Bridging Research and Practice: A Cognitively Based Classroom Intervention for Teaching Experimentation Skills to

(b) challenges in instructional practice can lead to new questions for basic research. Although some have lamented the substantial proportion of nonoverlapping work in these two areas (e.g., Strauss, 1998), there does, indeed, exist an active area of intersecting research, as indicated by 15 years of articles in *Cognition and Instruction* as well as by two volumes of the same name spanning a 25-year period (Carver & Klahr, in press; Klahr, 1976).

Nevertheless, with a few notable exceptions (e.g., Brown, 1992, 1997; Fennema et al., 1996; White & Fredriksen, 1998), most of the research in the intersection between cognition and instruction is carried out by researchers whose predilection is to conduct their work in either the psychology laboratory or the classroom, but not both. Consequently, reports of laboratory-based research having clear instructional implications typically conclude with a suggested instructional innovation, but one rarely finds a subsequent report on associated specific action resulting in instructional change. Similarly, many instructional interventions are based on theoretical positions that have been shaped by laboratory findings, but the lab procedures have been adapted to the pragmatics of the classroom by a different set of researchers (e.g., Christensen & Cooper, 1991; Das-Smaal,

This article is organized as follows. First, we describe the topic of the instruction—a procedure for designing simple controlled experiments—and its place in the elementary school science curriculum. Then, we summarize the laboratory training study that provided a rigorous basis for our choice of the type of instruction to be used in our classroom intervention. With this as a background, we describe our approach to creating a *benchmark 2urhson [(f/F1 1 T7.491746 0 TD (diS(2uark)-2*

CVS performance, but only a handful of the studies in his sample included young elementary school children (i.e., below Grade 5). The results of these studies, as well as more recent ones for that age range, present a decidedly mixed picture of the extent to which young elementary school children can understand and execute CVS (Bullock & Ziegler, 1996; Case, 1974; Kuhn & Angelev, 1976; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Schauble, 1996). Moreover, even when training studies show statistically significant differences between trained and untrained groups,[1] the absolute levels of posttest performance are well below educationally desirable levels.

BACKGROUND: A LABORATORY TRAINING STUDY

Given the importance of CVS and given that few elementary school children spontaneously use it when they should, it is important to know whether there are effective ways to teach it and whether age and instructional method interact with respect to learning and transfer. One of the most controversial issues in instruction is whether unguided exploration is more or less effective than such exploration accompanied by highly directive and specific instruction from a teacher. Chen and Klahr (1999) addressed this question in the psychology laboratory. They compared different instructional methods in a context in which children had extensive and repeated opportunities to use CVS and design, conduct, and evaluate their own experiments. A total of 87 second, third, and fourth graders were randomly assigned to one of three different instructional conditions:

1. *Explicit training* was provided in the training–probe condition. It included an explanation of the rationale behind controlling variables as well as examples of how to make unconfounded comparisons. Children in this condition also received probe questions surrounding each comparison (or test) that they made. A probe question before the test asked children to explain why they designed the particular test. After the test was executed, children were asked if they could "tell for sure" from the test whether the variable they were testing made a difference and also why they were sure or not sure.
2. *Implicit training* was provided in the no-training–probe condition. Here, children did not receive explicit training, but they did receive probe questions before and after each of their experiments, as described previously.
3. *Unprompted exploration* opportunities were provided to children in the no-training–no-probe condition. They received neither training nor probes, but

[1]Ross (1988) found a mean effect size of .73 across all of the studies in his sample.

comparison from the ramps task. It is a confounded comparison because all four variables differ between Ramp A and Ramp B.

Procedure

Part I of the laboratory study consisted of four phases: exploration, assessment, transfer-1, and transfer-2. In each phase, children were asked to construct experi-

of a set of pairwise experimental comparisons in a variety of domains. The child's task was to examine the experimental setup and decide whether it was a good or a bad experiment (this type of assessment was used extensively in the classroom study, and it is described subsequently).

Measures

The classroom study detailed in this article uses several measures from the laboratory study by Chen and Klahr (1999) so we describe them here. *CVS performance score* is a simple measure based on children's use of CVS in designing tests. *Robust use of CVS* is a more stringent measure based on both performance and verbal justifications (in response to probes) about why children designed their experiments as they did. *Domain knowledge* is a measure of children's domain-specific knowledge based on their responses to questions about the effects of different causal variables in the domain. We employ all these measures in the classroom study in addition to new measures that were specific to the classroom study, which we describe in more detail later.

Results of the Laboratory Training Study

Only children in the training–probe condition increased their CVS knowledge significantly across the four phases in the laboratory study conducted by Chen and Klahr (1999); that is, expository instruction combined with probes led to learning, whereas neither probes alone nor unguided exploration did so. However, Chen and Klahr found grade differences in students' ability to transfer CVS between tasks and domains. Although second-graders' CVS scores increased marginally immediately after instruction, they dropped back to baseline levels in the transfer phases (when they had to remember and transfer what they learned about designing unconfounded experiments from, for example, springs to ramps and sinking objects). However, the third and fourth graders who participated in expository instruction successfully transferred their newly acquired CVS skills to near transfer domains, whereas only fourth graders were able to retain the skill and show significantly better CVS performance (as compared to untrained fourth graders) on the paper-and-pencil posttest administered 7 months later.

For the purposes of this study, the most important results from Chen and Klahr (1999) were that (a) absent expository instruction, children did not learn CVS,[2]

[2]Although children did not learn the CVS strategy by experimentation alone, they did spontaneously learn a different type of knowledge—knowledge about the domain itself. In no condition was there any direct instruction on domain knowledge.

even when they conducted repeated experiments with hands-on materials; (b) brief expository instruction on CVS was sufficient to promote substantial gains in CVS performance; and (c) these gains transferred to both conceptually near and (for fourth graders) far domains.

MOVING FROM THE LABORATORY TO THE CLASSROOM: THE DESIGN OF A BENCHMARK LESSON

tive classroom environment? (b) What is the relation between students' experimentation skills and the acquisition of domain knowledge? (c) Will instruction focused on the design and justification of students' own experiments also increase their ability to evaluate experiments designed by others? (d) What additional research questions are raised during the move from the psychology laboratory to the classroom? Throughout the process of engineering the classroom learning environment, we conceptualized our task in terms of differences and similarities between lab and classroom with respect to pedagogical constraints, pragmatic constraints, and classroom assessment (Table 1).

Pedagogical Constraints

For an effective instructional intervention that involved only minimal changes from the instructional procedures used in the laboratory research, we maintained both the instructional objective (teaching CVS) and the proven instructional methodology (expository instruction) from the earlier laboratory study. The instructional materials were the same ramps as used in the laboratory study. Students designed experiments by setting up different variables on two ramps and comparing how far a ball rolled down on each ramp. Within these constraints, there were several important differences between the laboratory script and the classroom lesson.

Pragmatic Constraints

The move from the laboratory to the classroom environment required us to consider numerous pragmatic constraints. Instead of a single student working with an experi-

TABLE 1
Comparison of the Pragmatics and Instructional Methods in the Laboratory and Classroom Studies

Considerations	*Laboratory Study*	*Classroom Study*
Pedagogical constraints		
Instructional objective	Mastery of CVS	Mastery of CVS
Instructional strategy	Expository instruction of one student. Active construction, execution, and evaluation of experiments by solo student.	Expository instruction—group of students. Active construction, execution, and evaluation of experiments by group (unequal participation possible).

the classroom work; see the Methods section.) Students in each group made joint decisions about how to set up their pair of ramps but then proceeded to record individually both their setup and the experimental outcome in their laboratory worksheets (the recording process is explained in more detail subsequently).

able time, and the fit between the CVS topic and the normal progression of topics through the fourth-grade science curriculum. From these four classrooms, we recruited volunteers for pre- or postinstruction interviews. We received parental permission to individually interview 43 of the 77 students participating in the class-

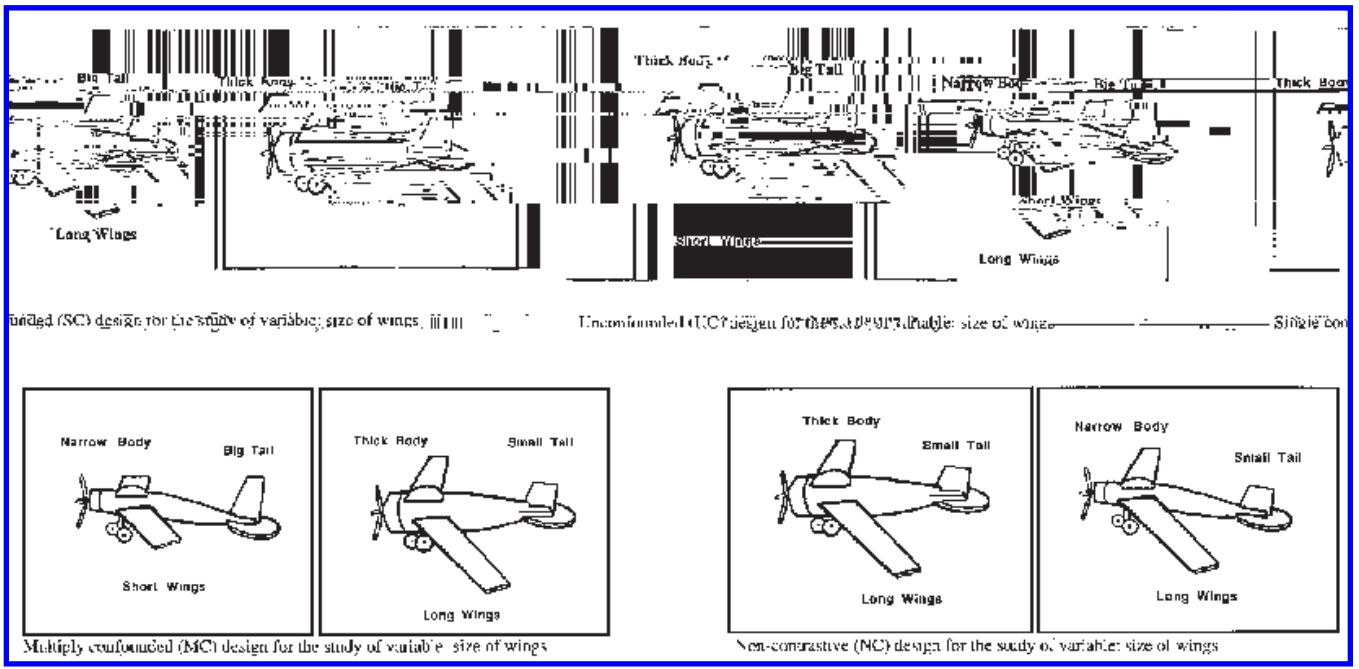

FIGURE 3 Comparison types used in experiment evaluation assessment booklet.

The interviewer followed the same script used in Chen and Klahr (1999). Students were asked to design and conduct nine experiments. The experiments were student-designed comparisons to decide whether a selected variable makes a difference in the outcome. After designing their comparisons, students were asked to justify these experiments. They also were asked the same questions employed by Chen and Klahr to indicate how certain they were about the role of the focal variable from the outcome of the experiment composed. They were asked, "Can you tell for sure from this comparison whether [variable] makes a difference? Why are you sure–not sure?" The entire session was recorded on videotape.

Experiment Evaluation Assessment

At the start of the first day of the classroom work, all students individually completed a paper-and-pencil experiment evaluation test on which they judged preconstructed experiments to be good or bad. Students were presented with 10-page test booklets in which each page displayed a pair of airplanes representing an experimental comparison to test a given variable. For each airplane, three variables were used: length of wings, shape of body, and size of tail. Figure 3 depicts some of the types of comparisons used on the experiment evaluation assessment.

Four different types of experiments were presented: (a) unconfounded comparisons, which were correct, controlled comparisons in which only the focal variable was different between the two airplanes; (b) singly confounded comparisons, in which the two airplanes differed in not only the focal variable, but also in one additional variable; (c) multiply confounded comparisons, in which the airplanes differed on all three variables; and (d) noncontrastive comparisons, in which only one variable was different between the airplanes, but it was not the focal variable. Stu-

The next phase of classroom work consisted of what we call *expository instruction* combined with exploration and application. This method of instruction con-

planations. The teacher asked the students to point out what variables were different between the two ramps and asked whether they would be able to "tell for sure" from this comparison whether the focal variable made a difference in the outcome.

2. *Model correct thinking.* After a number of conflicting opinions were heard, the teacher revealed that the example was not a good comparison. She explained that other variables, in addition to the focal variable, were different in this comparison and, thus, if there was a difference in the outcome, one could not tell for sure which variable had caused it. The teacher proceeded to make a good comparison to contrast with the bad one and continued a classroom discussion to determine why the comparison was good. (For simplicity of instruction—and to avoid drawing attention to other error sources—the teacher did not roll the balls during her instruction and focused on the logical aspects of designing good comparisons.)

3. *Test understanding.* Next, the teacher tested the students' understanding with another bad comparison and asked questions similar to those asked earlier.

4. *Reinforce correct thinking.* By pointing out the error in the bad comparison and providing a detailed account of the confounds in the bad test, the teacher reinforced students' correct thinking. The teacher created another good comparison and used the same method of classroom discussion as before to review why this test allowed one to tell for sure whether the studied variable makes a difference.

5. *Summarize rationale*. As a final step, the teacher provided an overall generalization for CVS with the following words:

> Now you know that if you are going to see whether something about the ramps makes a difference in how far the balls roll, you need to make two ramps that are different only in the one thing that you are testing. Only when you make those kinds of comparisons can you really tell for sure if that thing makes a difference.

Application experiments. The third phase of the classroom work allowed students to apply the newly learned CVS to another set of experiments. The students' activity in this phase was very similar to that of the exploratory experiment phase: setting up comparisons between two ramps to test the effect of different variables. The teacher's role in this phase was also similar to that during exploratory experimentation: The teacher facilitated collaborative work but did not offer evaluative feedback on students' experimental designs.

Measures

Our measures are designed to capture both the procedural and logical components of CVS. In addition to using all of the measures of the Chen and Klahr (1999) study, in the classroom study we introduced a new measure: certainty. We now give an overall summary of measures with associated scoring techniques.

CVS Performance Score

We measured students' CVS performance by scoring the experiments students conducted, that is, the way they set up the pair of ramps to determine the effect of a focal variable. Each valid, unconfounded comparison was given a score of 1, and all other invalid comparisons (singly confounded, multiply confounded, noncontrastive) were given a score of zero. This method was used for scoring both the individual interviews and the experiments students recorded on the laboratory worksheets. During the individual interviews, students conducted nine experiments for a maximum of 9 points. During classroom work, students conducted four experiments before and four experiments after instruction, so the maximum possible CVS score for each phase of classroom work was 4.

Robust CVS Use Score

During individual interviews, students were asked to give reasons for their experiments. A score of 1 was assigned to each experiment in which a student gave a CVS-based rationale in response to at least one of the two probe questions for that experiment. Robust CVS use was scored by measuring both CVS performance and the rationale the student provided for the experiment. This yielded a score of 1 for each valid experiment accompanied by a correct rationale. Maximum possible robust use score was 9. Robust use scores were computed for interviews only, as classroom worksheets did not ask for experimental design justifications.

Domain Knowledge Score

Students' domain knowledge was assessed by asking them to indicate which level of each variable made the ball roll farther down the ramp. Students were provided with a choice of the two levels for each variable (e.g., high–low, long–short) and were asked to circle their answer. Correct responses were scored as 1 and incorrect responses as zero.

Experiment Evaluation Score

Students' ability to evaluate experimental designs created by others was assessed with the pre- and postinstruction experiment evaluation tests (airplanes comparisons) described previously (Figure 3). Correctly indicating whether a given experimental comparison was good or bad gained students a score of 1, and incorrect evaluations were scored zero. In addition, individual students' patterns of responses to the 10-item experiment evaluation instrument were used to identify several distinct reasoning strategies.

Certainty Measure

The certainty score was not examined in the previous laboratory study. It is intended to capture the complexity of the type of knowledge students extracted from classroom experiences. In both individual interviews and classroom worksheets, probe questions asked students after each experiment whether they were certain of their conclusion about the role of the focal variable. To judge certainty, a simple yes–no response was then recorded after the question "Can you tell for sure from this experiment whether the [variable] of the [domain] makes a difference in [outcome]?" In the individual interviews, students also were asked to state their reasons for certainty. Answers to these questions were recorded and coded. To simplify procedures in the classroom, students were not asked to provide a rationale for their certainty on the worksheets.

RESULTS OF THE CLASSROOM STUDY

First, we present the results on students' knowledge about CVS, based on individual interviews and classroom worksheets. Second, we describe students' domain knowledge, that is, knowledge about which values of the variables make a ball roll farther, based on tests administered before and after classroom instruction. Third, we report on changes in students' ability to discriminate between good and bad experiments created by others. Fourth, we describe additional findings, such as students' experiment evaluation strategies and certainty of conclusions, that point to the inherent complexity of learning and teaching experimentation skills in elementary science classrooms and the various sources of error that can play a role during classroom learning.

CVS Performance and Robust CVS Use From Individual Interviews

CVS Performance

First, we looked at whether there were any changes in CVS scores during the preclassroom individual interviews. These interviews corresponded to the no-training–probe condition in which Chen and Klahr (1999) found only a marginally significant improvement for their fourth-graders. However, in this study, we did find some improvement across trials during the preinstructional individual interviews. Students conducted nine different experiments (three with each of three variables in either the springs or the sinking task) during these interviews. For ease of calculation, the scores for the three trials on each variable were collapsed into one score,

yielding a total of three scores—one for each variable. Mean performance scores improved from 17% correct on the first variable to 41% correct on the third, $F(2, 82) = 5.8$,

Analysis of CVS Performance From Classroom Activities

The nested design used in this study allowed us to measure several of the same constructs in both the individual interviews and the classroom (Figure 2). In this section, we describe the results of the inner pairs of pre–post measures, the results of classroom activities.

Analysis of Classroom Laboratory Worksheets

During classroom activities, students worked in 22 small groups. Although the students made their ramp setup decisions and built experimental comparisons together, each student individually filled out a laboratory worksheet. The analysis presented here is based on group performance because all the members of each

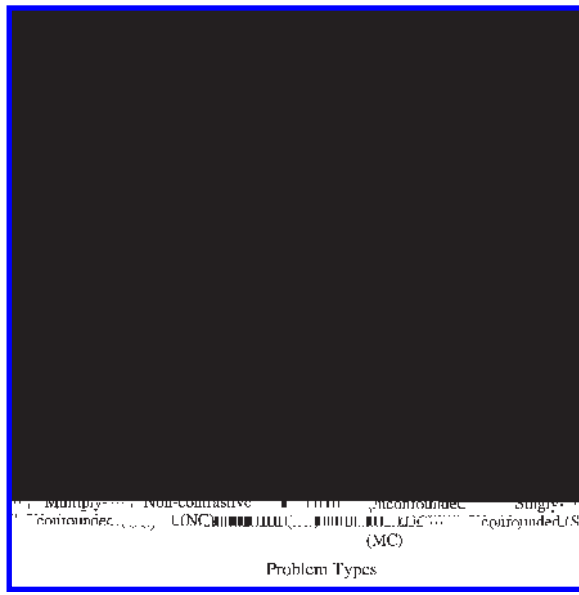

FIGURE 4 Students' ability to evaluate four different experimental designs.

mance on singly confounded and multiply confounded designs were significantly different compared to noncontrastive designs even after instruction, $t(73) = 2.2$, $p = .03$ (Figure 4).

These differences among the different problem types (some of which remained stable even after instruction) suggested that students might be using consistent but incorrect strategies to evaluate experimental designs. We hypothesized that students might have problems with three distinct aspects of evaluating experimental designs: (a) recognition of the focal variable, (b) action on focal variable, and (c) action on other nonfocal variables. Consistent CVS use can occur only if students recognize the focal variable of an experiment, change only that one variable between items compared, and keep all other, nonfocal, variables constant. All other possible combinations of actions yield incorrect experimental designs. Combining these three aspects yields five possible strategies:

1. *Vary only the focal variable and control other variables.* This is the correct CVS. Students using this strategy correctly recognize that only unconfounded problems are correct designs.
2. *Vary focal variable but ignore others.* Students who use this strategy judge all problem types, except the noncontrastive comparisons, as correct.
3. *Vary any (only) one variable and control all others.* Students using this strategy do not focus on the role of the focal variable; thus, they judge all

TABLE 4
Trends in Strategy Use Prior to Instruction and After Instruction from the Experimental Design Evaluation Test

				Answers to Problem Types				*Number of Students*	
Student Strategy	*Focus on Focal Variable?*	*Variable Changed?*	*Action on Other Variables*	*UC*	*SC*	*MC*	*NC*	*Pre*	*Post*
Vary only focal variable, control others (correct)	Yes	Focal variable only	Control them[a]	Good	Bad	Bad	Bad	20 (27%)	57 (77%)
Vary focal variable, ignore others	Yes	Focal variable	Ignore them	Good	Good	Good	Bad	12 (16%)	7 (9%)
Vary any one (only one) variable, control others	No	Any variable	Control them[a]	Good	Bad	Bad	Good	6 (8%)	1 (1.3%)
Vary at least one variable, ignore others	No	Any variable	Ignore them	Good	Good	Good	Good	14 (19%)	2 (2.7%)
Other (vary all? random?)								22 (30%)	7 (9%)
N								74[b]	74[b]

Note. UC = unconfounded design; SC = singly confounded design; MC = multiply confounded design; NC = noncontrastive design.

[a]Keep them the same. [b]Three students were absent during both the preinstruction and postinstruction experiment evaluation test.

Students' Certainty

Detailed examination of the interview and laboratory data linked students' certainty of the effect of the focal variable with the validity of their experiments. In both the individual interviews and the laboratory worksheets, students were asked to indicate their certainty in the role of the focal variable after each experiment. They were asked "can you tell for sure from this experiment whether the [focal variable] makes a difference in [the outcome]?" The type of experiment students composed (correct CVS or incorrect), students' statements of certainty, and the reasons for their certainty indicated during the interview were recorded and analyzed. Our expectation was that, with an increase in correct experimentation after instruction, students' certainty in their conclusions about the role of the focal variable would increase as well. However, this hypothesis was not confirmed. The subsequent sections detail students' certainty from individual interviews and from laboratory worksheets.

Individual interviews. Before instruction, 70% of the answers students gave after correct experiments indicated certainty in the role of the focal variable. After instruction, when students' CVS performance was nearly at ceiling (as discussed earlier), 84% of the answers after valid experiments indicated certainty. This change in certainty was not significant, $t(37) = 1.5$, $p = .14$. Thus, despite near-perfect CVS performance scores after instruction (97%), students remained uncertain about their inferences after 16% of these controlled experiments (Table 5).

Even more curious, when we analyzed the consistently good performers' reasons for certainty, we found an interesting pattern of responses. Recall that during the individual interviews after instruction, there were 20 consistent CVS performers, who created correct, CVS-based comparisons on at least eight of their nine experiments. Out of the 180 experiments these 20 students made, they composed 179 controlled experiments. After correct experiments, these consistently good performers gave the following rationales for certainty of inferences:

1. Experimental setup, indicating attention to CVS
2. Data outcome, indicating attention to the outcome of their tests
3. Prior theory, indicating answers based on students' existing domain knowledge
4. Combination of systemic setup and data outcome, indicating reasoning that is closest to the scientific way of evaluating experimental outcome

Those students who were certain that they could draw a valid inference from their correct experiment cited the experimental setup as a reason for their certainty on 37% of their answers. These students mentioned data outcome (22%), their prior domain theory (15%), or the combination of data outcome and prior theory (15%) less frequently than experimental setup as their reason for certainty. On the other hand, the students who indicated that they were uncertain after correct experiments formulated their rationale for certainty primarily based on the data outcome (38%; Table 6).

TABLE 5
Percentage of Correct Experiments and Certainty After These Experiments From Individual Interviews Before and After Instruction

	Correct Experiments (%)	*Certain After Correct Experiments (%)*
Before instruction	30	70
After instruction	97	84

TABLE 6
Consistent CVS-Performers' Certainty About the Role of the Focal Variable After Correct Experiments During Individual Interview After Instruction

		Stated Reason				
Certainty Level	*Number of Answers*	*System Setup (%)*	*Data Outcome (%)*	*System and Outcome (%)*	*Theory (%)*	*Other (%)*
Total	179	20	30	7.5	16	26.5
Certain	150	37	22	15	15	11
Uncertain	29	3	38	0	17	42

Laboratory worksheets. During laboratory work, students conducted experiments in small groups but then individually recorded their certainty after each experiment on their worksheets. Prior to instruction, students indicated certainty after 76% of their correct experiments. After classroom instruction—although the percentage of valid experiments increased significantly—the frequency of certainty after correct experiments remained the same: 76%. Thus, during classroom work, just as during interviews, the dramatic increase in CVS procedural knowledge was accompanied by a relatively constant level of uncertainty about the conclusions students could draw from their valid experiments.

Furthermore, we studied the relation between students' certainty and the number of correct experiments they composed during classroom work. Again, we looked at the certainty of consistently good CVS performers, making a distinction between those who did well on this score from the beginning of classroom work (prior to instruction) and those who became consistently good performers after instruction. We categorized the students who composed correct experiments for all their designs from the beginning of the classroom work as the *know-all-along* groups. The students who composed valid experiments on all their designs after instruction, but who did not consistently use the CVS strategy before instruction, were called the *learn-by-end* groups. We expected that the more experiments a group conducted with the CVS strategy, the higher their certainty would be, so that the students in the know-all-along group would display a larger gain in their certainty over time than would the learn-by-end group.

Although the certainty of the learn-by-end students increased by the end of the classroom work, this increase was not significant (M_{pre} = .51, M_{post} = .81), $t(12)$ = .97, p = .35. To our surprise, the overall certainty of the know-all-along group significantly decreased by the end of the classroom unit. Even though these students composed correct experiments 100% of the time, both prior to and after instruction, they were certain of the role of the focal variable in 87% of their answers before instruction and in just 73% of their answers after instruction, $t(27) = 2.56$, p = .017 (Table 7).

DISCUSSION OF CLASSROOM STUDY RESULTS

The main goal of this study was to determine whether an instructional methodology that produced substantial and long-lasting learning in the psychology laboratory could be transformed into an effective instructional unit for classroom use. Our results from the classroom study confirmed the findings of the prior laboratory study: Expository instruction embedded in exploratory and application experiments is an effective method to teach CVS. We found significant gains in students' ability to perform controlled experiments and provide valid justifications for their controlled designs, in their conceptual knowledge of the domain studied, and in their ability to evaluate experiments designed by others. We also found a few surprises and issues for further research on classroom learning and teaching.

CVS Performance and Justification in a Classroom Setting

As indicated by a series of independent but converging measures, expository instruction combined with hands-on experimentation was overwhelmingly successful and led to educationally relevant gains. As expected, CVS performance data collected from both the individual interviews and laboratory worksheets prior to and after instruction indicated significant performance increases. With respect to the consistency of students' CVS performance (their ability to perform correct experiments at least eight times out of nine trials during interviews), we also found a significant increase after instruction.

In addition, when we examined students' CVS performance prior to instruction, we found a significant increase in this measure due to experimentation alone. However, students' robust CVS use (correct performance with valid justification) did not increase by experimentation alone, that is, the complex skill of CVS performance combined with justification was not learned by experimentation alone. Expository instruction, however, provided significant (though not 100%) learning gains even on this stringent measure. This finding on robust CVS use started to

TABLE 7
Percentage of Correct Experiments and Certainty After These Experiments From Consistent CVS User Students' Classroom Worksheets Before and After Instruction

	Before Instruction (%)		*After Instruction (%)*	
	Correct Experiments	*Certain After Correct Experiments*	*Correct Experiments*	*Certain After Correct Experiments*
Know all along	100	87	100	73[a]
Learn by end	26	51	100[a]	81

[a]These were significant changes after instruction, $p < .05$.

highlight some of the difficulties inherent in classroom experimentation. Although both students' ability to create controlled experiments accompanied by valid justification and the consistent use of these justifications (eight times out of nine experiments) significantly increased after expository instruction, consistent CVS justification after correct experiments (consistent robust CVS use) remained well below consistent CVS performance. Of course, because CVS use is a necessary, but not sufficient, component of robust CVS use, robust CVS use can never exceed it. Nevertheless, we were struck by the size of the discrepancy between the two scores following instruction. The most likely explanation may be simply that although we explicitly taught children how to do CVS, we only indirectly taught them how to justify their procedures verbally. The importance of additional instruction on this aspect of scientific reasoning remains a topic for future investigation. Perhaps the provision of additional supports for scientific reasoning such as external representations (evidence maps, tables, and graphs) could improve students' ability to justify verbally their experimental designs and inferences (Toth, 2000; Toth, Suthers, & Lesgold, 2000).

Students' Experimentation Skills and Domain Knowledge

Even though specific domain knowledge was not explicitly taught, students' domain knowledge (i.e., about ramps) increased significantly after instruction on CVS. Our explanation of this domain knowledge increase is that data from valid experiments helped students learn about the correct role of each variable studied. Although Chen and Klahr (1999) found the same outcome during their laboratory study, these results are only preliminary, and further studies will help us more closely examine the relation between valid experimentation skills and domain knowledge learning.

Students' Ability To Evaluate Experiments Designed by Others

Individual students' ability to evaluate experiments designed by others increased significantly after classroom instruction. In addition, there was a significant increase in the proportion of students who could correctly identify a good or bad design at least 9 times out of 10. Thus, even brief expository instruction on CVS—when embedded in student experimentation—increased individual students' ability to evaluate designs composed by others. A detailed examination of experiment evaluation performance indicates that on both the initial and final tests students were most successful in judging unconfounded and noncontrastive experiments, and their main difficulties were in judging confounded experiments. This result prompted us to examine closely students' experiment evaluation strategies and, among other issues (briefly summarized in the next section), provided momentum for further studies in both the applied, classroom setting as well as the laboratory.

Surprises and New Directions for Research

Whereas our main focus in this study and elsewhere (Klahr et al., in press) has been the transition from the psychology laboratory to the classroom, our work in the classroom also yielded numerous ideas for further consideration in both laboratory and classroom research. These new issues include the presence of naive strategies employed by students during experimentation and the peculiar uncertainty about inferences during experimentation.

Students' Strategies of Experiment Evaluation

We found that a substantial proportion of students applied incorrect CVS strategies before instruction. Analysis of the experiment evaluation test revealed consistent differences in students' performance on the four problem types (unconfounded, singly confounded, multiply confounded, and noncontrastive) and, in turn, prompted our analysis of strategy use. Although the number of students employing correct evaluation strategies improved after instruction, "buggy" strategies did not disappear. This result suggests the need for a refined instructional methodology that is aimed directly at correcting specific aspects of these erroneous strategies.

Students' Certainty and Reasoning

An examination of students' certainty in their inferences and their reasoning during experimentation in individual interviews and classroom laboratory work revealed some unexpected and potentially important findings. One was that although students' CVS performance increased substantially, there remained a nontrivial proportion of valid experiments from which they were unable (or unwilling) to draw unambiguous conclusions. After instruction, approximately one sixth of the students in the individual interviews and one fourth of the students in classroom experimentation would not state that they were certain about the effect of the focal variable on the outcome of the experiment even after they conducted valid experiments. Because all ramp variables influenced the outcome measure, this finding was surprising. It led us to examine the reasoning behind students' certainty judgments after correct experiments.

With the detailed analysis of reasons given during individual interviews, we found that students' reasoning was distinctively different based on whether they were certain or uncertain in the inferences they could draw after correct experiments. On more than one third of the certain responses after correct experiments, students supported their conclusions by citing their use of CVS. Those students who were uncertain after correct experiments supported their judgment more often by using the actual outcome and their prior theories about the domain rather than

CVS-related logic. Furthermore, the analysis of students' individual records during classroom work revealed that certainty did not directly increase with more valid experiments performed; in fact, the certainty of those students who performed correct experiments both before and after instruction (the know-all-along students) decreased significantly.

We believe that these patterns of reasoning can be attributed to the fact that children face various error sources during experimentation. The control of variables strategy teaches students to overcome one error source: the logical error associated with the systemic setup of experiments. Other types of errors (e.g., measurement and random error) also can occur during experimentation and can make it difficult to draw clear inferences even after valid experiments. Consequently, although the learn-by-end groups—who were learning the CVS strategy during instruction and thus were not focused on other error sources—increased their CVS performance, they did not increase their overall certainty in the inferences they can draw from these correct experiments. We hypothesize that those students who possessed the CVS strategy prior to instruction were able to focus on error sources unrelated to the experimental setup during their experimentation and were more aware of data variability due to these error sources. In the face of these data deviations, the know-all-along students' certainty in the conclusions drawn from their valid experiments significantly decreased.

Clearly, the experiments conducted in the complex classroom settings imposed various sources of error other than the error associated with the setup of experiments, which was the focus of instruction. Although these error sources are important aspects of a rich understanding of experimentation, we did not include them in our highly focused instructional goals. Students struggled with these error sources, trying to combine them with their understanding of systematic, controlled experimentation. This led us to examine these additional sources of error during complex classroom experimentation. We recently conducted a second classroom intervention in which we studied further the role of errors such as measurement and random errors in addition to systemic error and the nature of students' conceptions of these error sources (Toth & Klahr, 2000). The results of this classroom study motivated the detailed (laboratory-based) examination of the same issues (Masnick & Klahr, 2000).

of these to classroom teaching and learning. Klahr and Chen, as experimental psychologists, traditionally studied learning in the psychology laboratory and built theories about the nature of scientific inquiry (Klahr, 2000; Klahr & Dunbar, 1988) and the essential components of learning such as analogical reasoning (Chen, 1996). Toth is a former science teacher with training in curriculum and instruction and experience working with teachers on classroom science learning challenges (Coppola & Toth, 1995; Levin, Toth, & Douglas, 1992; Toth, 2000; Toth et al., 2000). This diversity in backgrounds enabled us to move successfully between the fields of psychology and education and to use laboratory research to establish effective classroom practice. Thus, we were able to construct a sustained research cycle that contained three phases: (a) use-inspired, basic research in the laboratory; (b) classroom verification of the laboratory findings; and (c) follow-up applied (class-

and Mrs. Cheryl Little, who braved the rough waters of innovation and gave their time and energy to support our efforts in building an effective classroom learning environment. Numerous colleagues also were instrumental during the implementation of research and preparation of this document. We thank Jennifer Schnakenberg, Anne Siegel, Sharon Roque, and Jolene Watson for their invaluable assistance in data collecting, coding, and analysis. Leona Schauble, Bradley Morris, and two anonymous reviewers provided invaluable comments on earlier drafts of this article.

REFERENCES

Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2,* 141–178.

Brown, A. (1997). Transforming schools into communities of thinking and learning about serious matters. *American Psychologist, 52,* 399–413.

Bullock, M., & Ziegler, A. (1996). Scientific reasoning: Developmental 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 Planck Institute for Psychological Research.

Carver, S. M., & Klahr, D. (Eds.). (in press). *Cognition and instruction: 25 years of progress.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Case, R. (1974). Structures and strictures: Some functional limitations on the course of cognitive growth. *Cognitive Psychology, 6,* 544–573.

Chen, Z. (1996). Children's analogical problem solving: Effects of supe4olv53.nal,53.1(solvctures)-153.1(pro)-158.83obl

Klahr, D., Chen, Z., & Toth, E. E. (in press). From cognition to instruction to cognition: A case study in elementary school science instruction. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from professional, instructional, and everyday science*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–55.

Kuhn, D., & Angelev, J. (1976). An experimental study of the development of formal operational

Lawrence Erlbaum Associates, Inc. does not have electronic rights to Table A1. Please see the print version.