COGNITION AND INSTRUCTION, *18*(4), 423–459

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

Eva Erdosne Toth, David Klahr, and Zhe Chen
Department of Psychology
Carnegie Mellon University

This article describes the first cycle of a multiyear research project aimed at establishing a common ground between educationally relevant psychological research and educational practice. We translated a theoretically motivated, carefully crafted, and laboratory-based instructional procedure of proven effectiveness into a classroom intervention, making minimal modifications to the instructional components and adapting to the constraints of an elementary school science classroom. Our intervention produced significant gains in fourth-grade students' ability to create controlled experiments, provide valid justifications for their experiments, and evaluate experiments designed by others. It also raised several new questions about how students understand sources of error during experimentation and how that understanding is related to their level of certainty about conclusions that are supported by the experimental outcomes. We view this report as part of a continuing research cycle that includes 3 phases: (a) use-inspired, basic research in the laboratory; (b) classroom verification of the laboratory findings; and (c) follow-up applied (classroom) and basic (laboratory) research.

Two beliefs widely shared by readers of this journal are that (a) basic research in cognitive and developmental psychology can contribute to instructional practice, and

Requests for reprints should be sent to Eva Erdosne Toth, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213–3890. E-mail: etoth+@andrew.cmu.edu

(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, Klapwijk, & van det Leij, 1996). This division of labor between laboratory-based cognitive research and classroom research is understandable but, in our view, unnecessary and inefficient because much can be lost in the translation from the psychology laboratory to the classroom.

In this article, we describe an attempt to establish a more tightly coupled sequence of laboratory research and practical innovation conducted by the same set of researchers. The effort involved collaboration among researchers from both developmental and educational psychology, as well as the classroom teachers themselves. We translated a theoretically motivated, carefully crafted, and laboratory-based instructional procedure of proven effectiveness into a classroom intervention, making minimal modifications to the instructional components and adapting to the constraints of a real classroom.

The practice-inspired research cycle described here further supports both of the widely held beliefs cited previously (in the first paragraph). First, instruction based on prior laboratory research was educationally effective. Children learned and transferred what they had been taught. Second, once the instruction was situated in a real classroom, a new set of basic research issues arose, and they are currently under investigation. Because we view the move from the laboratory-based research environment to the classroom as fraught with potential pitfalls, we took a very small step—a "baby step," perhaps—in making this move. Nevertheless, or perhaps consequently, we were able to devise an effective curriculum unit that maintained what we believe to be the fundamental features of the laboratory instruction and that still is consistent with everyday classroom practice.

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 lesson* (diSessa & Minstrell, 1998), using the results of the laboratory study and considering the characteristics of the classroom environment. Next, we present the design, implementation, and results of a study that aimed to verify the laboratory findings in a classroom situation. Finally, we discuss some of the new issues raised during the implementation of classroom instruction, which resulted in follow-up research in both the applied (classroom) and the laboratory settings.

DESIGNING UNCONFOUNDED EXPERIMENTS: THE CONTROL OF VARIABLES STRATEGY

The ability to design unconfounded experiments and to derive valid inferences from such experiments are two fundamental aspects of scientific inquiry. There is wide agreement among science educators and policymakers that "even at the earliest grade levels, students should learn what constitutes evidence and judge the merits or strengths of the data and information that will be used to make explanations" (National Research Council, 1996, p. 122). Although the importance of teaching children to design, execute, and evaluate controlled experiments is emphasized in national standards, the methods of teaching these concepts and procedures are not well specified. At the heart of scientific experimentation is the ability to use the control of variables strategy (CVS). Procedurally, CVS is a method for creating experiments in which a single contrast is made between experimental conditions. In addition to creating such contrasts, the full strategy involves being able to distinguish between confounded and unconfounded experiments. Conceptually, CVS is based on the ability to make determinate inferences from the outcomes of unconfounded experiments as well as to understand the inherent indeterminacy of confounded experiments.

How well do elementary school children learn and use these concepts and the procedures associated with them? Neither the educational nor the psychological literature presents a clear answer to this question. For example, Chen and Klahr (1999) found that even in schools with strong science programs, fourth graders' performance on CVS tests—although well above chance—was less than 50% correct. Children in the second and third grade performed even worse. Ross's (1988) meta-analysis of more than five dozen CVS training studies from the 1970s and 1980s indicated that a variety of training methods can generate improvement in

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.

they did receive more opportunities to construct experiments than did children in the other conditions.

Materials, Procedure, and Measures of Laboratory Training Study

Chen and Klahr (1999) used three different domains in which children had to design unconfounded experiments: (a) springs, in which the goal was to determine the factors that affect how far springs stretch; (b) sinking, in which children had to assess the factors that determine how fast various objects sink in water; and (c) ramps (described subsequently). The three domains shared an identical structure. In each, there were four variables that could assume either of two values. In each task, children were asked to focus on a single outcome that could be affected by any or all of the four variables. For example, in the springs task, the outcome was how far the spring stretched as a function of its length, width, wire size, and the size of the weight hung on it. Each child worked with one of the three tasks on his or her first day in the study and then with the other two tasks on the second day. Task order was counterbalanced, as was the order of focal variables within each task. Here, we describe only the ramps task because that task also was used in the classroom intervention described in the next section. (Appendix A summarizes the features of all three domains.)

Ramps Task

In the ramps task, children had to make comparisons to determine how different variables affected the distance that objects rolled after leaving a downhill ramp. Materials for the ramps task were two wooden ramps, each with an adjustable downhill side and a slightly uphill, stepped surface on the other side (see Figure 1). Children could set the steepness of the downhill ramps (high or low) using wooden blocks that fit under the ramps in two orientations. Children could control the surface of the ramps (rough or smooth) by placing inserts on the downhill ramps with either the rough carpet or smooth wood side up. They also could control the length of the downhill ramp by placing gates at either of two different distances from the top of the ramp (long or short run). Finally, children could choose from two kinds of balls: rubber balls or golf balls. To set up a comparison, children constructed two ramps, setting the steepness, surface, and length of run for each and then placing one ball behind the gate on each ramp. To execute a test, participants removed the gates and observed as the balls rolled down the ramps and then up the steps and came to a stop. The outcome measure was how far the balls traveled up the stepped side of the ramp. Figure 1 depicts a

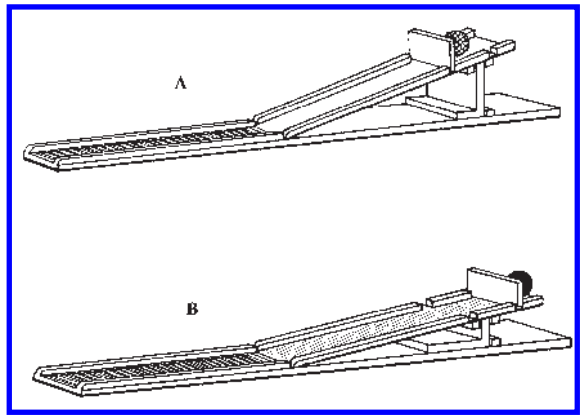

FIGURE 1 The ramps domain. On each of the two ramps, children can vary the angle of the slope, the surface of the ramp, the length of the ramp, and the type of ball. The confounded experiment depicted here contrasts (a) a golf ball on a steep, smooth, short ramp with (b) a rubber ball on a shallow, rough, long ramp (see Appendix A for additional information).

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 experimental test comparisons from which they could make a valid inference about the causal status of a variable of the domain (e.g., in the springs domain, the possible causal variables were spring length, width, wire size, and weight size; see Appendix A for details). The exploration phase established an initial baseline of children's ability to design unconfounded experiments in the first domain (e.g., springs). For the training–probe condition, the exploration phase was immediately followed by expository instruction on how to create controlled experiments. In the subsequent assessment phase, children were asked to design experiments to investigate a different variable in the same domain (e.g., if, in the exploration phase, the experiments focused on spring length, then the assessment phase focused on spring width). Transfer-1 and transfer-2 took place a week after exploration and assessment. Children returned to the laboratory and were asked to design unconfounded experiments in the other two domains that they had not investigated yet (e.g., in the current example, they would do experiments with ramps and with sinking objects).

Part II of the laboratory study was a paper-and-pencil posttest administered 7 months after the individual interviews. This experiment evaluation test consisted

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

Although the laboratory study demonstrates that brief expository instruction about CVS can produce substantial and long-lasting learning, the type of one-on-one instruction and assessment used in a typical psychology experiment—requiring strict adherence to a carefully crafted script—is clearly impractical for everyday classroom use. Furthermore, we were aware that Chen and Klahr's (1999) study had a relatively narrow focus when compared to the multiple goals that classroom teachers usually have when teaching about experimental design. Thus, we formulated the goal of translating, adapting, and enriching the laboratory instructional procedure so that it could be used as a benchmark lesson for a classroom unit, that is, based on the results of prior laboratory studies, we attempted to engineer a classroom learning environment (Brown, 1992) that students could refer to during future hands-on activities calling for the design of informative experiments. In addition, because we wanted to study the effectiveness of this translation, we recognized the need to include a variety of assessment procedures to inform us about the effectiveness of our instruction.

Our progress toward these goals can be divided into five steps: (a) teacher networking, (b) survey of curricula, (c) classroom observation of current practices, (d) development and implementation of a benchmark lesson, and (e) assessment of the effects of classroom instruction on children's ability to use CVS in designing and evaluating experiments.

At the outset, we established a small network of experienced elementary school science teachers, all of whom were already including some aspects of CVS instruction in their curricula. We met with these teachers in informal workshops and asked them to tell us about the content, methodology, and theory of their current CVS curricula. After informal discussions with the teachers about these issues, we visited their classrooms to conduct classroom observations to get firsthand exposure to the way in which teachers' theories and objectives for teaching CVS actually materialized during classroom instruction.

With this background, we began to craft a lesson plan based on the methodology used by Chen and Klahr (1999). In designing the lesson plan and its associated assessments, we were guided by the following questions: (a) Can fourth graders learn and transfer CVS when participating in expository instruction in a collabora-

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 experimenter, students in the classroom worked in groups of 3 to 5 students. (Assignment of students to groups was determined by the teachers on the basis of the ability of the different students to work together.) The groups were not differentiated based on students' general ability. Each group had its own set of materials with which to work.

Because the teacher could not keep track of the experimental setups of all the groups, we transferred this responsibility to the students. We provided students with worksheets that included a preformatted table representation to record ramp setups (see Appendix B). The method of filling out this table and the rest of the questions on the worksheet were explained to the class before experimentation. Thus, although students had to record the way in which they set up each pair of ramps, they did not have the additional responsibility of devising their own representations for the ramp setup. However, they did have to negotiate the mapping between the physical ramp setup and the tabular representation of that setup. They received detailed instruction on how to do this. During the classroom work, only the ramps domain was used. (The sinking and springs domains were used in the individual interviews before and after

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).
Materials	Ramps or springs or sinking	Ramps only during classroom work (springs and sinking during interviews)
Cognitive mechanism targeted	Analogical transfer	Analogical transfer and representational transfer with interpretive use of experimenter-provided representation
Pragmatic constraints		
Timing	Two 45-min sessions—during or after school	Four 45-min science classes
Teacher	Outside experimenter	Regular science teacher
Student grouping	Individual students	Entire classroom—organized into five groups of 3 to 4 students
Teacher–student ratio	1 : 1	1 : 20
Record keeping	By experimenter—not available for students	By students in experimenter-designed data sheets
Assessment	Domain knowledge test Experimenter's written record of comparisons made by students during individual interviews Videotaped record of student's answers to questions about comparisons during individual interviews with subset of subjects	Domain knowledge test Experimenter's written record of comparisons made by students during individual interviews Videotaped record of student's answers to questions about comparisons during individual interviews with subset of subjects Students' written records of comparisons made and responses given during classroom work Paper-and-pencil pre- and posttests for all students in participating classes

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).

Classroom Assessment

Although the method of assessment in the classroom was derived from assessments developed for the laboratory study, it was implemented slightly differently, and it also included new measures and analyses. Students' ability to compose correct experiments was measured in both the interviews and in the classroom work from the experimental comparisons they made (Table 1). In the laboratory study, this took place in a dialogue format between the experimenter and the individual student, in which the student composed experimental comparisons and the experimenter asked the student probe questions at the beginning and end of each experiment. In the classroom study, this measure was also derived from the students' worksheets (discussed previously). Experiment justification and certainty responses were coded from the videotaped interviews. During the classroom work, students recorded their ramp setups and indicated the certainty of their conclusions on their individual laboratory worksheets. However, students' experiment justification ability was not recorded on the classroom worksheets because we wanted to keep the timing and complexity of students' activities as similar as possible to those in the laboratory study. In both environments (the laboratory and the classroom), students were tested for their domain knowledge prior to and after instruction. In the laboratory study, this test was conducted as part of the dialogue between the experimenter and individual students. In the classroom, each student filled out a paper-and-pencil, forced-choice test to indicate their domain knowledge. A paper-and-pencil experiment evaluation test—based on the posttest used by Chen and Klahr (1999)—was given before and after instruction in the classroom. A detailed description of this new assessment is given subsequently.

METHODS OF CLASSROOM STUDY

Participants

Seventy-seven students from four fourth-grade classrooms in two demographically similar private elementary schools in southwestern Pennsylvania participated. One school was coeducational, and the other school was for girls. The participants included 50 female and 27 male students. Schools were selected from among the set of schools represented in our small teacher network on the basis of several pragmatic factors, including permission of school authorities, teacher interest and avail-

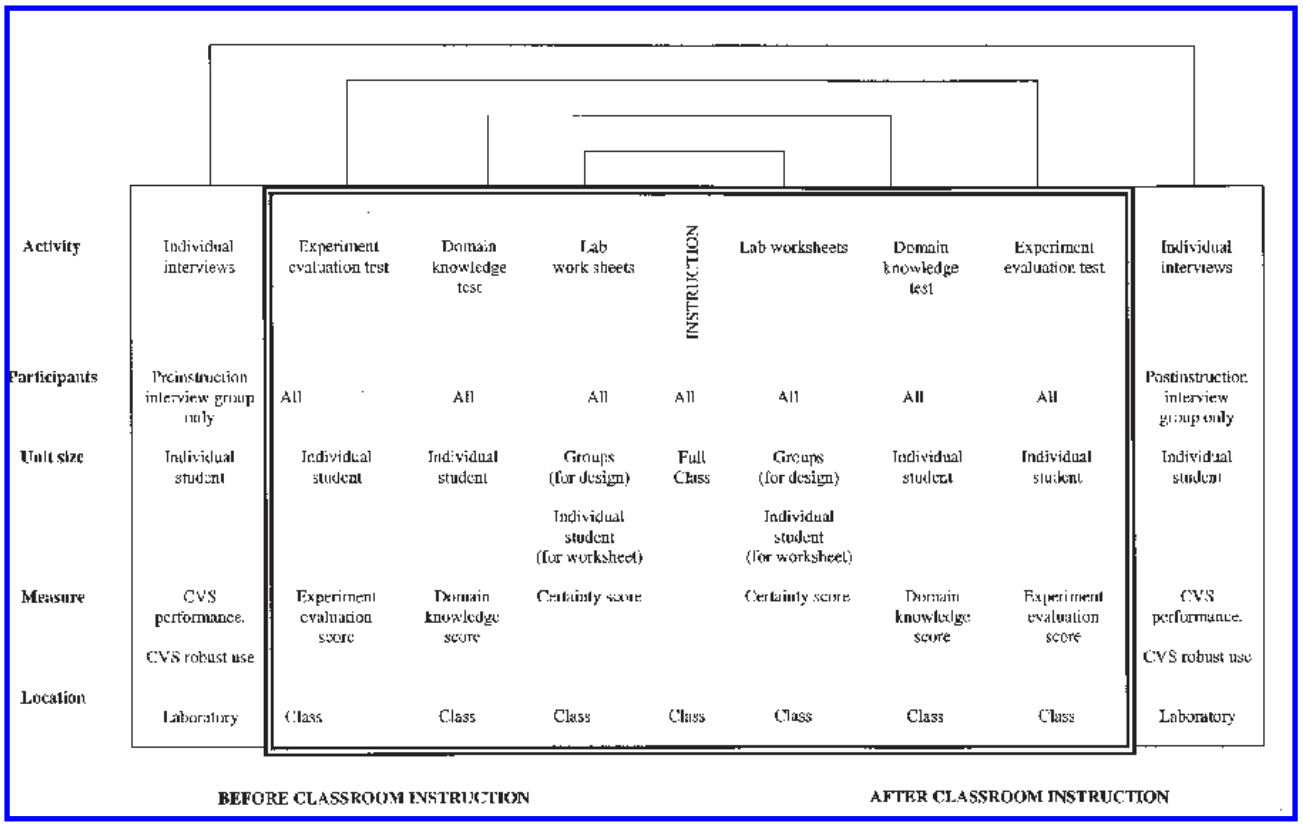

FIGURE 2 Schedule of various assessments before and after classroom instruction. All activities inside the outlined box took place in the classroom.

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 classroom study (31 girls and 12 boys, *M* age = 10 years old).

Research Design

The research design included a set of nested preinstruction and postinstruction measures (Figure 2). The "inner" set of evaluations—depicted inside the double bordered box in Figure 2—used several assessment methods, including a paper-and-pencil experiment evaluation test, a domain knowledge test, and students' written records of the experiments they conducted. These evaluations were administered by the teacher to all students, in class, before and after the instructional session. The "outer" set of assessments consisted of individual interviews.

Procedure

Individual Interviews

The initial and final assessments were based on individual interviews using the procedure similar to Part I of Chen and Klahr (1999). The pragmatics of conducting research in schools shaped the design of this outer evaluation because we could conduct the individual interviews only with those students for whom we had received parental permission. Twenty-one of the 43 volunteer students were randomly assigned to the individual preinstructional interview group and were interviewed before the classroom activities began. The rest were assigned to the individual postinstructional interview group and were interviewed after the classroom activities had been completed.[3] These individual pre- and postinstructional interviews—conducted outside the classroom in a separate room—included students' hands-on design of valid experiments, verbal justifications for the design, and the conclusions students drew from these experiments. One half of the students in both the pre- and postinstruction individual interview groups were randomly assigned to be assessed with springs and the other half with sinking objects.

[3]Because we wanted to avoid any potential reactivity between the individual assessments and students' response to the classroom instruction, we included only one half of the "permission" students on the individual pretest and the other half on the individual posttest. The assumption was that in each case, these students were representative of the full classroom and that there would be no reactivity. Subsequent analyses, reported later, supported both assumptions.

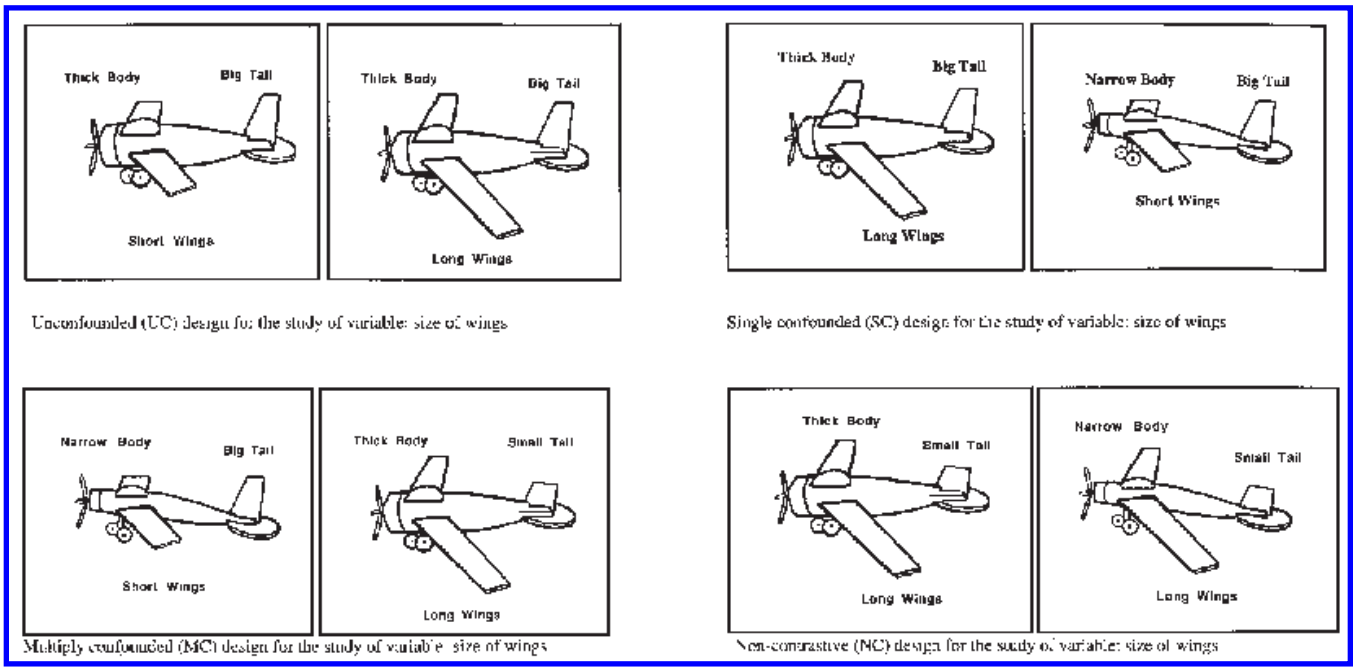

FIGURE 3 Comparison types used in experiment evaluation assessment booklet.

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. Students were asked to evaluate these comparisons by circling the words *bad test* or *good test* based on their judgment of whether the picture pairs showed a correct experiment to determine the effect of the focal variable. (Only unconfounded comparisons are good tests; all others are bad.) The experiment evaluation assessment was given before and after classroom instruction (Figure 2).

Classroom Activities

Classroom activities were led by the regular classroom teacher, with the first author in the role of nonparticipant observer. The teacher began with a short demonstration of the different ways in which the ramps can be set up and an explanation of how to map these ramp setups into the tables on the students' laboratory worksheets (described subsequently). Following the demonstration, there was a short (5 min) paper-and-pencil domain knowledge test to assess students' prior beliefs about the role of different variables.

The next phase of classroom work consisted of what we call *expository instruction* combined with exploration and application. This method of instruction consists of three parts: (a) exploratory experiments conducted in small groups, (b) whole classroom expository instruction, and (c) application experiments conducted in small groups. In essence, this methodology is what Lawson, Abraham, and Renner (1989) described as the *learning cycle,* although our instruction is a one-time event focusing on a procedural skill rather than repeated cycles to learn conceptual knowledge over an extended time.

Exploratory experiments. At the beginning of classroom work, students were asked to explore what makes balls roll further down ramps by experimenting. They conducted four different experiments—two to test each of two different variables. The students decided how to set up their ramps to make a good comparison to test whether each focal variable makes a difference in how far a ball will roll down the ramp. Students individually recorded their experimental setups and data into preformatted worksheets that had two sections for each experiment (Appendix B). The first section asked students to map their ramp setup into a table representation, and the second section included questions about the outcome of each experiment and about whether the experiment conducted would allow students to draw definite conclusions about the role of the current focal variable. Students were asked (a) "Does the [variable] make a difference? Circle your answer: Yes–No"; and (b) "Think about this carefully; can you tell for sure from this comparison whether the [variable] makes a difference? Circle your answer: Yes–No." During the classroom work, the students were not asked to provide explanations for the answers recorded in their worksheets. These four experiments, conducted in the exploration phase of classroom work, provided an additional baseline measure of students' preinstruction knowledge of CVS. While students conducted these exploratory experiments, the teacher's role was to facilitate group work and individual reflection. The teacher reminded the class to "make sure all team members agree on the setup before you roll the balls," "think about your answers carefully," and "record your thinking about this experiment individually." The teacher did not provide corrective feedback regarding CVS during exploratory experimentation.

Expository instruction. The second stage of classroom work included about 20 min of instruction to the entire class on how to create controlled experiments. The teachers followed these five steps:

1. *Initiate reflective discussion*. This was based on a bad comparison—a multiply confounded comparison between two ramps. After setting up this bad test, the teacher asked students whether it was a good or bad test. Rather than accepting a simple short answer (yes or no), she asked students to explain their beliefs. She provided time and opportunity for students' different views and, often conflicting, ex-

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, $F(2, 82) = 5.8$, $p = .005$ (postinterview scores were near ceiling and did not show this trend). Thus, prior to instruction, there was some "spontaneous" improvement in the CVS performance score during the interviews, although these scores remained far below children's ultimate levels of performance following expository instruction.

Next, we looked at the difference between the CVS scores from the preinstruction individual interviews and the postinstruction individual interviews. There was a dramatic increase from a mean of 2.7 out of 9 (30%) prior to instruction to a mean of 8.7 (97%) after instruction, $t(41) = 12.3$, $p < .0001$. With respect to individual students' performances, we defined as a *consistent user of CVS* any student who correctly used CVS on at least eight of the nine trials in the individual interviews. Only 1 of the 21 children (5%) in the individual pretest interviews met this definition, whereas 20 of the 22 children (91%) in the posttest individual interviews exhibited consistent performance, $\chi^2(1, N = 43) = 31.9$, $p < .0001$.

Robust CVS Use

Mean scores for robust CVS use (a measure indicating students' ability to design an unconfounded experiment and provide a CVS rationale) increased from a mean of .57 out of 9 (6.3%) in the preinterview to a mean of 7.0 (78%) in the postinterview, $t(41) = 11.7$, $p < .0001$. None of the 21 students interviewed prior to instruction was a consistent robust CVS user (i.e., able to create controlled experiments and justify their designs with a CVS rationale on at least eight of nine trials). After instruction, 12 of the 22 students interviewed (55%) were consistent robust CVS users, $\chi^2(1, N = 43) = 15.9$, $p < .0001$ (Table 2).

Following the same rationale we used when analyzing students' CVS performance score, we tested the possibility that the guided discovery inherent in our repeated experiments and interviewer questioning would lead to an increase in robust CVS use, prior to any instruction. However, no such increase was found. Thus, unlike the CVS performance scores, robust CVS scores did not increase during the preinstructional trials.

TABLE 2
Proportion of Consistently Good CVS-Performers and Consistently Good CVS-Robust-Users Prior to and After Instruction During Individual Interviews

Performers and Users	*Preinstruction*	*Postinstruction*[a]
Consistently good CVS-performers	1/21	20/22
Consistently good CVS-robust users	0/21	12/22

[a]The students in the preinstruction and postinstruction conditions were not the same students.

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 group were recording the same experiment.

Mean CVS performance scores for each group increased significantly, from a mean of 2.32 (58%) before instruction to a mean of 3.86 (97%) after instruction, $t(21) = 4.2$, $p < .0004$. An additional t test comparing mean scores for the first and last experiment conducted prior to instruction revealed no significant increase, that is, learning CVS by experimentation alone did not occur in the classroom during group work.

The consistency of CVS performance (correct design of all four experiments) increased from 45% of groups prior to instruction to 91% of groups after instruction. A chi-square test indicated a significant difference in the number of groups who were consistent CVS performers prior to and after instruction, $\chi^2(1, N = 44) = 10.48$, $p < .0012$. Because the student worksheets did not ask for individual justification of experiments, the classroom data provide no measure of robust CVS use.

Analysis of Domain Knowledge Test

An important aspect of our expository instruction was that it focused on the acquisition of the domain-general strategy of controlling variables and forming correct justifications for the validity of controlled experiments. At no point was there any expository instruction regarding the role of the causal variables in the ramps domain. However, we found a significant increase between preinstruction and postinstruction domain knowledge scores. Whereas 79% of the students provided correct answers to all three questions on the domain knowledge test prior to CVS instruction, 100% of the students got the maximum score after instruction, $t(71) = 4.3$, $p < .0001$.

TABLE 3
Summary of CVS Performance and Expertise, Robust CVS Use and Expertise, Experiment Evaluation, and Expertise and Domain Knowledge From Individual Student Records

Source of Individual Data	*Measure*	*Before Instruction*	*After Instruction*
Interviews[a]	Performance score (proportion correct)	0.30	0.97
	Proportion of consistent performers (correct on at least 8 out of 9 experiments)	0.05	0.91
	Robust CVS use score (proportion correct)	0.06	0.78
	Proportion of consistent robust-CVS users (correct on at least 8 out of 9 experiments)	0.00	0.55
Experiment evaluation test[b]	Experiment evaluation score (proportion correct)	0.61	0.87
	Proportion of consistently good evaluators (correct on at least 9 out of 10 experiments)	0.28	0.76
Domain knowledge test[c]	Mean domain knowledge score	0.79	1.00

[a]$N = 21$ before and $N = 22$ after. The students in the preinstruction and postinstruction interviews were not the same; [b]$N = 74$. This includes students who participated in pre- and postinstructional individual interviews; [c]$N = 75$ before and $N = 73$ after.

Analysis of Experiment Evaluation Test

Students' ability to evaluate experiments designed by others increased significantly following classroom instruction. Students were presented with a 10-item experiment evaluation test containing four different types of comparisons to judge (Figure 3). Mean experiment evaluation scores increased from 61% correct prior to instruction to 87% correct after instruction, $t(70) = 8.72$, $p < .0001$. The percentage of students who were consistently good evaluators of experiments designed by others, that is, those who could evaluate correctly at least 9 of the 10 comparisons, increased from 28% prior to instruction to 76% after instruction, $\chi^2(1, N = 74) = 31.2$, $p = .001$. Thus, a brief period of expository instruction significantly increased students' ability to evaluate the validity of experiments designed by others. Table 3 provides a summary of the effects of expository instruction on individual students' CVS performance, robust CVS use, domain knowledge, and experiment evaluation scores.

To ascertain that there was no effect of the interviews conducted prior to the experiment evaluation test—and to ascertain that the interviewed students were representative of the entire sample—we analyzed the data for possible differences between the pre- and postinterview subgroups. None of our measures conducted after individual interviews (CVS performance, certainty, domain knowledge, and evaluation mean scores) revealed any significant differences between interviewed students and the rest of the sample.

Surprises and New Issues for Teaching and Learning Scientific Inquiry Skills

The previous sections detailed the results of the translation between the laboratory and classroom learning environments and documented the effectiveness of our expository instruction methodology in the classroom setting. However, these sections projected a rather unidirectional view of our movement between the laboratory and the classroom. Our attempts to adapt the laboratory procedures to produce a research-based classroom learning environment provided a stark reminder that the classroom is quite different from the highly controlled laboratory (Klahr, Chen, & Toth, in press). While considering the instructional, practical, and assessment constraints of the classroom environment (as described earlier) and during the analysis of classroom data, we found a few new challenges and issues ripe for extended study in both the laboratory and the classroom. In the classroom, scientific experimentation usually happens in groups, and the thinking strategies and reasoning skills of students in these groups can be quite diverse. In the following section, we detail two of our findings that have the potential to affect the teaching of scientific inquiry skills: (a) individual student's strategies of experimentation and (b) certainty of students' inferences after valid experiments.

Students' Strategies of Experimentation

The experiment evaluation test, administered both prior to and after classroom work, measured students' ability to evaluate four different types of experiments designed by others: unconfounded, singly confounded, multiply confounded, and noncontrastive. Although an analysis of variance with preinstruction and postinstruction mean scores on the four problem types as repeated measures showed that students' mean scores increased significantly after instruction, $F(1, 70) = 76.5$, $p < .0001$, there was also a main effect of problem type, $F(3, 210) = 15.2$, $p < .0001$, as well as a significant interaction between time of assessment and problem type, $F(3, 210) = 10.3$, $p < .0001$.

Prior to classroom instruction, students were more successful in correctly evaluating unconfounded ($M = .83$, $SD = .29$) and noncontrastive ($M = .69$, $SD = .36$) designs than single confounds or multiple confounds ($M = .46$, $SD = .41$ and $M = .58$, $SD = .41$, respectively). All pairwise comparisons were significant, $p < .05$. On the postinstruction experiment evaluation test, students still had highest mean scores on unconfounded ($M = .91$, $SD = .22$) and noncontrastive ($M = .93$, $SD = .24$) designs and lowest mean scores on singly confounded ($M = .85$, $SD = .32$) and multiply confounded problems ($M = .85$, $SD = .34$). However, the difference between unconfounded and noncontrastive problems was not significant. Perfor-

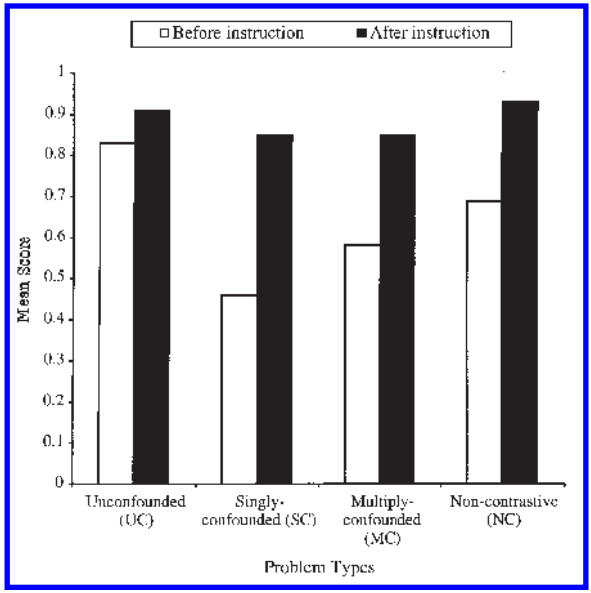

FIGURE 4 Students' ability to evaluate four different experimental designs.

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

unconfounded and noncontrastive problems as correct and those experiments that differ in more than one variable as incorrect.

4. *Vary at least one variable and ignore others.* Students using this strategy do not recognize the indeterminacy of confounded experiments (Fay & Klahr, 1996) and judge all comparison types as correct.

5. *Other.* This category was added as we hypothesized that there are additional strategies that our current measurement instrument did not allow us to detect, such as "vary all," "all good," or "random" strategies. Further modifications to our experiment evaluation instrument will help us specify the various strategies within this general category.

To determine the strategy used by each student, we matched all the answers each student gave on the experiment evaluation test to the answers expected with the use of each of the four strategies. We determined the best fit between our hypothesized strategies and the actual strategies students applied by giving a score of 1 on each of the 10 evaluation items of the test if a positive match between one of our hypothesized strategies and the student's answer on that test item could be determined. We assigned a score of zero to each nonmatch between students' actual response and the response yielded by the hypothesized strategies. We scored a student as using one of our hypothesized strategies if the student's responses matched at least 8 of 10 answers yielded by that strategy. There were five instances when a student's scores matched two possible strategies (with a score of 8 on each of two hypothetical strategies). In those cases, we categorized students according to the "weaker" strategy—the one conceptually most distant from correct CVS. We used this methodology to determine the strategy use of students both prior to and after instruction. Table 4 summarizes the various student strategies we identified, the judgment each strategy would yield on the various problem types, and the number of students who were matched to each category both prior to and after instruction.

The analysis of student strategies indicated that, prior to instruction, 27% of the students applied the correct, CVS strategy to evaluate experiments conducted by others, whereas after instruction, 77% of the students used this strategy. Before instruction, 16% of students used the "vary focal, ignore others" strategy, and 8% used the "vary any one, control others" strategy. The "vary at least one variable" strategy, which ignores the roles of both the focal variable and all other variables, was used by 19% of the students prior to instruction. The number of students applying strategies other than CVS dramatically decreased by the end of classroom work; however, 23% of students still employed incorrect strategies. We were not able to match the strategies of 30% of the students in the preinstruction and 9% of the students in the postinstruction test. The distribution of the different strategies was significantly different on the preinstruction and postinstruction experiment evaluation tests, $\chi^2(1, N = 77) = 32.6, p = .0001$.

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).

CONCLUSIONS

The research reported in this article was motivated by the current debate about the possibilities of interfacing research on learning conducted in psychology laboratories with educational practice (Glaser, 1990, 1991; Schoenfeld, 1999; Strauss, 1998). In contrast to the common practice of experimental psychologists and classroom researchers working in separate research groups, the authors of this article brought diverse backgrounds to the project, enabling the development and confirmation of effective instructional methodologies as well as the practical application

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 (classroom) and basic (laboratory) research.

Our experience suggests that the construction of a research-based classroom environment is not a simple one-way, one-step process. Although this article focused on the classroom verification phase of our cycle of studies, three principles emerge from our overall experience of bridging the laboratory and classroom worlds:

1. Translation (not transfer) is the best way to conceptualize the method of moving between the laboratory and classroom worlds. The process is not a straightforward transfer from the laboratory to the classroom (Klahr et al., in press), and research-based instructional practice can only be built by considering the constraints of classroom environment.
2. Bidirectional information flow is needed to establish the reciprocity of learning between the two environments. Both the laboratory and classroom environments have strengths in yielding research results on which subsequent interventions in both environments should build.
3. Iterative cycles of research and development in the form of multiyear, multiphase efforts are necessary to establish long-lasting results.

The systematic application of these three principles has aided us in our attempts to design a classroom environment that considers the needs and constraints of practitioners and incorporates up-to-date results from relevant educational and psychological research. As our research cycle in both worlds continues, we expect further refinement of these three principles of interfacing the worlds of research on learning and educational practice.

ACKNOWLEDGMENTS

Zhe Chen is now at the University of California, Davis.

This research was funded by the James S. McDonnell Foundation, CSEP program. We express our gratitude toward our teacher colleagues, Mrs. Linda Cline

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 superficial, structural and procedural features. *Journal of Experimental Child Psychology, 62,* 410–431.

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

Christensen, C. A., & Cooper, T. J. (1991). The effectiveness of instruction in cognitive strategies in developing proficiency in single-digit addition. *Cognition and Instruction, 8,* 363–371.

Coppola, R. K., & Toth, E. E. (1995). EarthVision: Teachers and students learn to use high performance computing together. In J. Willis, B. Robin, & D. Willis (Eds.), *Technology and teacher education annual* (pp. 191–196). Charlottesville, VA: Association for the Advancement of Computers in Education.

Das-Smaal, E. A., Klapwijk, M. J., & van det Leij, A. (1996). Training of perceptual unit processing in children with a reading disability. *Cognition and Instruction, 14,* 221–250.

diSessa, A. A., & Minstrell, J. (1998). Cultivating conceptual change with benchmark lessons. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp.155–188). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fay, A. L., & Klahr, D. (1996). Knowing about guessing and guessing about knowing: Preschoolers' understanding of indeterminacy. *Child Development, 67,* 689–716.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education, 27,* 403–434.

Glaser, R. (1990). The reemergence of learning theory within instructional research. *American Psychologist, 45,* 29–39.

Glaser, R. (1991). The maturing of the relationship between the science of learning and cognition and educational practice. *Learning and Instruction, 1,* 129–144.

Klahr, D. (Ed.). (1976). *Cognition and instruction.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes.* Cambridge, MA: MIT Press.

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 thought. *Child Development, 47,* 697–706.

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, 60*(3, Serial No. 245).

Lawson, A. E., Abraham, M. R., & Renner, J. W. (1989). A theory of instruction: Using the learning cycle to teach science concepts and thinking skills. *National Association for Research on Science Teaching Monograph* (No. 1).

Levin, S. R., Toth, E. E., & Douglas, C. (1993). Earth day treasure hunt: Developing successful activities on electronic networks. In G. Davies & B. V. Samways (Eds.), *Teleteaching* (pp. 557–562). New York: Elsevier.

Masnick, A. M., & Klahr, D. (2000). *Elementary school children's understanding of error in science experimentation.* Manuscript submitted for publication.

National Research Council. (1996). *National Science Education Standards.* Washington, DC: National Academy Press.

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.

Schoenfeld, A. (1999, April). *Are we ready for the 21st century? Notes on research methods, the preparation of new researchers and fundamental research questions.* Presidential address presented at the 1999 annual convention of the American Educational Research Association, Montreal, Canada.

Strauss, S. (1998). Cognitive development and science education: Toward a middle level model. In W. Damon (Series Ed.) and I. Sigel & K. A. Renninger (Vol. Eds.), *Handbook of child psychology: Vol. 4. Child psychology in practice* (5th ed., pp. 357–400). New York: Wiley.

Toth, E. E. (2000). Representational scaffolding during scientific inquiry: Interpretive and expressive use of inscriptions in classroom learning. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 953–958). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Toth, E. E., & Klahr, D. (2000, April). *Error errors: Children's difficulties in applying valid experimentation strategies in inquiry based science learning environments.* Poster presented at the annual convention of the American Educational Research Association, New Orleans, LA.

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2000). *Mapping to know: The effects of representational guidance and reflective assessment on scientific inquiry skills.* Manuscript submitted for publication.

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

APPENDIX A

Table A1 is from "All Other Things Being Equal: Children's Acquisition of the Control of Variables Strategy," by Z. Chen and D. Klahr, 1999, *Child Development, 70,* p. 1102.

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

APPENDIX B
Preformatted Recording Sheets Used in the Classroom

From "All Other Things Being Equal: Children's Aquisition of the Control of Variables Strategy," by Z. Chen and D. Klahr, 1999, *Child Development, 70,* p. 1103. Copyright 1999 by the Society for Research in Child Development. Reprinted with permission.