Journal of Educational Psychology
2010, Vol. 102, No. 1, 90–101

© 2010 American Psychological Association
0022-0663/10/$12.00 DOI: 10.1037/a0017972

Learning the Control of Variables Strategy in Higher and Lower Achieving Classrooms: Contributions of Explicit Instruction and Experimentation

Robert F. Lorch Jr., Elizabeth P. Lorch, William J. Calderhead, Emily E. Dunlap, Emily C. Hodell, and Benjamin Dunham Freer
University of Kentucky

Students (n = 797) from 36 4th-grade classrooms were taught the control of variables strategy for designing experiments. In the instruct condition, classes were taught in an interactive lecture format. In the manipulate condition, students worked in groups to design and run experiments to determine the effects of four variables. In the both condition, classes received the interactive lecture and also designed and ran experiments. We assessed students' understanding using a written test of their ability to distinguish valid from invalid experimental comparisons. Performance on this test improved from the pretest to the immediate posttest in all conditions, and gains were maintained at a 5-month delay. For students from both higher and lower achieving schools, gains ordered as follows: both > instruct > manipulate. However, students from higher achieving schools showed greater gains in all conditions. Item analyses showed that the interactive lecture improved students' understanding of the need to control irrelevant variables, and experimentation improved students' understanding of the need to vary the focal variable.

Keywords: science education, inquiry in science, explicit instruction, experimentation, scaling up

Reasoning in science, like reasoning in other complex domains, involves processes that are susceptible to various forms of error. However, unlike many other domains, science has developed a method of inquiry that guides logical reasoning. The experimental method in science is based on the *control of variables strategy* (CVS), which defines a procedure for designing interpretable experiments: The researcher must manipulate variables of interest while holding constant all other variables (Chen & Klahr, 1999; Ross, 1988). CVS also entails a method of deriving logical inferences from a valid experiment by comparing results for conditions that differ on only one variable. In short, CVS is a core skill in scientific reasoning.

Instruction in CVS has received a great deal of research attention because a theoretical account of the components of effective CVS instruction is likely to have implications for successful instruction of other topics in science. In fact, research on learning of CVS has been a focal point for a recent debate concerning the relative efficacy of explicit instruction compared with discovery-based experimentation (Dean & Kuhn, 2007; Klahr, 2005; Kuhn, 2005; Kuhn & Dean, 2005). Rather than focusing on the relative efficacy of explicit instruction and discovery by experimentation, the first goal of our study is to examine their potentially separate contributions to learning of CVS by fourth-grade students.

A second motivation for our study is more practical. Even though science educators and policymakers emphasize the importance of teaching CVS beginning early in elementary school (Kentucky Department of Education, 2006; National Research Council, 1996), most elementary school children are not very adept at designing experiments or drawing valid inferences from experiments (Bullock & Ziegler, 1999; Dunbar & Klahr, 1989; Kuhn, 1989; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Schauble, 1996; Schauble, Glaser, Duschl, Schulze, & John, 1991). Given the importance of CVS, can teaching interventions be developed that are effective and practical in the classroom?

Research by Klahr (Chen & Klahr, 1999; Klahr & Nigam, 2004; Li, Klahr, & Jabbour, 2006; Toth, Klahr, & Chen, 2000) and by Zohar (Zohar & Aharon-Kravetsky, 2005; Zohar & David, 2008;

Robert F. Lorch Jr., Elizabeth P. Lorch, Emily E. Dunlap, Emily C. Hodell, and Benjamin Dunham Freer, Children at Risk Research Cluster, Department of Psychology, University of Kentucky; William J. Calderhead, Department of Special Education and Rehabilitation Counseling, University of Kentucky.

The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. The research reported here was supported by U.S. Department of Education, Institute of Education Sciences Grant 1 R305 H060150-01 to the University of Kentucky. We are grateful to many people in the Fayette County public schools for their support of this research. George McCormick supported the initial research proposal, and the following directors of elementary schools gave their enthusiastic approval: Carmen Coleman, Julie Hawkins, and Fabio Zuluaga. Thanks to all of the principals and fourth-grade teachers who provided access to their classrooms. We are very grateful for the hard work of the teachers who coordinated our visits to the schools: Kathy Roberts, Nancy Underwood, Libbi Kirkland Sergent, Dilcia Reynolds Wahnsiedler, Micah Rumer, Kim Binzer, Bryan Quisenberry, Vicky Berger, Stacey Richardson, Greg Howell, Julie Jones, Sadie Jackson, Melodie Vereen, and Tasha Howard. We are extremely indebted to Lori Bowen, who serves as the director of science education in the elementary schools of Fayette County. Lori has been generous with her time and unfailingly enthusiastic in her support of our research efforts. Finally, we thank Peggy Keller for her advice on statistical analyses.

Correspondence concerning this article should be addressed to Robert F. Lorch Jr., Department of Psychology, University of Kentucky, Lexington, KY 40506-0044. E-mail: rlorch@email.uky.edu

Zohar & Peled, 2008) has yielded teaching interventions that produce rapid gains in CVS understanding by elementary and middle-school students. Further, the gains are long lasting and generally transfer well to tasks beyond the immediate teaching environment. However, most assessments of the two teaching interventions have involved individual instruction rather than classroom teaching. Thus, the second issue addressed in our study is how well Chen and Klahr's (1999) basic intervention "scales up" to the classroom (Ball & Forzani, 2007; McDonald, Keesler, Kauffman, & Schneider, 2006). Of particular interest is whether the intervention is effective in different school environments.

Contributions of Explicit Instruction and Experimentation to Learning CVS

Our theoretical focus is on the potentially separate contributions of explicit instruction and opportunities for experimentation to learning CVS. This issue may be placed in the context of a recurring debate about the importance of guidance during instruction (Dean & Kuhn, 2007; Kirschner, Sweller, & Clark, 2006; Klahr, 2005; Kuhn, 2005; Kuhn & Dean, 2005; Mayer, 2004). Many educational researchers have argued for an approach to science education based on "discovery" in which students are encouraged to reconstruct the procedures and domain knowledge of science through their own explorations of a domain (i.e., with minimal guidance from the teacher). According to this view, students achieve a deeper and more enduring understanding of the processes, logic, theory, and findings of science if they engage in the types of inquiry activities that characterize the practice of science (Bruner, 1961; Kuhn, 1989). In learning CVS, this approach has often taken the form of presenting students with opportunities to design and to conduct experiments. Students are not given explicit instruction or feedback, but they are presented with a structured environment where the relevant variables are specified, and the variable to be tested may or may not be specified (Chen & Klahr, 1999; Kuhn, Black, Keselman, & Kaplan, 2000; Kuhn & Dean, 2005; Kuhn et al., 1995; Zohar & David, 2008; Zohar & Peled, 2008). For example, students might construct conditions in a computer simulation to study factors affecting plant growth. They compare the outcomes of different simulations to identify the effects of a limited number of specified variables. They are not given explicit instruction in how to construct valid experimental comparisons; rather, it is expected that they will gradually develop the appropriate logic through their experimentation. Indeed, when provided with multiple opportunities for experimentation over time, elementary school students generally show gradual improvement in their ability to construct valid comparisons (Dean & Kuhn, 2007; Kuhn & Angelev, 1976; Kuhn, Schauble, & Garcia-Mila, 1992; Schauble, 1996; Schauble et al., 1991). However, experimentation without feedback produces little immediate improvement in understanding of CVS (Chen & Klahr, 1999; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008; Zohar & David, 2008; Zohar & Peled, 2008). These results are in sharp contrast to the rapid and enduring gains in CVS understanding that have been consistently observed for interventions that combine opportunities for experimentation with teacher-guided instruction in CVS (Chen & Klahr, 1999; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008; Toth et al., 2000; Triona & Klahr, 2003; Zohar & Aharon-Kravetsky, 2005; Zohar & David, 2008; Zohar & Peled, 2008).

The findings from studies on CVS learning are consistent with recent arguments that explicit guidance is critical to efficient, effective learning (Mayer, 2004; Kirschner et al., 2007). In the specific context of CVS, Chen and Klahr (1999) have argued that the relative ineffectiveness of discovery-based experimentation is due to the fact that the outcome of an experiment does not provide any information about the validity of the experimental design; that is, any experiment will have an outcome, but that outcome sheds no light on whether the experimental comparison was valid. The constraints typically placed on experimentation may aid some students in discovering CVS (Kuhn et al., 2000; Kuhn & Dean, 2005), but learning is much faster if students are systematically guided through the logic and/or the logic is explicitly presented for them.

Despite findings that discovery-based experimentation alone is a relatively ineffective method of CVS instruction, it may be an important component of the teaching interventions developed by Klahr and by Zohar. Both teaching protocols combine teacher-guided, explicit instruction of CVS with opportunities for exploratory experimentation by students. Although Chen and Klahr (1999) have argued that explicit CVS instruction is critical to the success of their intervention, it may be that the opportunity to conduct experiments also contributes importantly to its success. In fact, the combination of explicit instruction with student experimentation may be fundamental to the success of both Klahr's and Zohar's interventions. This suggestion was recently made by Zohar and David (2008). They hypothesized that the *linguistic component* of their intervention (i.e., explicit instruction based on verbal discussion) is critical to the development of metastrategic knowledge associated with CVS understanding. In addition, they hypothesized that opportunities to apply CVS in experiments are critical to the acquisition of the thinking strategy because the experience makes concrete what would otherwise remain abstract knowledge. This hypothesis is supported by Glenberg's findings of large gains in memory and comprehension of a narrative when elementary students engage in hands-on manipulation of materials that are the subject of the narrative (Glenberg, Gutierrez, Levin, Japuntich, & Kaschak, 2004; Glenberg & Kaschak, 2002) or of science instruction (Glenberg & Levin, 2006).

In the current study, we take up Zohar and David's (2008) suggestion that explicit instruction and opportunities for exploratory, hands-on experimentation both contribute importantly to CVS instruction. We base our test on Chen and Klahr's (1999) instructional intervention rather than Zohar's because Chen and Klahr's protocol more clearly separates the two components: The experimentation phase in Chen and Klahr's procedure does not involve any teacher guidance beyond specification of the goal of each experiment and the relevant variables in the domain; in contrast, the experimentation phase in Zohar's intervention is used as an additional opportunity for instructional guidance.

We compared three teaching interventions in 36 fourth-grade classrooms. One condition consisted of a replication of Toth et al.'s (2000) classroom adaptation of Chen and Klahr's (1999) intervention. The intervention included experimentation as a pretest and as a posttest of students' understanding of CVS. Between the two phases of experimentation, the intervention also included a short teaching protocol in which the teacher presented students

with examples of both confounded and unconfounded experiments using the same experimental apparatus as the pre- and posttests. The teacher used questions and discussion to reveal the characteristics of a valid experimental comparison, and the lesson concluded with an explicit statement of the CVS. Thus, this condition included both explicit, teacher-guided instruction and opportunities for experimentation by students (both condition).

A second condition omitted the teaching protocol, replacing it with additional opportunities for students to conduct experiments. Thus, this condition involved only hands-on experimentation with minimal guidance and no explicit instruction (manipulate condition).

The third condition omitted the pretest and posttest experimentation and presented students with only the teacher-guided instruction in CVS (instruct condition). Because the omitted experimentation component was not replaced with another activity, students in the instruct condition spent less time doing domain-relevant activities than students in the other two conditions. This confounding of instructional condition with time doing relevant activities was a deliberate choice rather than a design oversight because the options for eliminating the confounding would have qualitatively changed the research question of interest. The most straightforward way to equate time on CVS-relevant activity for the instruct condition would have been to add more examples to the explicit instruction component. Such a change would have required approximately tripling the amount of time in explicit instruction with the likely consequence that students would have become bored by the extensive repetition. An alternative to simply extending the interactive lecture component might have been to add a lecture that communicated domain knowledge without including additional instruction in CVS (cf. Zohar & David, 2008). If we had implemented either of these options, comparison of the instruct condition with the both condition would no longer isolate the contribution of the experimentation component. Rather, by the first option, the question would change to whether it is more effective to present extensive explicit instruction or to combine a shorter lesson of explicit instruction with opportunities for experimentation. By the second option, the question would change to whether it is more effective to combine explicit instruction with a lecture on the domain or with hands-on experimentation. These were not the questions in which we were interested.

Students in all three conditions were given a pretest, immediate posttest, and delayed posttest of their abilities to evaluate the validity of experimental comparisons in a variety of domains. On the basis of previous findings using similar procedures (Chen & Klahr, 1999; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008; Zohar & David, 2008; Zohar & Peled, 2008), we expected much greater learning gains from the pretest to the posttest in the both condition than in the manipulate condition. The observation of interest is whether explicit instruction and experience conducting experiments in the both condition make distinct contributions to learning of CVS. It is possible that teacher-guided, explicit instruction is sufficient for learning CVS, and experimentation adds little to students' understanding. In that event, performance in the instruct and both conditions should be equivalent. However, if Zohar's hypothesis is correct that explicit instruction and experimentation facilitate different components of the learning of CVS, then more learning should occur in the both condition than in the instruct condition.

Scaling Up the Teaching Intervention for the Classroom

The second goal of the study is to evaluate how well Chen and Klahr's (1999) teaching intervention translates into the classroom. Training studies demonstrate that instruction in CVS can be effective as early as the fourth grade, but there are serious obstacles to translating most laboratory methods into classroom practice. Most training studies have involved many sessions of minimally guided, individual experimentation over a period of several weeks. Thus, the instructional methods are too labor intensive and time intensive to be practical in the classroom if their sole purpose is teaching of CVS (Ross, 1988). The work of Klahr and his associates is a notable exception to these shortcomings (Chen & Klahr, 1999; Toth et al., 2000).

Chen and Klahr's (1999) teaching protocol is promising as a classroom intervention because it is very effective in one-on-one instruction in the laboratory, it can be adapted in a straightforward way to a classroom environment (Toth et al., 2000), and it is brief. Initial forays into the classroom have been quite successful, but they have been limited in some important respects. Two studies have demonstrated the feasibility of conducting the intervention in the classroom, but both lacked an appropriate control condition to evaluate the effectiveness of the intervention (Li et al., 2006; Toth et al., 2000). Most studies have included only relatively high-achieving classrooms (Klahr & Nigam, 2004; Toth et al., 2000). In the one study that focused on lower achieving classrooms (Li et al., 2006), the intervention was changed to bring the classroom to near-mastery levels of understanding of CVS before evaluation of learning occurred. This necessitated prolonging the teaching intervention, reducing its practicality as a classroom intervention. All of the studies from Klahr's laboratory have been limited to relatively few classrooms and relatively few students. Finally, Zohar and David (2008) conducted a study that addresses some of these limitations. It involved classroom instruction of both high and low achieving students, and it included an appropriate control condition. However, the classroom instruction was supplemented with relatively extensive individualized instruction and was relatively small in scale (i.e., 119 students in six classes at one school).

Our study addressed the shortcomings of previous studies in several ways. We included three, systematically manipulated variants of the teaching protocol so that we would have appropriate comparisons for evaluating the effectiveness of the basic intervention. We conducted instruction in 36 fourth-grade classrooms to evaluate the intervention when applied on a relatively large scale. Furthermore, to assess the robustness of the teaching intervention, we conducted instruction in schools that differed in average science achievement levels. We emphasize that the higher and lower achieving schools in our sample differ in many respects. Compared with students attending lower achieving schools, students attending higher achieving schools, on average, do better on standardized tests in other domains, come from economically more advantaged families, are racially less diverse, and spend more class time on the subject of science. Thus, the comparison allows us to assess the efficacy of the basic teaching intervention in different learning environments.

Method

Participants

Participants were fourth-grade students matriculating in public schools in Fayette County, Kentucky. A total of 36 classrooms from 12 schools participated in the study. Half of the classes were from five schools that in the preceding school year had achieved the highest average scores in the district on the science section of the state-mandated Kentucky Core Content Test (KCCT). The other half of the classes were from seven of the eight schools with the lowest average scores in the district on the same test.[1] A total of 797 students completed participation on at least one of the first 3 days of the procedure; 420 students attended higher achieving schools, and 377 attended lower achieving schools.

Not surprisingly, the higher and lower achieving schools differed in many ways besides their scores on the science section of the KCCT. Some of those differences are illustrated in Table 1, which shows that the schools differed in achievement in other domains, in diversity and the proportion of students on free or reduced lunch, and in enrollments. The schools also tended to differ in their support of science education, with the higher achieving schools generally devoting more instruction time to science than the lower achieving schools. In short, the comparison of higher versus lower achieving schools should be understood as a comparison of learning environments that are distinct with respect to the challenges they present for teaching science.

Design

Classes were drawn from schools at two levels of achievement (high vs. low) on the KCCT in science and were randomly assigned to one of three instructional conditions (instruct, manipulate, both). Steps were taken to balance possible influences of teacher assignment and school on instructional condition. Within any school that had three fourth-grade classrooms, the three classrooms were treated as members of the same *triplet*. Because not all schools had exactly three classrooms, triplets sometimes had to be formed from classrooms from two different schools. When that was necessary, classrooms with similar characteristics were assigned to the same triplet. Sets of classrooms that were members of the same triplet were all taught by the same instructor and received the same assignment of materials to test position. The classrooms in a triplet were assigned at random to instructional conditions. Thus, each triplet constituted a replication of the basic experimental design at a given level of school achievement.

Data for the main measures in the experiment were collected from individual students. However, students worked in small groups when experimenting, nesting those students within groups. All students and groups were, in turn, nested in classrooms, and classrooms were nested within triplets/replications at two achievement levels.

Materials

Materials for the experiment included sets of ramps and balls as well as two assessment instruments adapted from Chen and Klahr (1999).

Ramps. Ramps similar to those used by Chen and Klahr (1999) were constructed to be used by the students for experiments. Balls were rolled down a *down ramp* onto an *up ramp* containing a set of numbered steps to stop the balls and to provide a metric for the distance the balls rolled. Experiments using the ramps could vary with respect to four binary factors. A ramp could be set at one of two levels of steepness by manipulating the orientation of a support block. The surface of the down ramp could be rough or smooth, depending on which side of a carpet insert was placed face up in the ramp. Two alternative starting points were marked on the down ramp for rolling the ball. In addition, either a new yellow ball or an old white ball could be rolled during a trial.

Assessments. Two instruments were created to assess students' understanding of CVS. The assessment instruments were very similar to those used by Toth et al. (2000). The first was a paper-and-pencil *comparison test* of students' ability to evaluate the validity of experimental comparisons. The test was composed of five items from each of three domains (e.g., baking cookies, exercise, growing plants). Each domain was introduced with a page that described a question of interest (e.g., how different ways of baking might affect the taste of a cookie) and introduced three variables to be investigated (i.e., whether the cookies were baked for 5 or 10 min, whether they were sweetened with sugar or honey, and whether they used one or three eggs). The subsequent five pages in a domain each presented a comparison of two experimental conditions that could vary on any combination of three variables. For example, each of the five items in the baking cookies domain depicted a comparison of two ways in which a batch of cookies might be baked. A comparison was depicted with pictures of the values of the three variables and labels for those values. For each item, the required response was to circle whether the comparison was a *good* (i.e., valid) or *bad* (i.e., invalid) test of the effects of a specific variable.

Within each domain, two of the comparisons were valid (40%); the three invalid comparisons (60%) were composed of one doubly confounded comparison (the two pictures had different values on all three variables), one singly confounded comparison (the two pictures had different values on two variables), and one noncontrastive comparison (the two pictures differed on only one variable, but it was not the variable being tested). Because each student was tested at three different times, three versions of the comparison test were created, all with the same structure but composed of different domains. The assignment of test versions was counterbalanced across the three testing times. In addition, for each test, we created three different orderings of the last four comparisons within each domain. This was to discourage students from copying from one another during the group administration of the test.

The second instrument consisted of three booklets to record students' work designing and running experiments using the ramps. The pretest booklet, used in the manipulate and both conditions, had a cover page to record the student's name and the

[1] The lowest scoring school in the district was excluded from the experiment. We did include one class from the school in the initial 3 days of testing, but student behavior problems precluded successful administration of the procedure. The disruptions were likely due to the fact that the regular classroom teacher was absent all 3 days, and there was a different substitute teacher on each day.

Table 1
Characteristics of Participating Schools Based on Available Data in the 2005–2006 Academic Year

School	KCCT science[a]	KCCT math[a]	KCCT English[a]	Free or reduced lunch (%)	Caucasian students (%)	Total enrollment
High #1	117.6	120.9	107.1	8	81	691
High #2	112.6	123.1	112.0	11	87	273
High #3	105.0	108.8	101.8	12	76	526
High #4	104.4	98.9	98.8	29	72	581
High #5	100.1	116.5	107.7	2	87	717
Low #1	76.7	72.6	75.8	87	53	392
Low #2	76.0	52.2	61.0	90	28	262
Low #3	75.4	79.8	75.9	45	60	666
Low #4	72.1	57.5	68.0	63	41	416
Low #5	66.7	66.2	70.4	89	29	546
Low #6	65.7	67.0	61.9	87	42	284
Low #7	63.8	90.5	73.1	93	23	230

Note. KCCT = Kentucky Core Content Test; High = higher achieving schools; Low = lower achieving schools.
[a] Proficiency on each KCCT is a score of 100; maximum score is 140.

classroom teacher, followed by four pages with the same basic layout. At the top of the page was a question identifying the focal variable that students were to test. Below the question was a table that students were to use to plan their experimental design. The table listed the four variables and provided students with a choice of two values for each variable for each of their two ramps. Below the planning table was a second table to record how far the ball rolled on each of three trials for both ramps. Finally, students were asked to identify on which ramp the ball usually rolled further, to make a conclusion about whether the focal variable had an effect, and to judge their confidence in the conclusion. The booklet had pages for two tests of both *length of run* and *type of surface*. Paralleling the pretest booklet was a posttest booklet with the identical format, also used in the manipulate and both conditions. The only difference between the two tests was the selection of focal variables. All four focal variables were represented once in the posttest (i.e., steepness, type of ball, length of run, surface).

Finally, a third booklet was constructed for the manipulate condition only. Each page of this booklet was identical to the pretest and posttest except for two changes. First, no focal variable was identified at the top of the page; instead, a space was provided for students to indicate the goal of their experiment. Second, Question 5 (requesting a conclusion with respect to the effect of the focal variable) was replaced with a question asking students to respond to the open-ended question: "What did you decide from your experiment?"

Procedure

Classrooms were visited four times throughout the school year. Visits were scheduled on three consecutive days during the fall semester. On Day 1, any pretests were administered. On Day 2, depending upon condition, the instructional intervention was conducted, and ramps experiments were conducted by groups of students. On Day 3, students were given the comparison test. The fourth visit in the latter half of the spring semester was to administer another comparison test. All instruction and testing were administered to the class as a group.

The same two female graduate assistants did all of the instruction in all of the conditions. A given instructor did all the teaching to classes assigned to the same triplet, so each instructor taught each of the three instructional conditions the same number of times. One instructor taught eight triplets of 24 classes; the other instructor taught the remaining four triplets of 12 classes. In addition to the instructor, each classroom had one or two helpers to distribute/pick up materials and to help answer students' procedural questions when they conducted experiments.

Both condition. The both condition incorporated almost all of the components of the instruct and manipulate conditions, so it is described in detail.

Day 1. The 12 classrooms assigned to the both condition received two pretests on Day 1 of the procedure. The first was the comparison pretest. The instructor always read the first two pages of each item domain and then let students do the last four items in the domain on their own. The test took about 20 min to complete. Performance was scored for the total number of correctly answered questions (maximum = 15).

Following the comparison pretest, the students were assembled into groups of three to five. Each group was assigned a set of ramps, and one ramps pretest booklet was given to each group. A pair of ramps was used to identify the four variables and to illustrate how to manipulate each one. The instructor then gave an overview of the steps that each group was to follow in planning and conducting experiments. To aid in explaining the procedure, the instructor used an overhead projector to display a transparency of the first page of the ramps pretest. She showed the students how to record their experimental design in the planning table and how to record the results of the experiment on the second table on the page. Once the students appeared to understand the procedure, they began working on the first experiment to test the length of run variable.

Each group discussed how to set up their ramps to test the length of run variable. When they had recorded their plan, the instructor or a helper checked that the group had recorded a value for each variable and then gave them permission to conduct the experiment. No feedback was given regarding the validity of the plan. The students ran three trials, recording the distance the ball rolled on each ramp on each trial. When the experiment was complete, the group answered the questions about conclusions from the experiment.

When a group finished the first experiment, they designed and executed an experiment to test the surface variable. Following that, they did another experiment to test length of run, then a final experiment to test surface. Including instruction, the ramps pretest phase required between 30 and 45 min to complete. The entire procedure on Day 1 required from 50 min to 75 min to complete. Some classrooms had the flexibility to continue for 75 min, if necessary; in other classrooms, no more than 50 min could be allotted. Because of variability across classes in both the time available and the efficiency with which groups worked, not all groups completed all four pretests. However, the majority of groups (68%) did complete all pretests. There was no difference between the manipulate ($M = 3.44$) and both ($M = 3.55$) conditions in the number of pretests completed, although there was a tendency for students from the lower achieving schools to complete more pretests ($M = 3.77$) than students from higher achieving schools ($M = 3.28$), $t(22) = -1.938$, $SE = 200$, $p = .064$. In each of the four combinations of instruction and achievement level, the median number of pretests completed was 4. Thus, all students had extensive opportunity to conduct experiments.

Day 2. The procedure on Day 2 required approximately 50 min and consisted of two parts. First, the instructor conducted a lesson on the CVS. The lesson began by the instructor introducing a proposal for a ramps experiment to test whether the variable of steepness had an effect on how far the ball rolled. The proposed comparison of ramps was completely confounded (i.e., differed on all four variables). The students were asked to evaluate whether the experiment was a good test of steepness and to explain why or why not. The instructor then asked the class to compare the two ramps on each of the four variables. When they determined that the ramps differed on every possible variable, the instructor then asked them when the ball rolled further on one of the ramps, what might be the cause? This discussion culminated in the point that any of the four differences might cause a difference, so they could not tell from the results of the experiment whether steepness had an effect on how far the ball rolled. In turn, that meant that if the students wanted to determine for certain whether steepness had an effect, they must allow the ramps to differ only in steepness. The instructor then had the students tell her how the ramps should be set up to test steepness, and she concluded the exercise by stating the properties of a good test of steepness.

Following the first example, the instructor repeated the same sequence of instruction with a new experimental example. The second example was also an invalid experiment, although there was only a single confound in the design. After the second example, the instructor explicitly stated the CVS. The lesson took 15–20 min. Instruction was videotaped for later evaluation of the integrity of the delivery of the teaching protocol.

After the lesson, students assembled in their groups from the previous day and again did experiments with the ramps. The procedure was identical to that on Day 1. The focal variables in the four experiments were tested in the following order: steepness, type of ball, length of run, and surface. Most groups completed all four tests (72%), and no group completed fewer than three tests. There were no differences in the number of completed tests as a function of instructional condition or achievement level (mean number of experiments completed ranged from 3.63 to 3.86).

Day 3. The procedure on Day 3 consisted solely of administration of the comparison test. The test was identical in format to the corresponding test on Day 1, but the domains tested were different. It took approximately 20 min to administer the test.

Day 4. The Day 4 procedure was identical to that of Day 3.

Instruct condition. The procedure for the instruct condition was identical to that of the both condition, except that the students in the instruct condition never conducted experiments using the ramps. This difference in procedure led to the following specific changes relative to the both condition. On Day 1, there was no instruction in conducting ramps experiments and no group experiments using the ramps. On Day 2, the class first received an introduction to the ramps and the ways in which they could be manipulated, which had been covered on Day 1 in the both condition. Then the class received the identical lesson in the CVS that the both condition received, including the demonstrations and discussions of confounded and unconfounded experiments. Classes in the instruct condition did not do ramps experiments following the CVS lesson on Day 2. Thus, the procedure in the instruct condition was shorter on Day 1 (approximately 25 min) and Day 2 (approximately 30 min) than in the both condition. The procedure on Days 3 and 4 was identical to that described for the both condition.

Manipulate condition. The procedure of the manipulate condition was identical to the both condition, with one exception. On Day 2, classes in the manipulate condition did not receive the CVS lesson. In its place, the students were reassembled into their groups and were told to do whatever experiments they wished to try to learn how the four variables affected how far the balls rolled. The groups conducted their experiments for approximately 15–20 min (i.e., the duration of the CVS lesson in the other two conditions), recording what they did for each experiment. Following the period of open-ended investigation of the ramps, the students were then given the same ramps posttest to perform as students in the both condition.

Results

The videotapes of instruction in the both and instruct conditions were coded with respect to the 21 key components of the teaching protocol. Coding was done by three assistants who were unaware of the condition they were coding. A random sample of eight videotapes was independently scored by two raters. Of the 168 judgments, there were only three disagreements (98.2% agreement). With respect to the integrity of the treatment, there were 12 omissions from the teaching protocol across 460 opportunities (2.6%). In these few instances, the relevant information typically was communicated implicitly by the instructor. In short, treatment integrity was very good.

Comparison Tests

The primary measures of performance were the scores of the individual participants on the three comparison tests (pretest, immediate posttest, delayed posttest) of a student's ability to distinguish valid and invalid experimental comparisons. Each test was scored for the total number of correct responses on the basis of 15 items. Of the 797 students who participated in at least one of the first 3 days of the procedure, the data of 75 students (17.86%) from higher achieving schools and six students (1.59%) from lower achieving schools were excluded from all analyses because they

demonstrated on the pretest that they already knew the CVS. We adopted Chen and Klahr's (1999) benchmark of 13 or more correct responses (>85%) as our criterion for exclusion on the basis of pretest score. The analyses reported below are based on the data of 371 students from lower achieving schools and 345 students from higher achieving schools.

Analytical strategy. Because of their nested structure, we analyzed the data using multilevel modeling procedures (Raudenbush & Bryk, 2002); the software used was HLM 6.04 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004). The data had five levels of structure: repeated measures on each student, students, work groups, classrooms, and replications (roughly equivalent to schools). However, the data were analyzed with a three-level model that ignored work groups and replications. The group level was ignored because preliminary analyses showed no significant random variability at the group level. There was significant variability at the replications level of the model; however, with only six replications at each level of achievement, there was insufficient power to test possible interactions of achievement level with instructional condition. Therefore, achievement was coded at the classroom level.

The first level of the model consisted of the *repeated measures* on each student. The three tests were coded by two dummy variables: The prepost variable was coded 1 for the pretest and 0 for the other two tests; the postdelay variable was coded 1 for the delayed test and 0 for the other two tests. Note that the coding of these two variables establishes the immediate posttest as the common point of comparison for the pretest and delayed posttest, respectively.

There were no variables coded at the second, *subject level* of the model. However, this level was needed to partition the variability into within- and between-subjects variability.

The third level of the model was *classrooms*. The three instructional conditions were dummy coded at this level: The variable of manip was coded 1 for the manipulate condition and 0 for the other two conditions; the variable of both was coded 1 for the both condition and 0 for the other two conditions. This coding scheme makes the instruct condition the common point of comparison for the manipulate and both conditions, respectively. Finally, school achievement was coded at the classroom level (0 for lower achieving, 1 for higher achieving).

Number of correct responses. Table 2 summarizes the analysis of the data (Singer & Willett, 2003), and Figure 1 graphs the mean number correct (maximum = 15; chance = 7.5) as a function of test and condition. All reported tests are significant beyond the .05 level unless noted otherwise (Cohen, 1990).

Averaging across instructional conditions and school achievement (see Model 2 in Table 2), the mean number correct on the pretest was 7.495, which did not differ from chance performance of 7.5, $t(35) < 1$, *ns*. The slope value for the prepost variable was significant, $t(35) = -8.811$. The negative value of the slope means that performance was better by 2.088 points on the immediate posttest (coded 0) than on the pretest (coded 1). The negligible drop of .006 points from the immediate to the delayed posttest was not significant, $t(35) < 1$.

Because there were no indications of any differences between the immediate and delayed posttests and no significant residual variance associated with the postdelay variable (see Model 3 in Table 2), postdelay was dropped from the model. Looking at the results for prepost for the final model (see Model 4 in Table 2), it can be seen that the improvement in performance from the pretest to the combined posttests was affected by both the type of instruction and achievement level. There was less learning in the manipulate condition than in the instruct condition, $t(32) = 2.285$, and

Table 2
Parameter Fits (and Standard Errors) of Successive Models for the Comparison Tests

Model type	Variable	Parameter	Model 1	Model 2	Model 3	Model 4
Fixed effects						
Initial status	Intercept	γ_{000}	8.874** (0.185)	9.583** (0.262)	8.671** (0.260)	8.644** (0.227)
	Manip	γ_{001}			−1.303** (0.319)	−1.136** (0.318)
	Both	γ_{002}			1.049** (0.325)	0.949** (0.315)
	Achieve	γ_{003}			1.988** (0.253)	1.985** (0.247)
Rate of change						
Prepost	Intercept	γ_{100}		−2.088** (0.237)	−1.310** (0.244)	−1.306** (0.347)
	Manip	γ_{101}			1.022** (0.320)	0.874* (0.383)
	Both	γ_{102}			−1.007* (0.369)	−0.882* (0.425)
	Achieve	γ_{103}			−1.562** (0.286)	−1.552** (0.315)
Postdelay	Intercept	γ_{200}		−0.006 (0.151)	−0.060 (0.394)	
	Manip	γ_{201}			0.340 (0.410)	
	Both	γ_{202}			−0.199 (0.341)	
	Achieve	γ_{203}			0.014 (0.283)	
Variances						
Level 1	Within-Ss	σ^2_e	7.421	5.347	5.344	5.541
Level 2	Intercept	σ^2_0	1.890**	2.547**	2.537**	2.539**
Level 3	Intercept	σ^2_0	0.991**	2.008**	0.140*	0.244**
	Prepost	σ^2_1		1.446**	0.221*	0.433**
	Postdelay	σ^2_2		0.187	0.200	
Deviation			9,941.091	9,549.519	9,492.562	9,502.063
df			4	11	20	13

Note. These models predict number of correct responses on the comparison tests. All variables are dummy coded, as explained in the text.
* $p < .05$. ** $p < .01$.

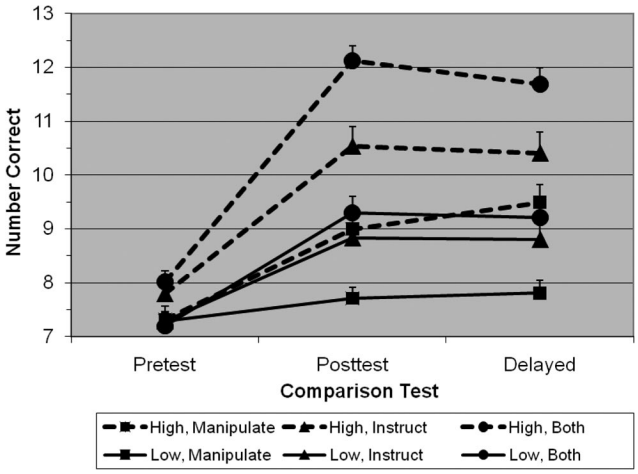

Figure 1. Mean number correct on the three comparison tests as a function of instructional condition and school achievement.

less learning in the instruct condition than in the both condition, $t(32) = -2.082$. In addition, learning was greater in the higher achieving schools than in the lower achieving schools, $t(32) = -4.935$. These effects on amount of learning were reflected in the mean performance levels on the posttests. Performance was lower in the manipulate condition than in the instruct condition, $t(32) = -3.578$, which, in turn, was lower than performance in the both condition, $t(32) = 3.007$. Finally, students from higher achieving schools outperformed students from lower achieving schools, $t(32) = 8.033$. Using Model 2 as a baseline, the instructional manipulation and school achievement accounted for 84.6% of the variance at the classroom level of Model 4.

Figure 1 shows the same ordering of means across instructional conditions for students from higher and lower achieving schools, but students from higher achieving schools appear to benefit more from the opportunity to conduct experiments with the ramps. In fact, when the interaction of achievement with instruction was tested against subject variability, the effect approached significance, $F(2, 671) = 2.850$, $p = .059$. Further, separate analyses of the data for the two levels of school achievement showed that experience doing experiments consistently improved the overall performance of students from higher achieving schools: Performance was above chance in the manipulate condition, $t(15) = 3.772$, and performance was higher in the both condition than in the instruct condition, $t(15) = 2.982$. In contrast, students from lower achieving schools did not show reliable benefits from doing experiments: Performance did not differ from chance for the manipulate condition, $t(15) < 1$, and performance did not differ reliably in the instruct and both conditions, $t(15) = 1.701$, $p = .089$.

If it really is the case that students from higher achieving schools benefit from experimentation, whereas students from lower achieving schools do not, that finding would have important theoretical and practical implications. However, two observations argue against interpreting the results as showing that the two student populations are affected in qualitatively different ways by the instructional manipulations. First, the ordering of means across tests and instructional conditions is identical for the higher and lower achieving schools (see Figure 1). Second, additional analyses were conducted on the comparison posttests, breaking down performance by item type; items illustrated valid experimental comparisons, confounded comparisons, or comparisons that contrasted a variable other than the focal variable (i.e., noncontrastive). These analyses show clear and similar differences between the instruct and both conditions for both student populations.

Percentage correct on each of the three types of items was analyzed as a function of achievement and instruction; achievement and instruction were coded as in previous analyses. The relevant data are summarized in Figure 2. As already established, students from higher achieving schools performed better than students from lower achieving schools. The new finding is that instructional condition interacted with the type of item. As shown in Figure 2, the instruct and manipulate conditions differed only with respect to performance on items depicting confounded comparisons, $t(32) = -6.534$. The instruct and both conditions performed similarly on confounded comparisons, but performance on both valid and noncontrastive comparisons was better in the both condition, $t(32) = 2.559$ and $t(32) = 3.542$, respectively. To describe this pattern of results in another way, the explicit instruction of the instruct and both conditions produced much better performance on the confounded comparisons than the manipulate condition. Further, combining the opportunity to perform experiments with explicit instruction (i.e., both condition) produced better performance on valid and noncontrastive comparisons than explicit instruction alone (i.e., instruct condition). The pattern of interaction of instruction with item type did not differ for higher and lower achieving schools (smallest $p > .2$).

To summarize, students from higher achieving schools benefit more than students from lower achieving schools from all three types of instruction. However, the pattern of effects of the instructional manipulation is the same for both participant populations.

Expert levels of performance. Mastery of the CVS means that a student (a) knows to focus on the variable to be tested, (b) knows that different values of the focal variable must be compared, and (c) knows that the values of all other variables must be held constant. Are students acquiring the entire strategy or just a partial understanding? To address this question, we categorized students as either being "experts" or not with respect to CVS. Consistent with our criterion for exclusion on the basis of the pretest, we defined *expert performance* as ≥13 correct on the posttest (Chen

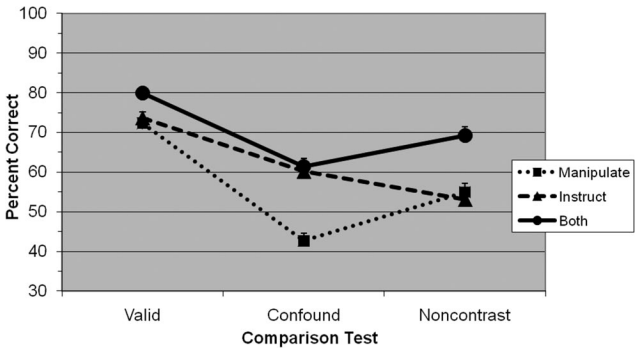

Figure 2. Mean percentage correct for three types of comparisons presented as a function of instructional condition.

& Klahr, 1999). Table 3 displays the proportions of students categorized as experts on the immediate and delayed posttests as a function of school achievement and instructional condition. The pattern of results for this measure mirrors the findings already reported for the comparison test scores computed across all students (see Figure 1). In fact, the pattern of significant effects for the proportion scores in Table 3 is identical to the effects reported in the preceding section for the mean number of correct responses.

These results raise the question of whether students who fall short of mastery levels are gaining even a partial understanding of CVS. To answer this question, we excluded the data of those students who attained mastery and analyzed the data of the remaining "nonexpert" students. Separate analyses were conducted on the data from the two achievement levels. For students from lower achieving schools, learning occurred but it was unimpressive. The number of correct responses on the comparison test increased by just 0.64 points from the pretest to the combined posttests, $t(835) = -4.072$. There were no effects of instructional condition on the amount of learning (all ps $\geq .407$). For students from higher achieving schools, the mean improvement from the pretest to the immediate posttest was 1.28 points—twice that of students from the lower achieving schools, $t(586) = -5.689$. Further, these students improved by an additional 0.78 points from the immediate posttest to the delayed posttest, $t(586) = 3.371$. However, like students from the lower achieving schools, there was no effect of instructional condition on the amount of learning observed (all ps $\geq .387$).

Integrating the results of the analyses of the experts with the analyses of learning by students who did not achieve expert status, the overall mean differences between the instructional conditions on the comparison tests are primarily due to differences in the proportions of students who mastered the CVS. This conclusion is based on the findings that the instructional manipulation had large effects on the proportions of students who attained expert levels of performance, but it had no reliable effects on the comparison test performances of the students who fell short of expert status.

Ramps Tests

In addition to individual scores on the comparison tests, students in the manipulate and both conditions worked in small groups to construct experiments with the ramps. As already noted, for reasons beyond their control, not all groups attempted all four ramps experiments (32.2% attempted fewer than four). Consequently, the dependent variable chosen for the analyses of the ramps data was the proportion of completed experiments that were valid experiments. Of the 159 groups conducting ramps experiments, 15 performed perfectly on the pretest, so their data were omitted from the analyses. In addition, the data for the ramps posttest were lost for one class (i.e., 7 groups). Thus, the final data set was composed of the data for 137 groups. We analyzed these data using a hierarchical linear model in which groups were nested within classes, and we coded achievement level and instructional condition at the classroom level. The coding of both factors was the same as in previous analyses.

Table 4 shows that the results for the group experiments are generally consistent with the findings for individual students on the comparison tests. First, there was no significant learning for groups from lower achieving schools in the manipulate condition, $t(21) < 1$. In contrast, groups from higher achieving schools did improve from pretest to posttest in the manipulate condition, $t(21) = 3.550$. Second, there was more learning overall in the both condition than in the manipulate condition, $t(21) = 6.848$. Finally, for each experimental design generated by a group, we observed whether the group (a) manipulated the focal variable and (b) controlled the nonfocal variables. Students in both conditions were more likely to manipulate the focal variable than to control the nonfocal variables, but the difference in performance on these two aspects of the design was greater in the manipulate condition ($M = 80.94\%$ on focal variable vs. $M = 55.56\%$ on nonfocal variables) than in the both condition ($M = 92.11\%$ and $M = 85.85\%$, respectively), $F(144) = 10.14$.

Discussion

Contributions of Explicit Instruction and Experimentation to Learning CVS

Students from higher achieving schools benefited more from every instructional condition than students from lower achieving schools, but the pattern of effects of instruction was very similar for the two populations of schools. The ordering of condition means with respect to overall performance on the comparison tests was identical. Likewise, the effect of instructional condition on the evaluation of valid, confounded, and noncontrastive comparisons was the same. Finally, comparison of the effects of the manipulate and both conditions on groups' designs of ramps experiments showed the same pattern for both school populations.

For students from both higher and lower achieving schools, learning was greater in the both condition than in the manipulate condition. This replicates findings of several previous studies (Chen & Klahr, 1999; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008; Zohar & David, 2008; Zohar & Peled, 2008) and demonstrates that combining explicit instruction with experimentation is much more effective than experimentation alone. How-

Table 3
Proportion of Students Scoring 13 or More Correct on the Comparison Posttests (Standard Errors in Parentheses)

	Lower achieving			Higher achieving		
Posttest	Manipulate	Instruct	Both	Manipulate	Instruct	Both
Immediate	.030 (.017)	.124 (.030)	.229 (.040)	.186 (.039)	.316 (.047)	.546 (.048)
Delayed	.058 (.025)	.127 (.032)	.227 (.043)	.250 (.044)	.359 (.050)	.477 (.049)
M	.044	.126	.228	.218	.338	.511

Table 4
Mean Proportion (and Standard Error) of Valid Ramps Experiments as a Function of Instructional Condition and School Achievement

Condition	Lower achieving	Higher achieving
Manipulate		
Pretest	.072 (.022)	.169 (.052)
Posttest	.130 (.036)	.419 (.067)
Difference	.058	.250
Both		
Pretest	.067 (.027)	.367 (.062)
Posttest	.512 (.075)	.880 (.037)
Difference	.445	.513

ever, our primary theoretical goal was to separate the contributions of explicit instruction and opportunities for experimentation to learning the CVS. Comparison of performance in the instruct condition with performance in the manipulate and both conditions demonstrates that explicit instruction and experimentation make distinguishable contributions to students' understanding of CVS.

Consistent with previous findings (Dean & Kuhn, 2007; Kuhn & Angelev, 1976; Kuhn et al., 1992; Schauble et al., 1991), there was unimpressive, but measurable, learning by students when they performed experiments of their own design with no explicit instruction and no feedback about the validity of their designs. Further, students who received both treatments outperformed those who only received explicit instruction. Most informative, however, is the pattern of performance on the valid, confounded, and noncontrastive items of the comparison tests. Figure 2 illustrates that explicit instruction selectively facilitated students' understanding of the need to control irrelevant variables. This is supported by the finding that in the two conditions involving explicit instruction (instruct and both) performance is much better on the confounded items than in the condition omitting such instruction (manipulate). Figure 2 also shows that compared with explicit instruction alone (instruct), combining experimentation and explicit instruction (both) results in selective improvement on the two item types that require attention to the focal variable (i.e., valid and noncontrastive items). This result suggests that conducting experiments serves to sharpen students' focus on the relevant variable and the need to manipulate it. This conclusion is consistent with findings from Kuhn's laboratory that students show better understanding of CVS if the experimenter designates a single variable to be tested in an experiment than if students are required to determine the goal of an experiment (Kuhn et al., 2000; Kuhn & Dean, 2005).

Our findings document distinct effects of the two types of instruction as implemented in our procedure. The effects we found for explicit instruction are generally consistent with Zohar and David's (2008) suggestion that such instruction provides students with an abstract understanding of CVS. It is particularly useful in helping students understand the necessity to control irrelevant variables (Chen & Klahr, 1999). The experience of conducting experiments helps consolidate understanding by providing concrete referents and an opportunity to actively implement the strategy (Glenberg et al., 2004; Glenberg & Kaschak, 2002; Glenberg & Levin, 2006). In particular, the hands-on experimentation seems to have caused the students to pay closer attention to the focal variable.

We conclude this section by acknowledging another possible account of the difference between the both and instruct conditions; namely, students in the both condition spent more time engaged in CVS-relevant activity than students in the instruct condition. We can rule out a simple version of the hypothesis that time engaged in relevant activity accounts for performance in our task. By that hypothesis, performance should have ordered as follows: both = manipulate > instruct. In fact, performance in the both and instruct conditions was better than performance in the manipulate condition. The confound also fails to explain why the both and instruct conditions produced different patterns of performance on the three item-types on the comparison posttests. The important point here is that what matters for learning is not time per se but how the time is used (Mayer, 2004; Rittle-Johnson, 2006). If we had added another CVS-relevant activity to the instruct condition to equate time, perhaps learning would have been equivalent to, or better than, the both condition. Such a result would be important as a demonstration that time spent conducting experiments is not necessarily the best use of instructional time. However, our question was not whether combining experimentation with teacher-led instruction was the most effective teaching intervention possible; rather, our question was simply whether combining experimentation with teacher-led instruction was more effective than explicit instruction alone.

Scaling Up the Teaching Intervention for the Classroom

An important motivation for the study was to examine how well Chen and Klahr's (1999) teaching intervention translates to the classroom. We distinguish three questions about how well the intervention scales up. First, is the intervention effective at producing immediate gains with respect to understanding CVS? Second, are the learning gains durable? Third, how does the intervention fare in different learning environments?

First, instruction in the both condition is relatively effective in promoting learning of CVS; averaging over the two school environments, performance rose from chance (50%) at pretest to 71.4% at the immediate posttest. The 80.5% level of performance by students in the higher achieving schools is very similar to performances observed by Chen and Klahr (1999) and Toth et al. (2000) with students of presumably similar abilities and educational advantages. For both the higher and lower achieving schools, most of the overall improvement in performance is attributable to the students who moved from chance performance on the pretest to mastery of CVS on the immediate posttest. In the lower achieving schools, 23% of the students achieved mastery; in the higher achieving schools, 55% achieved mastery.

Second, once learned, students maintained their understanding of CVS throughout the school year. The levels of performance on the delayed tests were strikingly similar to those on the immediate posttests, despite the fact that the delayed tests were conducted 4–5 months after instruction (Chen & Klahr, 1999; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008; Toth et al., 2000; Triona & Klahr, 2003). In short, children who achieved immediate mastery of CVS maintained their understanding.

With respect to the third question, students from lower achieving schools benefited less than students from higher achieving schools from all three forms of instruction. What accounts for the

substantial differences in overall learning by the two populations of students?

One difference between our two school populations is that higher achieving schools generally devoted more instructional time and resources to science. In contrast, a greater proportion of instructional time in the lower achieving schools was devoted to reading and math. As a result, students in higher achieving schools were probably better prepared for instruction in CVS and better equipped to handle the relatively unstructured learning environment represented by group experimentation with the ramps (Kirschner et al., 2006). This difference in preparation was evident in the difference in the numbers of students who demonstrated mastery of CVS on the comparison pretest: 75 of 420 students (17.9%) in the higher achieving schools compared with only six of 383 (1.6%) students from lower achieving schools. In addition, we conducted a post hoc analysis of the comparison pretest scores for the students who fell short of mastery, breaking down performance for the three types of items on the test. This analysis demonstrated different patterns for our two school populations. Students from higher achieving schools had a much higher percentage correct on pretest items depicting valid (67.28%) and noncontrastive (54.17%) comparisons than confounded comparisons (35.43%). This pattern indicates that at least some students were aware before the start of instruction that one of the variables is the focus of the comparison and that its values must differ. In contrast, students from the lower achieving schools showed a different pattern of results on the comparison pretest. They also scored best on valid comparisons (59.47%), but they performed much more poorly on both confounded (40.14%) and noncontrastive (44.04%) comparisons. This pattern suggests that students from lower achieving schools entered instruction with a general bias to judge a comparison as "good"; there is little indication of awareness of the focal variable. In short, students in the higher achieving schools had a more advanced understanding of what constituted a valid experimental comparison before they began our study.

In addition to different levels of background knowledge in science, students from the two school populations probably differed in other important ways. Our observations of science fair projects in lower and higher achieving schools suggested different levels of support of science education by the families of students in the two school populations. The students in the two populations of schools may have also differed in their confidence and motivation to achieve in the domain of science. Surely there are other differences. Some of these potential contributing factors are not ones that can be addressed in a brief teaching intervention, but others might be surmounted by changes in the intervention. Our future research goals concentrate on ways the basic intervention might be modified to achieve better results in lower achieving schools. In particular, we are interested in whether these students might benefit from modifications that lead to greater engagement in the topic. One method that might be effective is to have students conduct experiments individually rather than in groups, perhaps stimulating students to take greater ownership of the experiments. Another possibility is to modify the teaching protocol so that each student must respond to the instructor's questions about the experimental examples (e.g., by holding up a sign).

We conclude by acknowledging some important limits on the generalizability of our findings and conclusions. There is a big difference between designing an experiment in our procedure and designing an experiment for a fourth-grade science fair. Students in our procedure did not select the domain or the question for investigation, and the domain was much more highly structured in our procedure than in the science fair example. We specified the focal variable for students, and there is evidence that simple act in itself helps students to create valid experiments (Kuhn et al., 2000; Kuhn & Dean, 2005). In contrast, designing an experiment from scratch requires students to understand that they should focus on one factor at a time and choose a variable for manipulation that is appropriate to the question under investigation. We also specified all of the variables in the domain and their values. In addition, we limited the number of variables and restricted them all to have binary values. In a realistic experiment, there are many more variables, often with multiple or continuous values. All of these considerations acknowledge that there are substantial challenges to creating valid experiments that are not addressed in our procedure. A child who demonstrates perfect understanding in CVS may then create a science fair project that is confounded. Therefore, important questions for future investigations concern identifying components of instruction that help children to transfer basic understanding to complex situations. However, a child who learns CVS in the constrained environment we and many others have studied (e.g., Chen & Klahr, 1999; Kuhn et al., 2000; Zohar & David, 2008) has taken a significant step toward understanding the complexities of designing and interpreting experiments under less constrained circumstances (Klahr & Nigam, 2004).

References

Ball, D. L., & Forzani, F. M. (2007). What makes education research "educational"? *Educational Researcher, 36,* 529–540.

Bruner, J. S. (1961). The art of discovery. *Harvard Educational Review, 31,* 21–32.

Bullock, M., & Ziegler, A. (1999). Scientific reasoning: Development and individual differences. In F. E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich Longitudinal Study* (pp. 309–336). Munich, Germany: Max Plank Institute for Psychological Research.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

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

Dunbar, K., & Klahr, D. (1989). Developmental differences in scientific discovery processes. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 109–143). Hillsdale, NJ: Erlbaum.

Glenberg, A. M., Gutierrez, T., Levin, J. R., Japuntich, S., & Kaschak, M. P. (2004). Activity and imagined activity can enhance young children's reading comprehension. *Journal of Educational Psychology, 96,* 424–436.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9,* 558–565.

Glenberg, A. M., & Levin, J. R. (2006, November). *Embodiment in education.* Paper presented at the 47th Annual Meeting of the Psychonomic Society, Long Beach, CA.

Kentucky Department of Education. (2006, August). Core Content for Science Assessment (Elementary Version 4.1) [Computer software]. Retrieved from http://www.education.ky.gov/KDE/Instructional+Resources/

Curriculum+Documents+and+Resources/Core+Content+for+Assessment/Core+Content+for+Assessment+4.1/

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41,* 75–86.

Klahr, D. (2005). Early science instruction: Addressing fundamental issues. *Psychological Science, 16,* 871–872.

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

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96,* 674–689.

Kuhn, D. (2005). What needs to be mastered in mastery of scientific method? *Psychological Science, 16,* 873–874.

Kuhn, D., & Angelev, J. (1976). An experimental study of the development of formal operational thought. *Child Development, 47,* 697–706.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction, 18,* 495–523.

Kuhn, D., & Dean, D., Jr. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16,* 866–870.

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, 60*(4, Serial No. 245), 1–128.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction, 9,* 285–327.

Li, J., Klahr, D., & Jabbour, A. (2006). When the rubber meets the road: Putting research-based methods to test in urban classrooms. In *Proceedings of the Seventh International Conference of the Learning Sciences: Making a difference* (pp. 418–424). Mahwah, NJ: Erlbaum.

Mayer, R. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist, 59,* 14–19.

McDonald, S.-K., Keesler, V. A., Kauffman, N. J., & Schneider, B. (2006). Scaling-up exemplary interventions. *Educational Researcher, 35,* 15–24.

National Research Council. (1996). *National science education standards.* Washington, DC: National Academy Press.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *Hierarchical linear and nonlinear modeling* (2nd ed.). Lincolnwood, IL: Scientific Software International.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development, 77,* 1–15.

Ross, A. J. (1988). Controlling variables: A meta-analysis of training studies. *Review of Educational Research, 58,* 405–437.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32,* 102–109.

Schauble, L., Glaser, R., Duschl, R., Schulze, S., & John, J. (1991). Students' understanding of the objectives and procedures of experimentation in the science classroom. *The Journal of the Learning Sciences, 4,* 131–166.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press.

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

Toth, E. E., Klahr, D., & Chen, Z. (2000). Bridging research and practice: A research-based classroom intervention for teaching experimentation skills to elementary school children. *Cognition and Instruction, 18,* 423–459.

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 and Instruction, 21,* 149–173.

Zohar, A., & Aharon-Kravetsky, S. A. (2005). Exploring the effects of cognitive conflict and direct teaching for students of different academic levels. *Journal of Research in Science Teaching, 42,* 829–855.

Zohar, A., & David, A. B. (2008). Explicit teaching of meta-strategic knowledge in authentic classroom situations. *Metacognition Learning, 3,* 59–82.

Zohar, A., & Peled, B. (2008). The effects of explicit teaching of metastrategic knowledge on low- and high-achieving students. *Learning and Instruction, 18,* 337–353.

Received September 23, 2008
Revision received September 22, 2009
Accepted October 8, 2009 ■