaluations.

But what effect did the enhanced instruction have on students' far transfer performance? To answer this question, we compared students' performance on three experimental evaluation questions that were common in the delayed posttests given in Phases 1 and 2. One point was given for correct evaluation (i.e., as a good or bad way to find out about the target variable) and one was given for converting the experiment into an unconfounded comparison (6 point maximum). Using Grade Equivalent reading comprehension scores and ramps pretest scores as covariates, there was no significant difference between the types of instruction for this delayed far transfer performance measure. Students given the original instruction tended to perform somewhat better than those given the enhanced instruction (M = 4.00, SD = 2.04; M = 3.10, SD = 1.45), $F(1, 32) = 2.67$, $p = .11$. However, this comparison includes students in Phase 1 who received tutoring in addition to classroom instruction and thus were given more instruction than their Phase 2 counter-parts. Because all students who were not tutored in Phase 1 demonstrated mastery on the ramps posttest, we compared them to just those students in Phase 2 who also demonstrated CVS mastery on the ramps posttest. For just these ramps mastery students, those given the original instruction actually out-performed students given the enhanced instruction on the common delayed Story posttest items, *F*(1, 9) = 39.27, *p* < .001[3] (Figure 4). We will discuss possible explanations for this difference later.

Figure 4. Mean delayed Story posttest scores for near-transfer mastery students by phase.



Comparison of human- to TED-delivered "enhanced" instruction:

Since the human-delivered enhanced instruction led to higher ramps posttest scores for the lowest-knowledge students, we will discuss the Phase 2 comparison of Human- with TED-delivered instruction to ensure that there were no decreases in learning rate across delivery modalities. Poorer outcomes in the TED than in the human condition necessitate identifying the source(s) of the discrepancy and using this information to improve the TED-delivered instruction.

Failure of TED with enhanced instruction to produce high CVS learning and transfer rates necessitates identifying predictors of learning and mastery that are plausible causal factors and revise TED instruction accordingly. We did not anticipate that factors associated with learning and transfer would differ between human- and TED-tutored conditions, but tested for interactions with condition to ensure that this was the case.

As discussed previously, in the human-delivered instruction condition, students received spoken instruction and gave oral responses and worked with physical apparatuses (e.g., ramps) when designing experiments. In the TED-delivered condition, the instruction was presented via audio human voice-over; key points of the audio-delivered instruction were also presented on-screen in text form. Students typed all

[3] For low-knowledge students who did not design any CVS experiments on the ramps pretest, there was no significant difference between instruction (p = .49). However, mid-knowledge students who set up either one or two CVS experiments on the ramps pretest given the original instruction out-performed mid-knowledge students who were given the enhanced instruction (p = .05).

responses in text boxes. Students in this condition interacted with virtual apparatuses to design experiments. Previous experiments have found no significant differences in performance when students interacted with virtual or physical apparatuses (Triona & Klahr, 2003; Klahr, Triona, & Williams, 2007), so we did not expect that to be a source of any learning differences. The content of instruction was the same in both the human-tutor and TED-tutor conditions.

Results: In an ANCOVA with ramps posttest score as the dependent variable, classroom and condition (TED-tutored or Human-tutored) as the independent variables, and ramps pretest and reading comprehension score as covariates, there was a significant condition by classroom interaction, $F(1, 17) = 6.17$, $p = .02$. In classroom L3, there was a significant difference in favor of TED-tutored over Human-tutored students (M = 3.67, SD = 0.82; M = 1.4, SD = 0.89, respectively), $F(1, 9) = 19.30$, $p = .002$. However, in classroom L4, there was no difference between the TED- and Human-tutored conditions on the ramps posttest (M = 2.83, SD = 1.17; M = 3.00, SD = 1.73, respectively), $p = .67$.

For immediate far transfer performance on the immediate Story posttest, there was no significant condition by classroom interaction, nor was there a significant main effect for either classroom or condition. Similarly, on the delayed far transfer Story posttest, there was no significant condition by classroom interaction or main effect of condition, but there was a strong trend for classroom, favoring students in L3, $F(1, 16) = 9.84$, $p = .08$. Thus, the computerized version of the enhanced instruction produced similar near and far transfer outcomes as the same human-delivered instruction. However, transfer mastery rates (shown in Table 2), especially far transfer rates, were still generally much lower than what was acceptable to us. Thus, we sought to identify the predictors of transfer.

Table 2. Summary of Phase 2 mastery results.

	Near transfer mastery (ramps post)[a]	Immediate far transfer mastery[b]	Percentage immediate far transfer mastery conditional on near transfer mastery	Delayed far transfer mastery with time delay (Story follow-up post)[b]	Percentage far transfer mastery with time delay (follow-up Story post) conditional on near transfer mastery
Human	30.8%	7.7%	25%	7.7%	0%
TED	64.3%	14.2%	22.2%	14.2%	25%

[a] At least 3 of 4 unconfounded ramps set-ups. [b] At least 5 out of 6 unconfounded set-ups.

Predictors of transfer

Our first step in determining the best predictors of CVS learning and transfer was to find out which initial knowledge and standardized measures were most highly correlated with posttest performance for the two low-SES classrooms. In a forward regression, of ramps pretest, Story pretest, and standardized (CTB/TerraNova) reading comprehension, science, nonverbal "IQ", and verbal (or deductive) reasoning national percentile scores, only reading comprehension scores were significantly related to ramps posttest scores ($r = +.47$, $p = .03$). This relationship did not differ by classroom or condition (Human- or TED-tutored). The same result was found for L1 of Phase 1: using same variables in a forward regression, only reading comprehension was significantly related to ramps posttest scores ($r = +.87$, $p < .001$). In both cases, because instruction was presented orally by the teacher or with audio voice-over, we believe that a more general comprehension ability underlies the relationship between reading comprehension and near transfer.

Of both pretest and all standardized measures, only deductive reasoning was significantly positively related to the measure of immediate far transfer, the immediate Story posttest ($r = +.58$, $p = .006$). Again, there were no interactions with classroom or condition, so this relationship did not differ based on those factors. The verbal deductive reasoning measure assesses the skill of deriving a conclusion based [only] on the information given (in two or three sentences). The following is an example of a verbal deductive reasoning item from a practice booklet (Level 2, for U.S. grades 4-5):

A fire must have heat, air, and fuel or it will not burn.
Wood can be used as fuel for a fire.
The scouts made a campfire.

- The scouts used wood to make their campfire.
- The scouts toasted marshmallows over their fire.
- *The campfire had heat, air, and fuel.*
- The campfire burned for a long time.

Similarly, using the immediate Story posttest in addition to the same independent variables as above in a forward regression, only deductive reasoning and immediate Story posttest were significantly related to delayed Story posttest score (r = +.49, p = .04, for both variables). And again, these relationships did not differ by classroom or condition. Similarly, in classroom L1 of Phase 1, where all standardized measures were available, using all these independent variables (with the exception of the immediate Story posttest, which was not administered in Phase 1), only deductive reasoning was significantly related to the delayed Story-evaluation posttest (r = +.82, p = .001). The relationship between deductive (rather than inductive) reasoning and transfer has been previously reported (e.g., Novick, 1995). Novick hypothesized about the mechanism of analogical transfer in algebra word problems that could explain this relationship: "Because mapping can be described as a process of constraint satisfaction in which the existence of certain correspondences implies the existence of other, related correspondences (Holyoak & Thagard, 1989), I predicted that deductive reasoning would be a reliable predictor of success at mapping." (p. 13).

We believe deductive reasoning may be related to far transfer of CVS for a different reason. Deductive reasoning involves integrating given information and drawing conclusions based on just that information. Similarly, understanding the underlying logic of the CVS procedure involves integrating given information about the experimental set-up and results to form valid causal conclusions—i.e., understanding that only one variable can be changed across conditions because *only then can one know that that is the cause of any differences*. Because this understanding involves the integrative process of deductive reasoning, it may explain the relationship between deductive reasoning and far transfer.

Consistent with this prediction, prior research (Matthews & Rittle-Johnson, 2009) has found that conceptually-oriented explanations were predictive of procedural transfer and higher quality explanations were positively related to performance. Therefore, we performed a finer-grained analysis and coded for students' "highest quality" responses—those that demonstrated a complete understanding of the determinate nature of an unconfounded set-up (or the indeterminate nature of a confounded experiment). For example, when given the probe: "Imagine the balls rolled different distances. Could you tell for sure that the surfaces caused the difference?", one TED-tutored student responded: "Yes. Because everything is the same and if there is a difference it's because of the surface". This response explicitly demonstrates an understanding of the determinate causal relationship between the target variable and outcome. In contrast, the following response to the same probe, though correct, is of lower quality because it does not explicitly express the link between the variable and outcome differences: “Yes. Because everything is the same except [surface].”

Whether or not students explicitly expressed this relationship during the experimental evaluation portion of instruction was more highly related to their deductive reasoning scores than any other pretest or standardized measure. Furthermore, expression of this relationship was more highly related to deductive reasoning than other coded measures (e.g., number of correct responses, number of engineering and variable-effect responses, the difference between number of correct and engineering/variable-effect responses). Thus, students’ deep understanding of CVS is closely tied to the measure of deductive reasoning.

With reading comprehension and ramps pretest scores as covariates in an ANCOVA, whether students expressed this full relationship was not positively related to ramps posttest score (in fact, when reading was covaried, expression of this relationship was negatively related to ramps posttest CVS score). Thus, this deeper understanding did *not* predict immediate near transfer performance. However, for immediate far transfer, with both deductive reasoning and expression of the determinate relationship in an ANCOVA, only whether or not students expressed this relationship during the experimental evaluation phase was significantly related to immediate Story posttest performance. Deductive reasoning was no longer significantly related to far transfer performance. Thus, this CVS logic understanding may explain the relationship between deductive reasoning and far transfer.

Similarly, in an ANCOVA with immediate Story posttest and deductive reasoning as covariates, delayed far transfer as assessed by the delayed Story posttest was only predicted by whether or not students explicitly expressed the determinate relationship, and not by immediate Story posttest performance or deductive reasoning. Nor did it interact with condition, and thus was predictive of far transfer performance for both human- and TED-tutored students as anticipated. Thus, again, this measure of deep conceptual understanding was predictive of far transfer performance, and a good causal factor candidate. This may explain the far transfer advantage—especially for ramps mastery students—of Phase 1 original instruction over Phase 2 enhanced instruction. In Phase 1, teachers asked students whether or not the experiment could allow them to make definite inferences about the target variable and whether they could know for sure that the target variable would cause the outcome if more than one variable was different across conditions more frequently than in Phase 2, where each question was asked twice per evaluated experiment. Thus, it is possible that increased opportunities to think about and understand the determinate relationship between the set-up and outcome in Phase 1 led to improved far transfer performance.

Implications for TED development:

One way to address the finding of the relationship between reading comprehension and CVS near transfer is to make the experimental evaluation portion of the instruction more interactive and adaptive. Increasing the interactivity and adaptivity of this instructional portion—and therefore presenting the explanations more incrementally than in Phase 2 instruction—may help to ensure that students both attend to and understand the full content of this critical instruction. Remediation will be given when student responses indicate engineering or variable-effect misinterpretations. Evidence of the effectiveness of these modifications includes both improved near transfer performance and a weaker relationship between reading comprehension and CVS learning than was found in past instruction.

One way to address the finding of the relationship between expression of an understanding of the causal relationship between the set-up and outcome in TED development is to more frequently prompt students to think about and express the determinate relationship between the experimental set-up and outcomes during the evaluation portion of instruction. Prompting students to express the determinate relationship can be done at other points during instruction as well, for example, when students set up confounded experiments on the ramps and Story tests, they can be given feedback prompting them to consider why their design cannot lead to determinate inferences about the causal variable. Improved far transfer performance and significant relationships between students' articulations of determinate relationships and far transfer found in this and future evaluation iterations will support its status as a causal factor in CVS far transfer.

In addition, to test our hypothesis that the video introduction of the instructional goal and prerequisite information was responsible for the pretest by treatment interaction we found when comparing Phases 1 and 2, half of participating students will be given the full instruction and half will be given instruction without the video. If we find a similar interaction, TED instruction will be adapted to students' ramps pretest score.

Once these modifications are implemented in TED, we will continue our design-based cycle of evaluating and revising TED until it is capable of supporting all students in developing robust and transferable conceptual and procedural knowledge about the fundamentals of experimental design.

References

Anderson, J. R. & Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser, (Ed.), Advances in instructional psychology: Educational design and cognitive science (Volume 5), pp. 1-34. Mahwah, NJ: Lawrence Erlbaum Associates.

Chen, Z. & Klahr, D. (1999). All other things being equal: Children's acquisition of the control of variables strategy. *Child Development, 70*(5), 1098–1120.

Holyoak, K. & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295-355.

Klahr, D. & Li, J. (2005). Cognitive Research and Elementary Science Instruction: From the laboratory, to the classroom, and back. *Journal of Science Education and Technology, 4*(2), 217-238.

Klahr, D. & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15*(10), 661-667.

Klahr, D., Triona, L. M., & Williams, C. (2007). Hands On What? The Relative Effectiveness of Physical vs. Virtual Materials in an Engineering Design Project by Middle School Children. *Journal of Research in Science Teaching , 44*, 183-203.

Matthews, P. & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology, 104*(1), 1-21.

Novick, L. (1995). Some determinants of successful analogical transfer in the solution of algebra word problems. *Thinking and Reasoning, 1*(1), 5-30.

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

Strand-Cary, M. & Klahr, D. (2008). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development, 23*(4), 488-511.

Triona, L. M. & Klahr, D. (2003). Point and Click or Grab and Heft: Comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition & Instruction, 21*, 149-173.