Child Development, September/October 1999, Volume 70, Number 5, Pages 1098–1120

All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy

Zhe Chen and David Klahr

The ability to design unconfounded experiments and make valid inferences from their outcomes is an essential skill in scientific reasoning. The present study addressed an important issue in scientific reasoning and cognitive development: how children acquire a domain-general processing strategy (Control of Variables Strategy or CVS) and generalize it across various contexts. Seven- to 10-year-olds ($N = 87$) designed and evaluated experiments and made inferences from the experimental outcomes. When provided with explicit training within domains, combined with probe questions, children were able to learn and transfer the basic strategy for designing unconfounded experiments. Providing probes without direct instruction, however, did not improve children's ability to design unconfounded experiments and make valid inferences. Direct instruction on CVS not only improved the use of CVS, but also facilitated conceptual change in the domain because the application of CVS led to unconfounded, informative tests of domain-specific concepts. With age, children increasingly improved their ability to transfer learned strategies to remote situations. A trial-by-trial assessment of children's strategy use also allowed the examination of the source, rate, path, and breadth of strategy change.

INTRODUCTION

The ability to design unconfounded experiments and make valid inferences from their outcomes is an essential skill in scientific reasoning. An important issue in cognitive development is whether early elementary school children are capable of understanding the logical basis underlying the processes used to create and interpret unconfounded experiments and how they learn and generalize this strategy across various domains.

In this article, we focus on one such domain-general strategy that we call the "Control of Variables Strategy" (CVS). We define CVS in both procedural and logical terms. Procedurally, CVS is a method for creating experiments in which a single contrast is made between experimental conditions. The full strategy involves not only creating such contrasts, but also being able to distinguish between confounded and unconfounded experiments. The logical aspects of CVS include the ability to make appropriate inferences from the outcomes of unconfounded experiments as well as an understanding of the inherent indeterminacy of confounded experiments. In short, CVS is the fundamental idea underlying the design of unconfounded experiments from which valid, causal, inferences can be made. Its acquisition is an important step in the development of scientific reasoning skills because its proper use provides a strong constraint on search in the space of experiments (Klahr, 1999; Klahr & Dunbar, 1988).

Previous studies present a mixed picture of the extent to which elementary school children can understand and execute CVS. A recent study by Kuhn, Garcia-Mila, Zohar, and Andersen (1995) revealed that late elementary schoolchildren have only a fragile grasp of the concepts and skills that underlie the logic of CVS. In a variety of scientific discovery tasks, participants explored the effects of several variables on different outcomes. Even after 20 sessions spread over 10 weeks, fewer than 75% of adults' and 25% of fourth graders' inferences were valid. Other studies also have shown similar findings (e.g., Bullock & Ziegler, 1999; Schauble, 1996). Given that CVS is a fundamental scientific reasoning skill and given that few elementary schoolchildren spontaneously use it when they should, it is important to know whether there are any effective ways to teach CVS and whether age and instructional method interact with respect to learning and transfer.

Despite the centrality of this question, only two previous studies (Case, 1974; Kuhn & Angelev, 1976) shed any light on it, but neither of them was designed to *directly* address the issues raised in the present study. Case's pioneering work indicated that the majority of field-independent 7- and 8-year-olds could be taught to use CVS after a four-session training period spread over 4 weeks. In the Kuhn and Angelev study, fourth and fifth graders improved their reasoning about CVS after being exposed to problems requiring variable control and systematic combination of variables during a 15-week intervention program. Children who received explicit instructions, how-

ever, did not perform better than those given exposure alone.

These two early studies leave us with many unanswered questions: (1) What are the relative contributions of direct instruction, "mere exposure," and probe questions in facilitating children's understanding of CVS? (2) How well can early elementary school children use CVS in designing experiments? (3) Do children at various ages learn and transfer CVS differently? (4) Does training in CVS lead to an improvement in domain knowledge? (5) What is the nature and extent of individual differences in learning—particularly with respect to the initial emergence and stable use of a newly acquired strategy? These questions are the foci of the present study.

Thus, the primary aim of this research was to determine the conditions under which children can learn CVS. Whether, how, and when children acquire scientific reasoning strategies such as CVS are critical issues whose resolution has important implications for both cognitive development and instruction. A second aim was to determine the extent to which, once taught, children can transfer CVS to situations beyond the specific context in which they acquired the strategy. For example, after learning how to design unconfounded experiments to determine various factors in the stretching of springs, are children able to utilize CVS in creating valid experiments dealing with balls rolling down ramps? Or do they apply the strategy only to situations that share very similar features? Previous studies indicate that, when solving analogous problems, younger children tend to focus on perceptual or superficial features in mapping a source problem to a target problem, whereas older children are better able to override superficial similarities between problems and perceive structural similarity (e.g., Gentner, 1989; Holyoak, Junn, & Billman, 1984). Other studies, however, suggest that young children are not alone in their difficulties to perceive and map analogous relations between problems; adults also fail to apply useful source solutions in solving target problems (e.g., Reed & Bolstad, 1991; Ross & Kilbane, 1997).

Most studies on analogical reasoning focus on children's ability to apply specific solutions to isomorphic problems (e.g., Brown, 1989, 1990; Gick & Holyoak, 1983; Ross, 1989). In our study, the isomorphism between problems was at a deep conceptual level, based on the underlying logic of CVS, whereas the surface features were designed to correspond to three levels of transfer distance. In the present work, as in most discussions of transfer "distance," the metric is undefined, although the relative ordering is unambiguous. *Very Near Transfer* is defined as the application of CVS to test a new aspect of the same materials used in the original learning problem. *Near Transfer* is defined as the use of CVS to solve problems using a set of different materials that are still in the same general domain as the original problem. *Remote Transfer* refers to the application of CVS to solve problems with domains, formats, and context different from the original training task after a long delay. We expected that most children would be able to use CVS when the transfer distance was relatively near but that as transfer distance increased, children would experience more difficulty.

Our third aim was to contrast discovery learning with two levels of instruction on the acquisition of CVS. Discovery learning has been considered an effective approach for the acquisition of domain-specific knowledge. Its advocates argue that children who are actively engaged in the discovery of a strategy are more likely to be successful in applying it than those who passively receive direct instruction (e.g., Jacoby, 1978; McDaniel & Schlager, 1990). On the other hand, explicit instruction may be necessary when a strategy is difficult for children to discover by themselves. For example, Klahr and Carver (1988) found that a brief period of explicit instruction in how to debug computer programs was more effective than hundreds of hours of discovery learning. The relative impact of these different approaches might depend on the content of the learning tasks. Discovery learning might be effective when problem outcomes provide informative feedback (e.g., Siegler, 1976). Unguided experimental designs, however, typically do not provide informative feedback concerning their quality. This lack of feedback might render the discovery of process skills such as CVS particularly difficult for early elementary schoolchildren.

We focused on two types of instruction: *explicit* training (using examples and direct instruction to teach the general strategy) and *implicit* training via probes (providing systematic questions following children's activities) in hands-on experimentation in which extensive and repeated opportunities to use the strategy were provided. We expected these two types of instruction to yield different levels of learning and transfer. Whereas probe questions alone might not be adequate to foster the acquisition and use of CVS, the combination of both explicit instruction and probe questions might be an effective approach to promote the mastery of CVS.

Our fourth aim was to explore developmental differences in acquisition and transfer as manifested in the interaction between age and type of instruction. Explicit instruction might be effective for both younger and older children. Younger children might

not benefit from probe questions alone, however, whereas older children might already possess implicit knowledge concerning CVS, with the result that systematic questioning might be sufficient to elicit the use of the strategy. In addition to developmental differences in the relative effectiveness of different instructional methods, there might be developmental differences in children's transfer abilities. Previous studies have demonstrated that even young children can detect analogous relations between isomorphic problems and transfer learned solutions (e.g., Brown & Kane, 1988; Goswami, 1991, 1996). Other studies have shown that older children tend to be better able to override superficial dissimilarities and focus on structural features when mapping from one domain to another (e.g., Chen 1996; Chen & Daehler, 1992; Gentner & Toupin, 1986). We predicted that, although early elementary schoolchildren might prove capable of transferring CVS, only older children would display the ability to transfer the strategy to remote situations.

Our fifth aim was to examine the learning process by employing a microgenetic method. Microgenetic methods offer an effective approach for exploring the change process by providing detailed data concerning how new approaches are discovered and how children generalize the new strategies after their initial discovery (Siegler & Crowley, 1991; Siegler & Jenkins, 1989). The microgenetic approach typically involves a span of observation from the initial use of a strategy to its consistent use, as well as intense analyses of qualitative and quantitative changes (Bjorklund, Coyle, & Gaultney, 1992; Siegler, 1996). Because the method utilizes trial-by-trial assessments of ongoing cognitive activities, it facilitates precise analyses of how children change their strategies with experience and with instruction. This design allowed detailed analyses of children's learning of CVS in different conditions.

The sixth and final aim of this study was to determine whether the use of a domain-general strategy such as CVS would result in increased domain-specific knowledge (e.g., knowledge about springs or ramps). Although most research on children's scientific reasoning has tended to focus on either conceptual change (e.g., Carey, 1985; Chi & Ceci, 1987; Vosniadou & Brewer, 1992; Wellman & Gelman, 1992, 1998) or inference processing (e.g., Dunbar & Klahr, 1989; Fay & Klahr, 1996; Kuhn, Amsel, & O'Loughlin, 1988), a few studies (Schauble, 1996; Schauble, Glaser, Duschl, Schulze, & John, 1991) have focused on *both* aspects of scientific reasoning and on the mutual influence of changes in domain-specific knowledge and domain-general reasoning process skills. Whereas domain-specific knowledge would affect reasoning process strategies (e.g., unconfounded designs and valid inferences) by influencing both hypothesis and experiment spaces (Klahr & Dunbar, 1988), reasoning process skills in turn influence the acquisition of domain-specific concepts. Nevertheless, little is known about this interaction. The specific issue addressed in this study was whether the acquisition of CVS would contribute to children's acquisition of domain-specific information. We predicted that children who were trained to use CVS would be more likely to design unconfounded tests and make valid inferences about the outcomes of the tests and would therefore acquire more domain-specific content. In contrast, children who were not trained to use CVS and who designed mainly confounded experiments in the nontraining conditions would be less likely to change their initial understanding of the domain, because the results of their experiments are typically uninformative, at best, and misleading, at worst.

In summary, the present research was designed (1) to determine whether early elementary schoolchildren can gain a genuine understanding of CVS in a context that requires them to design unconfounded tests and make valid inferences, (2) to determine the extent to which children transfer a learned strategy, (3) to examine what type of training would be most effective for both learning and transfer, (4) to explore possible developmental differences in the learning and transfer of CVS in elementary schoolchildren, (5) to examine the rate, path, and breadth of strategy change under various conditions, and (6) to explore the relations between strategy use and the acquisition of domain-specific knowledge.

Children within this age range were selected because few studies have explored the understanding and use of scientific reasoning strategies in children at this age level. Previous research (e.g., Klahr, Fay, & Dunbar, 1993; Kuhn, Schauble, & Garcia-Mila, 1992; Penner & Klahr, 1996) suggested that fifth or sixth graders begin to develop strategies for generating experiments and interpreting test outcomes in multivariable contexts and that, although early elementary schoolchildren typically do not use CVS spontaneously when designing tests, they may well possess the ability to understand the rationale of the strategy and to transfer it across problems in moderately complex domains. Moreover, previous studies suggested that children's abilities to understand and apply CVS might undergo rapid changes during this period, thus allowing the examination of developmental differences in the use and application of CVS. In addition, studying children at these ages fills the gap between the research with older children (e.g., Kuhn et al., 1995; Schauble, 1996) and younger ones such as 5-

and 6-year-olds (Fay & Klahr, 1996; Sodian, Zaitchik, & Carey, 1991).

The present study consisted of two parts. Part I included hands-on design of experiments. Children were asked to set up experimental apparatus so as to test the possible effects of different variables. The hands-on study was further divided into four phases. In Phase 1, children were presented with materials in a source domain in which they performed an initial exploration followed by (for some groups) training. Then they were assessed in the same domain in Phase 2. In Phases 3 and 4, children were presented with problems in two additional domains (Transfer-1 and Transfer-2). Part II was a paper-and-pencil posttest given 7 months after Part I. The posttest examined children's ability to transfer the strategy to remote situations. Children who had participated in Part I and an equivalent number of their classmates who had not were given a posttest in which they were asked to solve problems in five domains—all different from the Part I domains—that involved evaluating whether each of a series of paired comparisons was a good test of the effect of a specific variable. The Part II tasks differed from the earlier problems in several important respects, including contextual differences (different "experimenters" and different settings in which tests were administered) and task dissimilarities (different formats: hands-on versus pencil and paper; strategy generating versus evaluating and different content: mechanical versus other types of problems).

METHOD

Part I: Hands-On Study

Participants

Participants were 87 second (*mean* age = 7,10), third (*mean* age = 9,0) and fourth (*mean* age = 9,9) graders (57 girls and 30 boys) from two private elementary schools in southwestern Pennsylvania. The students were volunteers recruited through a mailing to the parents of children in these grades. Children in each grade were randomly assigned to one of three conditions. The mean ages for the Training–Probe, No Training–Probe, and No Training–No Probe conditions were 9,1 (n = 30), 9,8 (n = 30), and 9,3 (n = 27), respectively.

Design

We used a 3 (condition) × 3 (grade) × 4 (phase) design with phase as a within-subjects measure. Children were asked to make a series of paired comparisons to test particular variables of each problem in four phases of the study: Exploration, Assessment, Transfer-1, and Transfer-2. In each phase, participants were asked to make comparisons in one task to find out whether or not a variable made a difference in the outcome (e.g., make a comparison that shows whether the length of a spring makes a difference in how far it stretches). Conditions differed in whether children received explicit instruction in CVS, and whether they received systematic probe questions concerning why they designed the tests as they did and what they learned from the tests. Three isomorphic tasks were used: Spring, Slope, and Sinking. Task order was counterbalanced, as was the order of focal variables within each task.

In the Training–Probe condition, children were given explicit instruction regarding CVS. Training occurred between the Exploration and Assessment phases. 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 was executed asked children to explain why they designed the particular test they did. 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. In the No Training–Probe condition, children received no explicit training, but they did receive the same series of probe questions surrounding each comparison as were used in the Training–Probe condition. Children in the No Training–No Probe condition received neither training nor probes.

Materials

In each of the three tasks, there were four variables that could assume either of two values. In each task, participants were asked to focus on a single outcome that was affected by all 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 weight. Each child worked with one of the three tasks on their first day in the study (Exploration and Assessment phases) and then with two other tasks on their second day (Transfer-1 and Transfer-2). Table 1 summarizes the features of all three hands-on tasks.

Springs task. In the springs task, children had to make comparisons to determine the effects of different variables on how far springs stretch. Materials consisted of eight springs varying in length (long and short), coil width (wide and narrow), and wire diameter (thick and thin). The springs were arranged on a

Table 1 Problem Domains Used in Part I

	Domain		
	Springs	Slopes	Sinking
Primary materials	Eight springs that vary on three variables A frame for hanging two springs Two sets of weights, a heavy pair and a light pair	Two ramps, each with adjustable angle and "starting gate" location Two sets of two balls, golf and rubber (squash) Two two-sided surface inserts (for ramps) with different coefficients of friction	Two water-filled cylinders, with two drop heights indicated Eight objects that vary on three variables Scooper and magnet for retrieving sunken objects
To be determined	What factors determine how far a spring will stretch?	What factors determine how far a ball will roll down a ramp?	What factors determine how fast an object will sink in water?
Variables: 2 independent values for each of 4 variables[a]	• length long, short • coil diameter wide, narrow • wire diameter thick, thin • weight size heavy, light	• angle high, low • starting gate short, long • surface smooth, rough • ball golf, rubber	• shape cube, sphere • material steel, Teflon • size large, small • height high, low
Dependent measure	Length of extension (or distance from base of rack) when weight is added	Distance ball rolls at end of ramp	Speed of sinking in water (or which reaches bottom first)
Subject activity			
Experimental design	From set of 8 springs: • Select 2 springs • Hang springs on rack hooks • Select weights to go with each spring	For each of 2 ramps: • Select one of two angles • One of two surfaces • One of two starting positions • Select one of two balls to run	From set of 8 objects: • Select 2 objects • For each object, select one of two heights from which to drop object
Experiment execution	Hang weights on springs Observe amount of stretching (or distance from base)	Release gates (not necessarily simultaneously), allowing balls to roll Observe distance balls roll after leaving ramp	Simultaneously drop each object into water-filled cylinder Observe relative sink rates (or arrival times at bottom of cylinder)
Notable aspects of domain and procedure	All variables investigated are integral to selected spring Choice is from among pre-existing springs having a "cluster" of variable values Experiment is easy to set up and execute (no timing issues) Measurement is easy (stable outcome)	Variables are independent, object is constructed from choice of values for each variable Comparison objects are constructed; variable values are not clustered Outcome is evanescent (if based on speed), or stable (if based on final distance)	All variables investigated are integral to selected object Choice is from among pre-existing objects having a "cluster" of variable values Easy to set up (simply choose two objects and heights) Simultaneity necessary at start of drop Outcome must be observed instantly, otherwise it is lost

[a] Children were asked to investigate the first three variables listed in each task. The remaining variable was identified by the experimenter at the outset, but the participants were never asked to investigate its effect.

tray such that no pair of adjacent springs made an unconfounded comparison. A pair of "heavy" and a pair of "light" weights also was used. Heavy and light weights differed in shape as well as weight, so that they could be distinguished easily. To set up a comparison, children selected two springs to compare and hung them on hooks on a frame and then selected a weight to hang on each spring. To execute a comparison, participants hung the weights on the springs and observed as the springs stretched. The outcome measured was how far the springs stretched down toward the base of the frame.

Slopes task. In the slope 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 slope task were two wooden ramps, each with an adjustable downhill side and a slightly uphill, stepped surface on the other side. Children could set the steepness of the downhill ramps (steep and low) using wooden blocks that fit under the ramps in two orientations. Children could determine the surface of the ramps (rough or smooth) by placing inserts on the downhill ramps either carpet side up or smooth wood side up. They also could determine how far the balls rolled on the downhill ramp by placing gates at either of two positions different distances from the top of the ramp (long or short run). Finally, participants could choose from two kinds of balls, rubber squash balls and golf balls. To set up a comparison, participants 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 comparison, 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 measured was how far the balls traveled up the stepped side of the ramp. Figure 1 depicts a comparison from the slop task. It is a completely confounded comparison because all four of the variables differ.

Sinking task. In the sinking task, children had to determine which variables affect how fast objects sink in water. Materials for the sinking task consisted of eight objects differing in size (large and small), shape (spheres and cubes), and material (metal and plastic), and two clear cylinders filled with water. Guides for dropping objects from two different heights above the water (high and low) were attached to the cylinders. To set up a comparison, participants selected two objects they wished to compare and told the experimenter from which height they would like each object to be dropped. To execute a comparison, children observed as the experimenter held the objects at the specified heights above the water and then dropped them simultaneously. Children were asked to determine which object in the comparison sank faster in the water, and the outcome measured was which object hit the bottom of the cylinders first.

Procedure

The procedure was divided into four phases spread over two days: (1) Exploration and training (for the Training−Probe condition only), (2) Assessment, (3) Transfer-1, and (4) Transfer-2. Phases 1 and 2 took place on Day 1 and Phases 3 and 4 on Day 2. Two sessions of about 40 min each were used on two different days. Day 2 was separated from Day 1 by approximately 1 week (exactly 7 days for 87% of the cases; $M = 7.7$, $SD = 2.4$). Participants were interviewed individually in a quiet space in their school, and all children's activities, including their designs and explanations, were videotaped for later coding and analysis. A

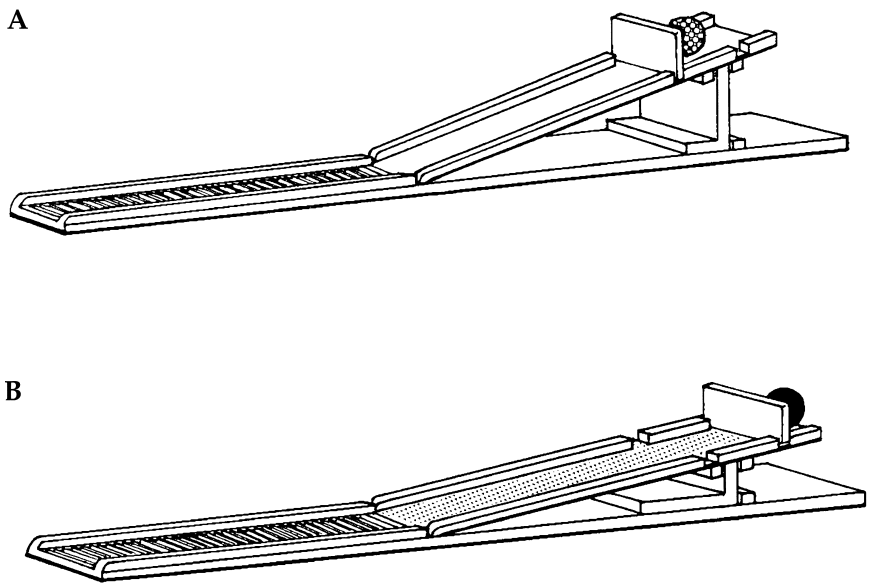

Figure 1 The Slopes Domain. On each of the two slopes, children can vary the angle of the slope, the surface of the ramp, the length of the ramp, and the type of ball. The confounded experiment depicted here contrasts (A) the golf ball on the steep, smooth, short ramp with (B) the rubber ball on a shallow, rough, long ramp. See Table 1 for additional information.

Table 2 Procedure Table

		Condition		
		Training–Probe	No Training–Probe	No Training–No Probe
Day 1	Phase 1a—Exploration			
	Cover story, Task 1	X	X	X
	Identify variables A, B, C, and D[a]	X	X	X
	Initial conceptual understanding	X	X	X
	Produce two comparisons each for A and B	X	X	X
	Explanations (probes)	X	X	—
	Phase 1b—Training			
	Training on variables A and B	X	—	—
	Phase 2—Assessment			
	Produce two comparisons each for C and B	X	X	X
	Explanations (probes)	X	X	—
	Final conceptual understanding	X	X	X
Day 2	Phase 3—Transfer-1			
	Cover story, Task 2	X	X	X
	Identify variables E, F, G, and H	X	X	X
	Initial conceptual understanding	X	X	X
	Produce two comparisons each for E and F	X	X	X
	Explanations (probes)	X	X	—
	Final conceptual understanding	X	X	X
	Phase 4—Transfer-2			
	Cover story, Task 3	X	X	X
	Identify variables I, J, K, and L	X	X	X
	Initial conceptual understanding	X	X	X
	Produce two comparisons each for I and J	X	X	X
	Explanations (probes)	X	X	—
	Final conceptual understanding	X	X	X
	Similarity questions	X	X	X
	Final Training for School A		X	X

[a] Capital letters refer to variables used. All four variables were identified for the children, and their prior beliefs about all four were elicited. Subjects were then asked to make comparisons, typically for only two of the variables in each phase.

short description of activities in each phase of the interview for a child in the Training–Probes condition is given below. Table 2 gives an overview of the activities children completed in each condition.

Phase 1a—Exploration (Day 1). A brief cover story introduced the first task (e.g., springs) and all variables (e.g., length, width, wire size, and weight) that might affect the outcome in that task. Children were first asked to identify the variables to ensure that they could correctly map our verbal descriptions of each variable to the physical dimensions of the materials. Next, their conceptual knowledge was assessed by asking them to indicate their understanding of the causal factors in the domain. More specifically, they were asked which of the two levels of each variable would have a greater effect on the outcome. (For example, in the Sinking domain, they were asked whether they thought that a sphere or a cube would sink faster, and in the Slopes domain, they were asked whether they thought that a ball would roll farther after it left a smooth ramp or a rough ramp.) For each of two target variables identified by the experimenter (A and B, where, for example, A = spring length and B = spring diameter), children (1) produced two comparisons to generate relevant evidence (production task), and (2) answered probes about their choice of comparisons and what they could tell from the outcomes (explanations).

Phase 1b—Training (Day 1). Children in the Training–Probe condition were provided with explicit instruction concerning CVS. During training, the participants were given both negative (confounded) and positive (unconfounded) examples (designed by the experimenter) and were asked to make a judgment of whether each example was a good or bad comparison and to explain why. The experimenter then explained whether and why each example was a good or bad comparison.

Phase 2—Assessment (Day 1). For each of two target variables (C and B, one new and one old variable), children were asked to produce two comparisons and explain the reasoning behind their choice of objects and their conclusions.

Phase 3—Transfer-1 (Day 2). A cover story introduced the second task (e.g., slopes) and all variables that might affect the outcome in that task. Children were asked to identify the variables and say how they thought each variable would affect the outcome (initial domain knowledge). For each of two target variables, children were asked to produce two comparisons and explain the reasoning behind their choices and conclusions as in the earlier evaluative phases.

Phase 4—Transfer-2 (Day 2). This phase was identical to Phase 3, except that the third task (e.g., sinking) was introduced.

The procedure for the No Training–Probe condition was the same as in the Training–Probe condition, except that these subjects did not receive training between Exploration and Assessment. In addition, children in the No Training–Probe condition made three comparisons for each variable identified in the Exploration and Assessment phases, to compensate for the longer time on task that Training required in the Day 1 procedure. In the No Training–No Probe condition, children were only asked to produce comparisons; they did not have to explain the reasoning behind their choices or conclusions. These children produced four comparisons for each variable identified in the Exploration and Assessment phases and three comparisons for each variable identified in Transfer-1 and Transfer-2, to compensate for the extra time on task afforded by both Training and use of probes.

After children solved all the problems, they were asked a series of questions about the similarity between the task from Day 1 and the transfer tasks completed in Day 2. They were asked (1) if anything about the transfer tasks on Day 2 reminded them of the Day 1 task, (2) to explain how the three tasks were alike and/or different, and (3) whether they learned anything on Day 1 that helped them to work on the transfer tasks on Day 2.

The procedures at both schools were essentially the same, except for the following differences: (1) at School A, children in the two No Training conditions did receive training at the end of the hands-on study, after the Transfer-2 phase, and (2) at School B, in addition to being asked about their domain knowledge of each task (how they thought each variable would affect the outcome) before making any comparisons, children also were asked about the effects of each variable on the outcome after they completed each task (i.e., after Assessment, Transfer-1, and Transfer-2). The addition of these questions in School B allowed comparison of children's domain knowledge of the tasks before and after they made comparisons.

Part II: Posttest

The posttest was designed to examine children's ability to transfer the CVS strategy they learned in the hands-on study to relatively remote situations. We consider the application of CVS in the posttest as "Remote Transfer" for several reasons: (1) there was a long-term delay (7-month time interval) between the hands-on experiences and the posttest, (2) there were substantial contextual differences in that "experimenters" and settings in which tests were administered differed between the hands-on phases and posttest, and (3) the tasks also differed both in format (generating tests in the hands-on tasks versus evaluating tests on paper in the posttest) and in content (mechanical versus other types of domains).

Participants and Timing

Approximately 7 months after the completion of Part I, 55 fourth and fifth graders in School A received the posttest. Mean ages were 9,9 for the fourth graders ($n = 28$) and 10,8 for the fifth graders ($n = 27$). Participants in the Experimental group included 24 of the 29 students who had received training earlier (9, 8, and 7 in the Training–Probe, No Training–Probe, and No Training–No Probe conditions, respectively). Recall that all students who participated in the earlier hands-on study at school A were trained in CVS, through examples either early (between Exploration and Assessment for the Training–Probe condition) or later (after Transfer-2 for the other two conditions). The Control group consisted of 31 students who had not participated in Part I of the study.

Design and Materials

The posttest consisted of a 15-page packet containing three problems in each of five domains: plant growth, cookie baking, model airplanes, drink sales, and running speed. The domains were chosen to represent a wide range of contexts in which CVS can be applied, from the natural and social sciences to everyday situations such as cooking. The range of domains also presented varying levels of distance from the domains of the physical sciences used in the original CVS training tasks. Each domain involved three 2-level variables. For example, in the plant growth domain, plants could get a little or a lot of water, a little

or a lot of plant food, and a little or a lot of sunlight. Children were asked to evaluate comparisons that tested the effect of one target variable.

Each domain was introduced on a page of text identifying the variables and outcome for that problem and specifying the target variable. The next three pages depicted comparison pairs, one pair to a page, which students were to rate as good or bad tests of the target variable. The comparisons were of four types: unconfounded comparisons, comparisons with a single confound, comparisons in which all three variables had different values, and noncontrastive comparisons in which the target variable was the same in both items in the pair. Within each domain, one of the three comparisons shown was unconfounded, whereas the other two items were chosen from the three types of poor comparisons.

Comparison pairs were presented both in text and as pictures. At the top of each page was a description of the comparison in words, explaining that the pictures represented a comparison between two situations to determine if the target variable affected the outcome. Then the conditions in the two pictures were described, focusing on each variable in turn (e.g., "they gave plant A lots of water and they gave plant B a little water; they gave plant A lot of plant food and they gave plant B a little plant food; they gave plant A lots of sunlight and plant B no sunlight"). Boxes in the center of the page showed the variables in the comparison pictorially. Written labels were used with the symbols in the picture boxes to ensure that the level of each variable was clear. Figure 2 shows a typical page from the test booklet.

At the bottom of each comparison page, students read instructions reminding them of the target variable and asking them to circle "good test" if they felt the pictures showed a good way to find out about that variable and to circle "bad test" if they felt it was a bad way.

Several of the tasks used in the hands-on study and posttest were loosely adapted from previous studies: the Spring Problem from Linn (1980) and Bullock and Ziegler (1994); the Sinking Problem from Penner and Klahr (1996); the Cookie Baking Problem from Tschirgi (1980); and the Model Airplane Problem from Bullock and Ziegler (1999).

Procedure

The tests were administered during science class by the children's regular science teachers. All fourth graders had one teacher, and all fifth graders had another teacher. All children in each grade were tested on the same day. The teachers did not inform the students of any relationship between the posttest and the earlier hands-on experiment, and the researchers were not present during posttest administration.

The teachers first went over two example pictures with the students so the children learned to identify variables in the pictures. Children were instructed to read the descriptions and look at the pictures carefully. The teachers read aloud the description for only the first problem topic and first comparison. For the remainder of the problems, children worked on their own and at their own pace. For each comparison, students had to read the description at the top of the page, evaluate the comparison, and then circle "good test" or "bad test" at the bottom of the page. Most children took about 20 min, and all were finished within 30 min.

RESULTS

Measures

Four major dependent variables were measured: (1) *CVS score*: a simple performance measure based on children's use of CVS in designing tests, (2) *Robust use of CVS*: a more stringent measure based on both performance and verbal justifications (in response to probes) about why children designed their experiments as they did, (3) *Strategy similarity awareness*: based on children's responses to questions about the similarity across tasks, and (4) *Domain knowledge*: based on children's responses to questions about the effects of different causal variables in the domain.

Use of CVS

Children's use of CVS was indexed by their selection of valid comparisons. An example of a valid design to test the effect of the wire dimension is a pair that differs only in the focal variable (e.g., wire diameter), with all other variables (coil width, length, and weight) kept constant. Invalid designs included: (1) noncontrastive comparisons in which the focal variable was not varied and one or more other variables were varied, and (2) confounded comparisons in which the focal variable as well as one or more other variables were varied. Each valid comparison was given a score of 1. All invalid comparisons were given a score of 0. Because children made four comparisons in each phase, the CVS use scores for each phase could range from 0 to 4.

Robust Use of CVS

Children's responses to the probe questions "Why did you set up the comparison this way?" and "Can you tell for sure from this comparison?" were coded and several types of explanations were identified:

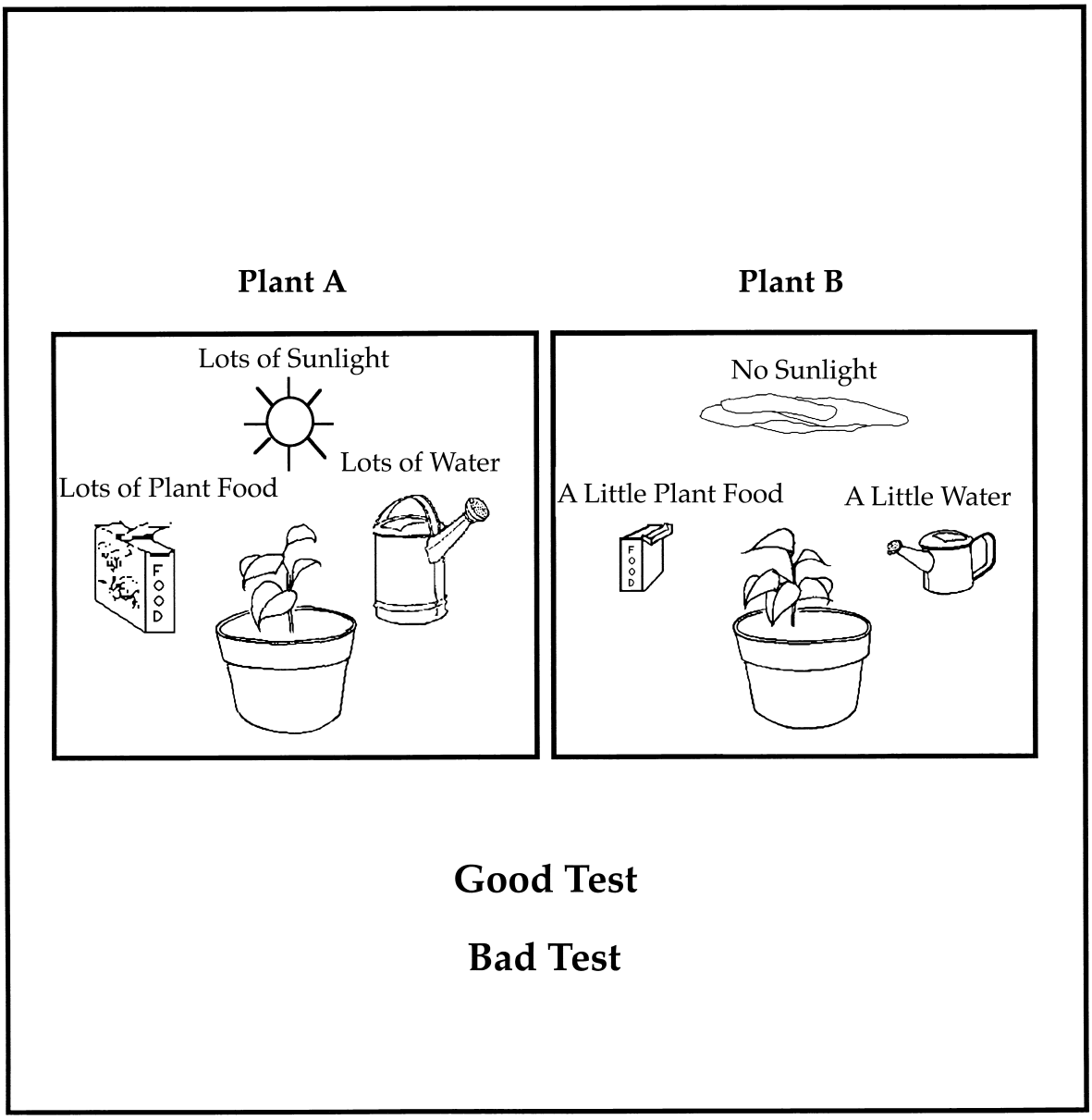

Figure 2 An example of a page from the Posttest Problem Booklet.

(1) Explanations that included mentions of CVS (e.g., "You just need to make the surface different, but put the gates in the same places, set the ramps the same height, and use the same kind of balls"); (2) Explanations that included controlling some but not all of the other relevant variables (e.g., "Cause they're both metal but one was round and one was square"); (3) Explanations that mentioned a comparison within the focal variable (e.g., "Cause I had to make the surfaces different"); and (4) Explanations that were irrelevant to CVS. A second observer independently coded 220 responses randomly sampled from the full set of 960. Interrater reliability was 96%.

Note that when children explained their designs and interpreted their test outcomes they did not simply repeat the terminology learned during training. CVS mention during the Assessment phase (immediately following Training, and in the same task domain) required a different contrastive dimension than was used during training, and correct CVS mention during Transfer 1 and Transfer 2 had to go far beyond simple repetition of terminology, and had to make the correct mapping between the under-

lying logic of CVS and the new variables in the new task domain.

Children received a Robust CVS Use score of 1 only for those trials for which they produced an unconfounded design *and* provided an explanation or interpretation that mentioned the control of all other variables. Other trials received a score of 0. Again, because children made four designs in each phase, the range of Robust Use scores was 0 to 4.

Strategy Similarity Awareness

To determine whether children were aware of the strategy level similarity among the tasks, their responses to the similarity questions also were examined. Children perceived the similarities among the problems at different levels. Some children cited CVS as a similarity among the problems (e.g., "We were trying to find out things, and I made everything the same except for the thing you were trying to find out"). Other children cited the similarity in procedural activities between tasks (e.g., "We made comparisons"). Others stated content similarities (e.g., "There were balls"). Still others had irrelevant or "I don't know" responses. Those who mentioned CVS as a task similarity received a score of 1, and others, 0. Interrater reliability based on a sample of 20 children was 100%.

Domain Knowledge

Domain knowledge of each task was assessed by asking children how they thought each variable would affect the outcome both before and after they designed and implemented their tests. Children's correct prediction/judgment of each variable was given a score of 1, and for an incorrect prediction/judgment, a score of 0 was assigned.

Preliminary analyses revealed no significant differences between schools or sex differences in performance in using CVS, both in initial phase and later phases, and in their domain knowledge of the three tasks. Data were thus combined over schools and genders.

Grade Differences in Initial Performance in Using the Control of Variables Strategy

Children's initial performance was measured by the proportion of unconfounded comparisons (out of 4) they produced during the Exploration phase. The first step in the analysis was to compare children's performance to chance. Rather than calculate the chance level of producing an unconfounded comparison of the focal variable from among *all* possible comparisons, we calculated the more conservative (i.e., higher) probability of randomly producing an unconfounded comparison from just the set of possible *contrastive* comparisons (i.e., those for which the test items differ with regard to the variable of interest). This analysis is based on the assumption that most children understood that comparisons should, at the least, differ along the focal variable. Indeed, the majority of children's designs in the Exploration phase (84%) were contrastive (73%, 87%, and 89% for second, third, and fourth graders, respectively). Moreover, these levels of contrastive designs are themselves all significantly above chance ($p < .05$, .001, and .001 for second, third, and fourth graders, respectively). The chance probability of producing an unconfounded, contrastive, comparison of a given focal variable[1] was determined to be .083. In all three grades, children's proportion of unconfounded comparisons in the Exploration phase (26%, 34%, and 48% in second, third, and fourth grades, respectively) was significantly higher than chance, ps $< .01$. A one-way analysis of variance (ANOVA) revealed significant grade differences in initial performance, $F(2, 84) = 3.53$, $p < .05$. Fisher's post hoc test indicated a significant difference between fourth and second graders, $p < .01$, and a marginally significant difference between the fourth and third graders, $p = .067$. Although, as just noted, children performed above chance, they still often designed confounded experiments in the Exploration phase, and younger children were more likely to do so than older children.

Acquisition and Transfer of CVS

To determine whether and how children in different conditions change their strategies in designing ex-

[1] For example, in the Springs domain, if the focal variable is length, then any spring can be Spring A and once Spring A is chosen, only four of the remaining seven springs will result in a contrastive comparison, because three of them are the same length as Spring A. Thus, regardless of how the weights are chosen, the chance probability of choosing a contrastive comparison is $4/7 = .57$. For computing the probability of an unconfounded comparison, given the assumption that we have a contrastive comparison, the calculation is as follows: From among the four contrastive springs that could be chosen as Spring B, only one is unconfounded with respect to Spring A (i.e., it is the one that is the same as A on all dimensions except length). When weights are chosen, any of the four weights (recall that there are two heavy and two light weights) can be hung on Spring A, but, given that choice, only one of the three remaining weights will result in an unconfounded comparison. So the probability of an unconfounded comparison, conditional on a contrastive comparison, is $(1/4)(1/3) = 1/12 = .083$. Similar calculations can be made in the other two domains, except that in the Slope Problem, the contrastive probability is 4/8, rather than 4/7, because subjects could construct ramp B to be identical to ramp A.

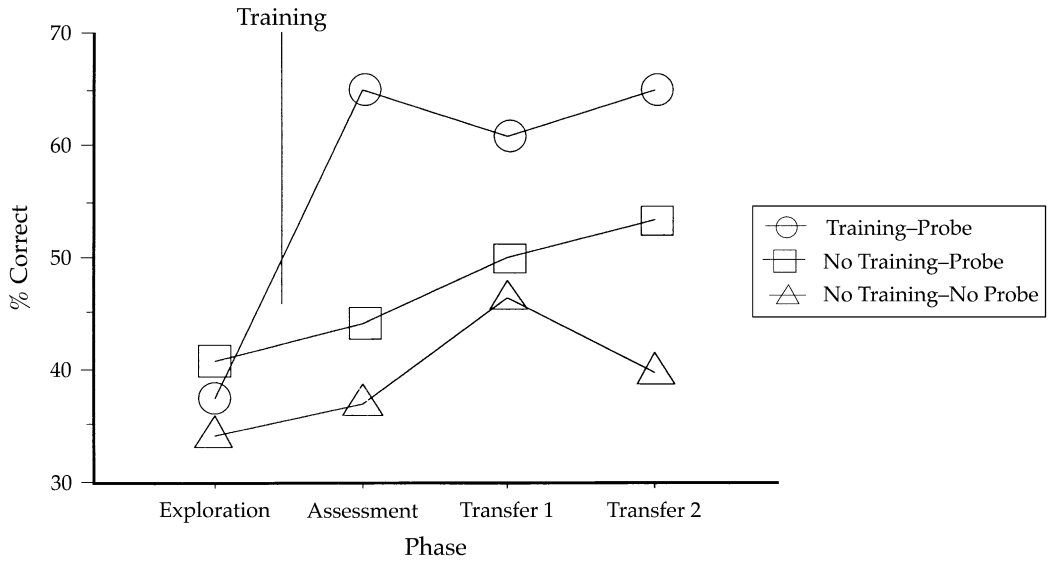

Figure 3 Percentage of trials with correct use of CVS by phase and condition.

periments, the frequency of CVS use in each phase was examined (Figure 3). A 3 (condition) × 3 (grade) × 4 (phase) ANOVA was performed with phase as a within-subjects variable. The analyses revealed a main effect for phase, $F(3, 234) = 6.51$, $p < .001$, indicating that children improved their performance over the course of the four hands-on phases, and an effect for grade, $F(2, 78) = 7.15$, $p < .005$, indicating that older children outperformed younger children. The interaction between condition and phase also was significant, $F(3, 234) = 2.25$, $p < .05$, suggesting that only in some conditions did children improve in using CVS. Separate one-way analyses of variance (ANOVAs) for each condition revealed that only children in the Training–Probe condition increased their performance over phases, $F(3, 87) = 12.8$, $p < .001$. Paired comparisons showed that children did better in the Assessment, Transfer-1, and Transfer-2 phases than in the Exploration phase, but there were no differences in mean CVS scores among the three later phases. In contrast, children's performance in the No Training–Probe and No Training–No Probe conditions did not significantly improve over phase.

Although the interaction between grade, condition, and phase was not significant, we further analyzed grade differences in performance improvement within each condition for three reasons: (1) such grade differences were hypothesized at the outset, so planned contrasts are appropriate; (2) possible grade differences have important implications in educational practice; and (3) second graders seemed to follow patterns different from those of third and fourth graders. For children in the Training–Probe condition (Figure 4A), one-way ANOVAs revealed that only the third and fourth graders improved their performance over phases, ps $< .005$. Paired comparisons indicated that both third and fourth graders performed better in each later phase than in the Exploration phase, ps $< .01$. A one-way ANOVA on the second graders' performance revealed a marginally significant improvement over phases, $p = .10$. Paired comparisons showed that the difference in the performance between the Assessment and Exploration phases was marginally significant, $p = .084$, and that their transfer performance was not significantly higher than the exploration performance.

In contrast, in the No Training–Probe condition (Figure 4B), a one-way ANOVA for each grade level did not reveal a main effect for phase, although paired comparisons showed a marginally significant difference between the Exploration and Transfer-2 phases among fourth graders, $p = .075$. In the No Training–No Probe condition (Figure 4C), neither a one-way ANOVA nor paired comparisons showed performance differences over phase for any grade level.

In order to assess transfer in individual students, we defined a "good experimenter" as a child who produced at least 7 out of 8 unconfounded comparisons during Transfer-1 and Transfer-2, and then we computed the proportion of children who became "good experimenters" between Exploration and Transfer. First, we eliminated children who were already performing at ceiling (4 out of 4 unconfounded experiments) during Exploration. No significant differ-

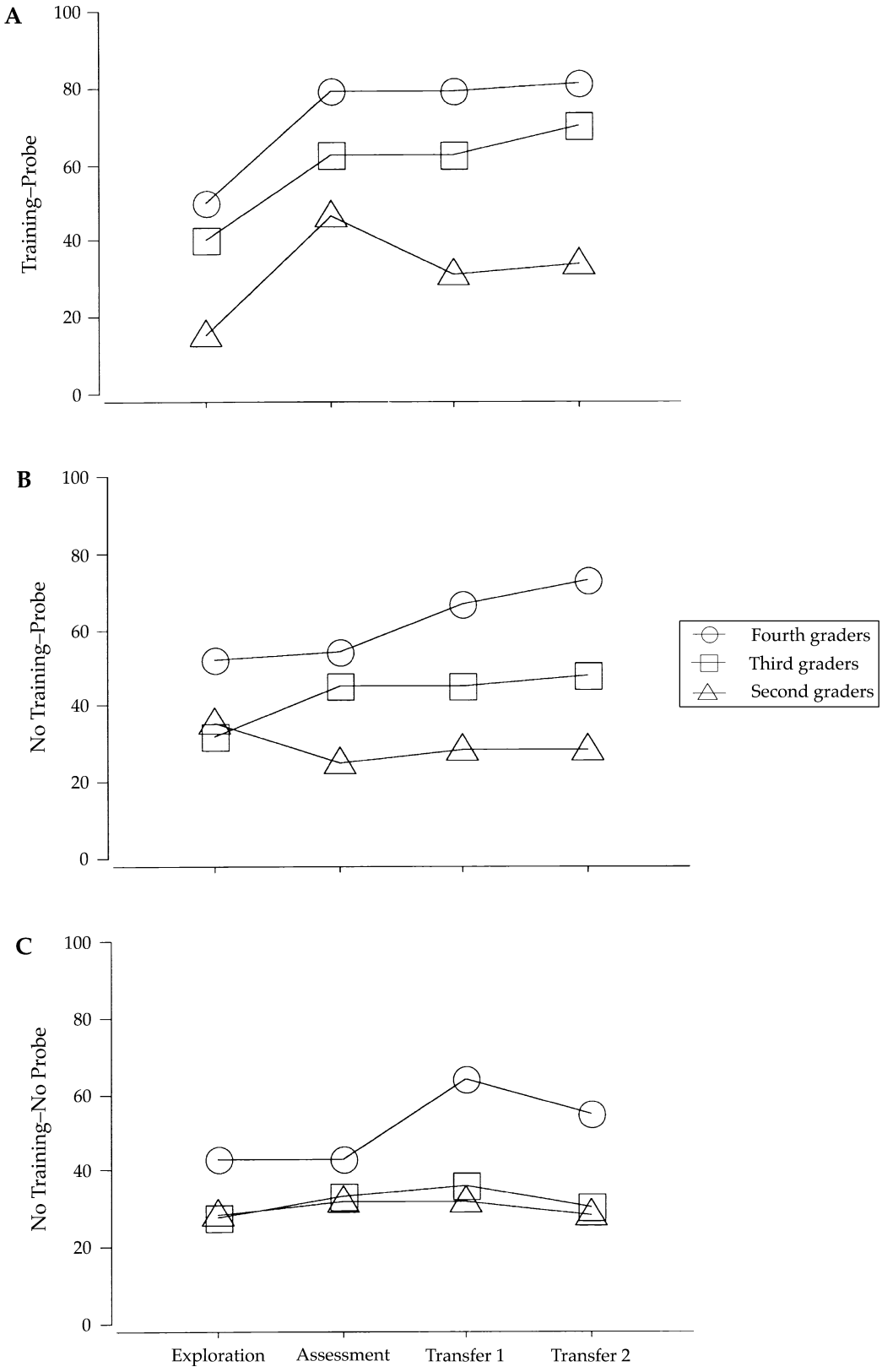

Figure 4 Percentage of correct CVS usage by phase, grade, and condition.

ences were obtained in the proportions of children who were thus eliminated from the Training–Probe, No Training–Probe, and No Training–No Probe conditions (10%, 10%, and 11%, respectively). Among the remaining 78 children, 44% (12/27) of the children in the Training–Probe condition, 22% (6/27) in the No Training–Probe condition, and 13% (3/24) in the No Training–No Probe condition were good experimenters. An overall χ^2 test indicated significant differences among the conditions, $\chi^2(2, N = 78) = 7.05$, $p < .05$. Separate χ^2 tests revealed that more children were good experimenters in the transfer phases in the Training–Probe than in the No Training–Probe condition, $\chi^2(1, N = 53) = 3.00$, $p = .083$ (marginally significant) and the No Training–No Probe condition, $\chi^2(1, N = 52) = 6.25$, $p < .05$.

Interpretation of Similarities across Tasks in Terms of CVS

Children's strategy similarity awareness also was examined. They could mention CVS when they were asked about similarities between the three tasks at the end of the hands-on study (Part I). Only School B data were included in this analysis because all children at School A were trained eventually (children in the two No Training conditions were trained after the Transfer-2 phase) before being asked to compare the problems. More children in the Training–Probe condition (5/20) mentioned CVS as a similarity between problems than did children in the No Training–Probe condition (1/20) and the No Training–No Probe condition (1/18). A Fisher exact test contrasting the Training condition and the two No Training conditions (i.e., the No Training–Probe and No Training–No Probe conditions combined) revealed that children in the Training–Probe condition mentioned CVS more often when comparing the problems than did children in the No Training conditions, $p = .04$.

Relations between the Use of CVS and Domain Knowledge

An important issue concerning the function of training in CVS is whether children's domain-specific knowledge—i.e., their understanding of the effects of the variables associated with springs, ramps, and sinking—improved as a result of training. The present study was designed primarily to examine elementary schoolchildren's ability to learn and transfer CVS, and neither the training nor the probe questions were directed toward, or contingent upon, the children's understanding of the content of the tasks. Any change in children's beliefs about the effects of the variables on the outcomes in all three tasks, however, is of obvious interest, and questions about children's initial and final domain knowledge allowed investigation of this issue. We expected that those children who designed more informative (i.e., unconfounded) comparisons would be more likely to gain accurate information about the causal variables in the domain than those who designed confounded experiments.

All the participants' initial conceptual knowledge was assessed in both Schools A and B, and no significant differences in performance were obtained: children made correct predictions about the effects of variables on 76%, 76%, and 79% in the Training–Probe, No Training–Probe, and No Training–No Probe conditions, respectively. Only data from School B were included in the conceptual change analyses, however, because the final domain knowledge was not assessed in School A. A 2 (time of assessment: initial versus final) × 3 (condition) ANOVA with domain knowledge as within-subject measure revealed a main effect for time of assessment, $F(2, 54) = 6.64$, $p < .05$, indicating that after designing the tests and observing the outcomes of the experiments, children improved their domain knowledge of the tasks. The interaction between condition and phase was also marginally significant, $F(2, 54) = 2.49$, $p = .093$ (See Figure 5).

Further paired comparisons between conditions indicated that only children in the Training–Probe condition significantly improved their domain knowledge, $t(18) = 4.62$, $p < .001$, whereas those in the two No Training conditions did not. Children in the No Training–Probe condition slightly, but not significantly, improved their domain knowledge. Children in the No Training–No Probe condition remained the same in their domain knowledge. These patterns are consistent with their performance in using CVS and indicate that proficiency in designing comparisons may lead to improved domain knowledge of a task.

To determine the relation between children's CVS performance and their *final* domain knowledge, two stepwise regressions were performed, one across conditions and the other within the Training condition. The dependent measure was children's final domain knowledge, and the independent measures were grade and "good experimenters" (who received a score of 1 when they designed at least 12 unconfounded tests out a total of 16 trials; others, 0). Only one variable entered the equation: The good experimenter measure accounted for 9% of the variance in children's final domain knowledge across conditions, and accounted for 21% of the variance within the Training condition. In additional stepwise regressions, none of these variables predicted children's *ini-*

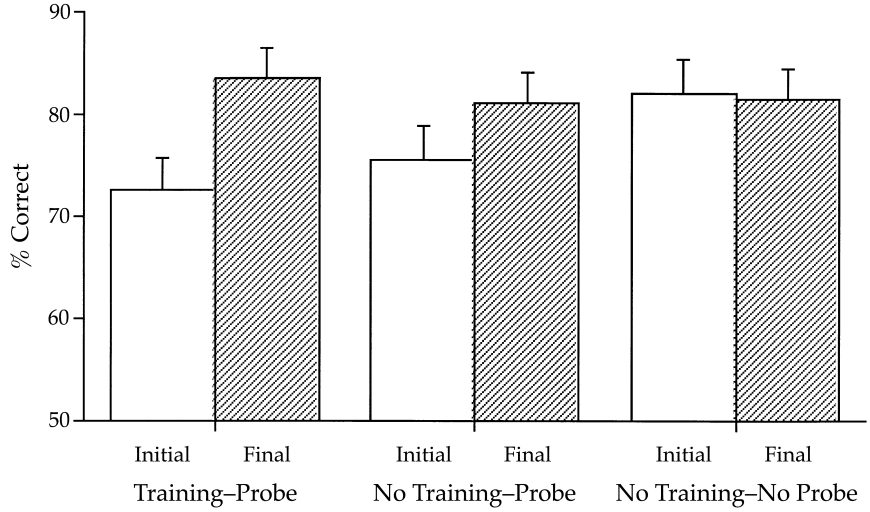

Figure 5 Initial and final conceptual understanding for each instructional group.

tial domain knowledge either across conditions or within the Training condition.

Individual Differences in Strategy Change

In order to examine the detailed time-course of the acquisition and use of CVS, we analyzed, on a trial-by-trial basis, the proportion of participants who generated robust use of CVS on each trial (Figure 6). Children in the No Training–No Probe condition were not included in this analysis because they did not receive probe questions and thus could not be expected to mention CVS. The majority of children started out with poor Robust Use scores. Immediately after training, however, children showed substantial improvement in the robust use of CVS (from about 15% to over 50%), and this level of performance remained throughout the transfer phases. In contrast, children in the No Training–Probe condition continued to perform at their initial low levels.

But even this fine-grained analysis conceals the fact that the pattern of strategy change was highly variable in both conditions. At the bottom of Figure 7, we have depicted specific examples of seven characteristic patterns of Robust CVS Use over the four phases of Part I. The patterns have been classified into two major groups: "Gain-by-end" and "Lose-by-end." Each major group contains distinctive subpatterns. Each of the four patterns in the Gain-by-end group displays improved performance between Exploration

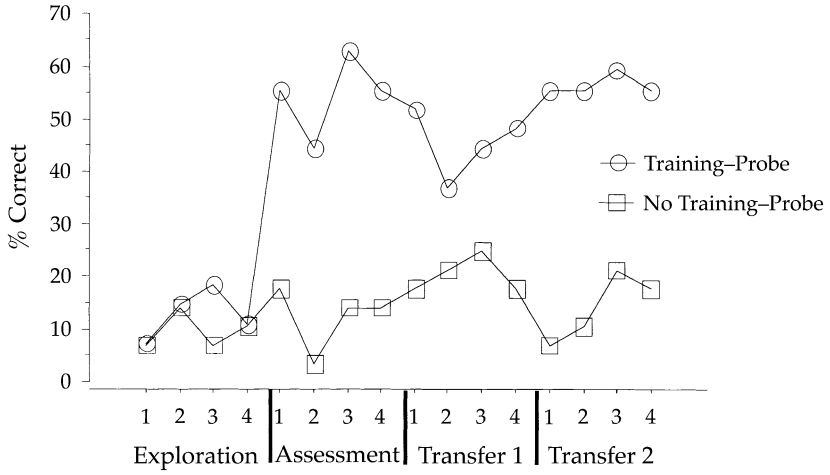

Figure 6 Percentage of children in Training–Probe and No Training–Probe groups who both mentioned and correctly used CVS (Robust Use of CVS) on each trial.

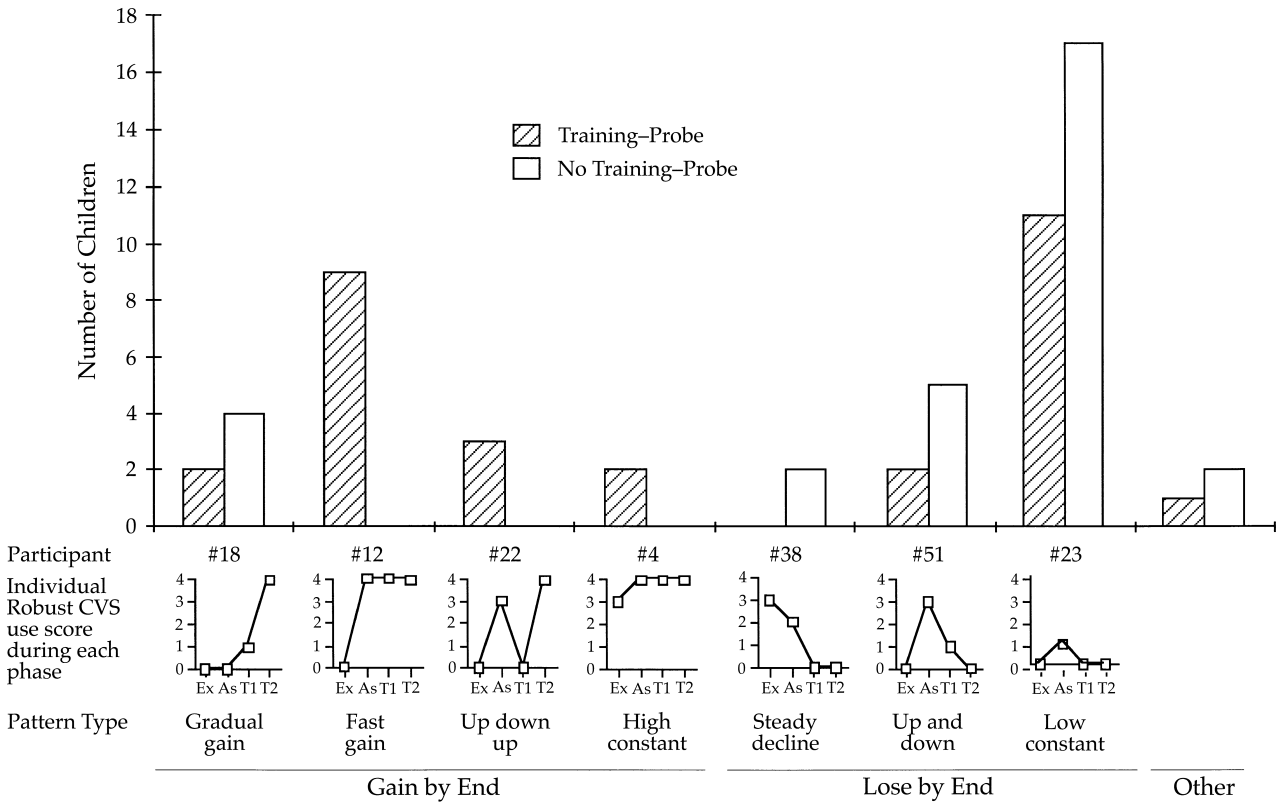

Figure 7 Number of children displaying each type of pattern of Robust Use of CVS across the four phases of Part I. For each pattern type, the results from a specific participant are displayed. These examples depict the number of robust use trials (out of four) during each phase: Ex, Exploration; As, Assessment; T1, Transfer-1; T2, Transfer-2. The pair of bars above each pattern shows the number of children in the Training–Probe and No Training–Probe group whose robust use scores fit that pattern.

and Transfer-2, but along different paths. The first pattern (Gradual gain) does not show improved performance until the transfer phases. The next (Fast gain) shows improved performance right after training, and the high performance continues. The next pattern (Up down up) shows an increase, followed by a decrease, and then a final increase. The final Gain-by-end pattern (High constant) starts with high scores and remains there. There are three Lose-by-end patterns. The first (Steady decline), shows declining performance over phases, following a high start. The second (Up and down) increases and then decreases. The third pattern (Low constant) starts low and remains low.

Each child's pattern of robust use over the four phases was classified according to one of these types of patterns. As shown in Figure 7, even though there is high variability among subjects in which pattern best fit their scores, the distributions differ between the two conditions, $\chi^2(7, N = 60) = 19.57$, $p < .01$. Over half of the children (53%) in the Training–Probe condition and only 13% of the children in the No Training–Probe condition fit one of the "Gain-by-end" patterns. In particular, 30% of the Training–Probe children, but none of the No Training–Probe children, fit the fast gain pattern.

Further evidence against the possibility that the better performance of trained children is simply because the training provided them with the appropriate vocabulary comes from an analysis of the extent to which children in each condition demonstrated robust use (and therefore did explicitly mention CVS) *for at least one trial.* In both conditions, more than half of the children correctly used and mentioned CVS at least once, with 70% of children in the Training–Probe condition and 63% of those in the No Training–Probe condition doing so. Thus, the performance difference cannot be attributed to a lack of access to the appropriate terminology. The analyses also revealed that the conditions differed in the percentage of trials in which CVS was not used after it was used the first time: 68% in the No Training–Probe condition and 24% in the Training–Probe condition. Moreover, during the final three phases, 90% (17/19) of the un-

trained children reverted to less-advanced strategies (defined as at least one trial in which robust CVS was not used after having been used), whereas only 55% (12/21) of the trained children did so. These results suggest that both training and experience consolidated the use of CVS, thereby reducing the frequency with which children returned to less-advanced strategies. Children in the Training–Probe condition increased their consistency in using CVS more readily and to a greater extent than did those in the No Training–Probe condition.

Posttest Performance

The central issue in the posttest was whether children are able to transfer the learned strategy to remote problems with a long (7 month) delay. Posttest data were collected only in School A, and therefore only third and fourth graders were included. Recall that in School A all children who participated in the hands-on interviews were trained in CVS, either early in the procedure or at the end of the hands-on study. All children who participated in the hands-on interview are now considered the Experimental group, whereas their classmates who did not participate make up the Control group.

The main dependent measure was number of correct responses to the 15 posttest problems. A correct response was given a score of 1, and an incorrect one, a score of 0. The mean proportion of correct responses in both conditions is presented in Figure 8. A 2 (group) × 2 (grade) ANOVA yielded a main effect for condition, $F(1, 51) = 6.61$, $p < .05$, and a marginally significant main effect for grade, $F(1, 51) = 2.88$, $p = .096$. The interaction between condition and grade was also marginally significant, $F(1, 51) = 2.77$, $p = .10$. Post hoc tests revealed that fourth graders in the Experimental conditions outperformed those in the Control condition, but no significant differences between conditions were found for third graders.

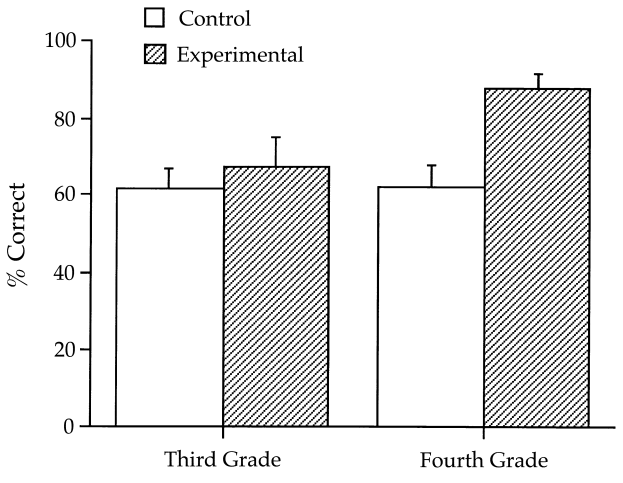

Figure 8 Percentage of correct posttest answers by grade and condition.

Another measure of remote transfer involved the percentage of "good reasoners" in the Experimental and Control conditions. Children who made 13 or more correct judgments out of a total of 15 problems were considered good reasoners. Forty percent of the third and 79% of the fourth graders in the Experimental group were categorized as good reasoners, compared to 22% of the third and 15% of the fourth graders in the Control group. Separate χ^2 tests indicated that the difference between groups in percentage of good reasoners was significant only for the fourth graders, $\chi^2(1, N = 55) = 10.78$, $p < .001$, but not for the third graders.

DISCUSSION

Children's performance in the exploration phase was consistent with previous findings concerning elementary schoolchildren's ability to use CVS. The mean CVS score in the exploration phase was higher than chance, primarily because a small proportion of children (about 15%) already knew the strategy (i.e., received perfect scores in the exploration phase). This initial performance seemed to be somewhat higher than older children's performance in previous studies (e.g., Bullock & Ziegler, 1999; Kuhn et al., 1992; Schauble, 1996), which showed that the majority of fifth and sixth graders produced mainly confounded experimental designs. The relatively high initial performance might have been due to the nature of the present tasks, where the specific variables were clearly identified, and children were directed to design tests of specific variables instead of designing experiments in a self-directed context.

The present results also showed that with appropriate instruction, elementary schoolchildren are capable of understanding, learning, and transferring the basic strategy when designing and evaluating simple tests. Children in the Training–Probe condition increased their use of CVS from 34% of the trials in the Exploration phase (before training) to 65% in the Assessment phase (after training), and to 61% and 64% of the trials in Transfer 1 and 2 phases, respectively. These results also show that explicit training within domains, combined with probe questions, was the most effective way to teach CVS. In contrast, providing probes alone did not significantly increase CVS use. Elementary schoolchildren demonstrated an im-

pressive ability to apply learned strategies across problems. Strategy training also facilitated the acquisition of domain-specific knowledge.

Developmental differences in learning and transfer of scientific reasoning skills were also evident. Second graders, like older children, learned CVS when the transfer tasks were within the same domain as the initial training. Third graders demonstrated the ability to transfer CVS across problems within the domain of mechanics (i.e., when reasoning about springs, slopes, and sinking) and with a short-term delay (1 week). Only fourth graders displayed remote transfer. Finally, although explicit training was an effective source of strategy change, the change was gradual, and even after CVS was used, children continued to use ineffective strategies for some trials. Explicit instruction facilitated consistent use of CVS.

Effects of Explicit Instructions on the Acquisition of CVS

The extensive literature on the relative effects of direct and explicit instruction and guided discovery on children's learning and transfer has yet to produce a consensus on how they exercise their effects on knowledge acquisition. Discovery learning, which encourages learners to figure out concepts or principles on their own, has been found more effective than direct instruction in the acquisition of concepts and rules both in adults and children (e.g., Guthrie, 1952; Katona, 1940). Discovery learning requires learners to engage in deeper cognitive processes such as induction and generalization of rules (Andre, 1986; Jacoby, 1978; McDaniel & Schlager, 1990).

The probe questions that are associated with this type of strategy learning and that were used in this study, however, did not prove helpful in guiding children in discovering CVS. Only a small proportion of children in the No Training–Probe condition improved their performance in designing unconfounded tests or citing CVS. These results are in accord with recent microgenetic studies on scientific reasoning skills (e.g., Kuhn et al., 1995; Schauble, 1996), arithmetic strategies (e.g., Siegler & Jenkins, 1989) and other concepts such as number conservation (Siegler, 1995), which revealed that, absent direct instruction, children typically show only very gradual improvement in the accuracy of their solutions and the quality of their explanations. In contrast, explicit instruction has been viewed as limited and less effective because it tends not to require children's active processing, and, thus, knowledge acquired via explicit instruction is assumed to be of limited generality. Nevertheless, the present results indicate that explicit instruction, combined with probes, can be highly effective and efficient in teaching children a scientific reasoning skill such as CVS. Note that providing explicit instruction is not necessarily associated with passive learning. Children in the present study did not receive instruction passively. The instructions were given only after children had opportunity to explore the task and then in combination with probe questions designed to facilitate children's active processing. Thus, even though explicit instruction tied to specific domains (e.g., springs), children who received such direct training were able to transfer the strategy across problems (e.g., to slopes or sinking).

There is a consensus among researchers that optimal instructional approaches may vary as a function of the types of knowledge and skills being acquired and that the effectiveness of different instructional approaches also depends on learners' abilities (e.g., Dillon, 1986; Linn, 1986). Consistent with these views, the present results, when combined with previous findings, suggest that the power of each approach depends on the learning content (e.g., concepts or subject matter knowledge versus processing skills or strategies) and children's age and knowledge base. This issue is further discussed in the section about Educational Implications, below.

Near and Remote Transfer of Scientific Reasoning Strategies

Where do ideas for designing experiments in novel domains come from? In an investigation of the thinking processes of scientists in world-class molecular biology labs, Dunbar (1997) identified two important processes. First, analogical reasoning was the primary source of new experiments. Second, and more important, such analogies tended to be "local" rather than "global." That is, scientists used variants of known procedures in which the analogical mappings from the known to the new experiment were very simple, rather than more abstract analogies (such as the famous solar system–atomic structure analogy). Thus, it is clear that analogical reasoning plays a central role in the real world of scientific discovery. The extensive literature on analogical transfer and problem solving (Brown, 1989; Gentner & Markman, 1997; Goswami, 1991, 1996; Holyoak & Thagard, 1995) indicates that analogical problem solving involves several major cognitive processes. First, students need to construct a representation of the source problem(s); second, when encountering a similar problem, students need to access the relevant source information and notice the similarity shared by the problems; third, the key components of the problems need to be mapped, so

that the source solutions or strategies can be extended; and fourth, the relevant solution needs to be implemented in the new context or domain.

In the case of CVS, the relevant processes require the acquisition of the strategy in a specific domain and then the accessing, mapping, and implementation of that strategy from one context to another. Children in all grades did well on very near transfer. Here they were taught how to use CVS to test one variable (e.g., spring length) in a domain and then they were asked to design experiments to test another variable in the same domain (e.g., spring width). As transfer distance increased, however, grade differences began to appear. In all but the very near transfer problems, children had to solve problems different from the original learning task, in either the same or different general domains, and with either a short or long delay. In order to succeed, children had to retrieve the strategy, map the problems, and implement the strategy in a new context with new operations. Third and fourth graders proved capable of transferring a newly learned strategy to other tasks in the same general domain. Second graders, however, experienced difficulty in mapping the original task and the newly encountered task, and in implementing the strategy in designing tests. In the Transfer-1 and -2 phases, second graders' performance in designing tests in the Training–Probe condition (35% were unconfounded designs) was not higher than that in the No Training–Probe condition (29%). On the other hand, these children mentioned CVS far more often in the Training–Probe condition (35% of trials) than in the No Training–Probe condition (0%). These data suggest that second graders might not have trouble accessing the learned CVS, presumably because they were tested by the same experimenter in the same context, with only a week between the original and transfer tasks.

In transferring CVS to remote problem situations, third graders experienced difficulty mainly in accessing the strategy learned 7 months before in a very different context. It was unlikely that they had difficulty implementing CVS because the posttest was an evaluation task and did not involve manipulating devices. Thus, once the child had an idea of controlling variables, applying this idea in evaluating the posttest comparisons was relatively straightforward.

Developmental Differences in Strategy Acquisition and Transfer

The present study revealed two main aspects of developmental differences in the acquisition of CVS. The first involved the effects of explicit and implicit training on the use of CVS. Explicit training benefited both younger and older children's use of CVS; they learned from the direct instruction and designed more unconfounded tests after training. Although implicit training via probe questions was not very effective in improving children's understanding of CVS, older children—but not younger children—appeared to benefit from such questions. Presumably this is because older children already possess rudimentary skills related to basic scientific inquiry, and systematic questioning, thus, was somewhat helpful—though not powerful—in eliciting the use of CVS. Factors such as age, experience, and education might all contribute to the differential effects of probes.

The second developmental difference lies in the extent to which a learned strategy can be transferred. The general findings in the literature indicate that students often encounter obstacles in drawing and using analogies to solve problems. In particular, younger children's representations of source problems tend to be tied to the specific, original learning context, and thus they experience difficulty in perceiving the underlying similarity between analogous problems (e.g., Flick, 1991; Gentner & Gentner, 1983). The present results are consistent with previous findings concerning children's ability to solve problems by analogy and indicated that third and fourth graders successfully applied CVS across problems, whereas only fourth graders used the learned CVS in solving problems with different formats and in different domains after a long delay. In contrast, second graders proved able to use CVS only within the original problem. Further investigation is needed to determine precisely what obstacles younger children experienced in transferring the strategy. Older children transferred CVS more effectively than younger children to the target problems even when they performed equally well in the source domain. Although previous research revealed preschoolers' rudimentary ability to avoid and recognize confounds in very simple evidence evaluation tasks (Fay & Klahr, 1996; Sodian et al., 1991), the present results indicate that the development of a full grasp of experimental processing strategies and the ability to learn these strategies improves with age (e.g., Kuhn et al., 1995; Schauble, 1990, 1996; Siegler, 1995; Siegler & Jenkins, 1989).

Process of Change

The present results also shed light on how children change their processing skills in designing experiments. Siegler (1995, 1996) identified several important dimensions of strategy change. The issue regarding the *path of change* concerns whether one form of thinking is followed directly by a qualitatively dis-

tinct level of understanding, or whether there is a transitional period during which multiple strategies coexist. Our analysis indicated that children engaged in a variety of ways of reasoning before and during training, at both the group and individual levels. Even those children who changed their strategies used various types of explanations before consistently using a more-advanced strategy. The patterns of change over the phases also differed among children. Some children changed their strategies from a less sophisticated one to a more advanced one, while others moved in the opposite direction. Among those who improved their designs and explanations, some improved their performance in the Assessment phase right after training, but performed worse when they encountered new problems. Others improved their performance in the Assessment phase, performed poorly on the first transfer problem, and then improved again when they encountered the second transfer problem. Still others more consistently used the CVS strategy once it was learned during training.

With respect to the *rate of change*, these results also indicate that, absent explicit training, there were no sudden shifts from one strategy to another. Gradual change was evident in that less sophisticated strategies were used even after an advanced strategy was used, and over the four phases, children, particularly the fourth graders, in the No Training–Probe condition slightly improved their experimental strategies. None of the No Training–Probe children showed "fast gain"—where they improved rapidly from the Exploration phase to the Assessment phase. Although children in this condition increased their consistency in using CVS, the number of children using robust CVS for at least one trial in each phase did not increase over phases. In contrast, 9 of the 14 children who improved their performance from Exploration to Transfer-2 fit the "fast gain" pattern. These findings suggest that the speed of strategy change depends on whether explicit instruction is provided as well as on children's age.

The present findings also allowed the investigation of issues concerning the *breadth of change.* In the present study, the majority of children who received training used CVS effectively in designing tests in the same problem (though with different variables) in which training was provided. More important, they not only extended the learned strategy to new problems of the same general mechanical domain, but also to other problems with very different contents and under different testing contexts.

Finally, the present study showed that explicit instruction is a major *source of change* in designing experiments. Although previous studies (e.g., Kuhn et al., 1995; Schauble, 1996) revealed that older children and adults learned from self-directed experimentation, the present results indicated that younger children (second to fourth graders), at the group level, did not improve in their use of CVS, at least over a relatively short period, through self-directed experimentation in problems where feedback was not immediately apparent. Explicit instruction combined with probe questions, however, proved to be effective in facilitating the learning of advanced experimental strategies.

Effects of CVS on the Acquisition of Domain-Specific Knowledge

Although the strategies children use in designing tests and in making inferences have been hypothesized to constrain how new knowledge is acquired, little empirical research has been devoted to the relation between the acquisition of domain-general process skills and domain-specific understanding. The construction or modification of mental models, rules, or concepts depends on the informative feedback of experimental outcomes and valid inferences drawn from them. If the tests are confounded, it is impossible to obtain accurate feedback concerning the effects of a particular variable, and thus it is difficult for children to refine their domain knowledge. On the other hand, unconfounded comparisons and valid inferences allow children to detect misconceptions and hence improve their domain-specific knowledge. In this study, children in the Training–Probe condition improved their domain-specific knowledge, whereas children in the other conditions did not. These results indicate that acquisition of a domain-general skill such as CVS can, in turn, facilitate the acquisition of domain-specific knowledge such as the role of causal variables in a variety of physical domains.

Educational Implications

The present findings have implications for important and recurring topics in science education. First, even early elementary schoolchildren can be trained to understand and use CVS. Young children not only are capable of transferring specific solutions across problems, they also can master scientific process strategies by transferring them to tasks with domain contents very different from the original learning task. Compared to older children, second graders did not apply CVS effectively to solve other problems. Yet, it is important to note that the training was quite brief and was provided only on one occasion. One approach to increase the likelihood of successful trans-

fer is to facilitate the construction of an abstract, generalized schema. Problem schemas often are formed through induction as a result of experiencing various instances of the general solution principle or rule. One critical factor facilitating schema construction is the opportunity to process diverse instances that share a similar goal structure or solution principle (e.g., Brown, Kane, & Echols, 1986; Chen & Daehler, 1989; Gentner, 1983, 1989; Gick & Holyoak, 1983). Providing such opportunity to second graders may prove effective in promoting more remote transfer of CVS.

Second, the power of each type of instruction depends on the learning content. Extensive work has been done on the power of discovery learning in the acquisition of mental models, rules, and concepts (e.g., Gentner & Gentner, 1983). When the tasks or problems generate outcomes that provide clear feedback, children are capable of modifying their initial mental model and discovering a rule or principle (Siegler, 1976). For most problems such as balance scale tasks, feedback returning from children's own performance and the observation of problem outcome (whether the outcome is consistent with their own prediction) will indicate how successful a strategy turns out to be. In contrast, the nature of CVS makes it difficult for self-correction to take place in a self-directed context for early elementary schoolchildren. The present results indicate that even systematic, guided probing did not facilitate children's understanding and learning of CVS, presumably because the outcome of each children's experiment did not provide informative feedback to indicate a confounded design or an invalid inference. Only explicit training pinpointing the rationale was effective in the acquisition of CVS.

There was a type of discovery learning that occurred in this study, however, but it was with respect to domain knowledge, rather than CVS itself. Our results demonstrate that, when children designed unconfounded experiments, they were able to correctly interpret the outcomes of those experiments, and thereby revise their initial misconceptions about the effects of each variable. This was particularly evident in the Training–Probe condition where children designed fewer confounded tests. Thus, direct instruction about a process skill facilitated discovery learning about domain knowledge.

Finally, the power of each type of instruction depends on children's age and initial knowledge. Systematic probe questions did not elicit children's use of CVS. Although younger children, particularly second graders, did not benefit from the probe questions at all, older children, especially fourth graders, used CVS somewhat more often in the No Training–Probe condition than those in the No Training–No Probe condition. Older children benefit from probes presumably because they have at least a rudimentary understanding of CVS. When children possessed only knowledge about CVS, they did not spontaneously use it in designing tests and making inferences in a new domain. When probe questions were provided, they guided the children in designing strategies and in considering their own comparisons in terms of CVS.

Conclusions

Elementary schoolchildren's relatively poor performance using CVS is not due to their inability to understand the rationale; they are capable of gaining a genuine understanding of CVS and transferring the strategy when designing and evaluating simple tests. Explicit training within domains, combined with probes, proved to be effective in facilitating the acquisition of CVS. Providing probes (without training), however, only slightly improved children's ability to design unconfounded experiments and make valid inferences. Receiving direct instruction concerning CVS not only improved the use of CVS but also facilitated conceptual change in the domain because the application of CVS led to unconfounded, informative tests of domain-specific concepts. Important developmental differences in the transfer of CVS were evident. With age, children demonstrated increased ability to transfer learned strategies to remote situations. Second graders transferred CVS only to very near situations; third graders were able to transfer CVS to both very near and near situations; and fourth graders were successful in transferring the strategy to remote situations. The present study also started to identify the difficulties that younger children have when transferring strategies. Younger children experienced difficulties in accessing the learned strategy and implementing it in various transfer situations.

ACKNOWLEDGMENTS

Different phases of this work were supported in part by an NIH Post-Doctoral Fellowship (MH19102) to the first author, by a grant from NICHD (HD 25211) to the second author, and by a grant from the James S. McDonnell Foundation (96-37). We thank Jennifer Schnakenberg for her participation in the preparation of the training scripts, as well as her careful and skillful data collection and transcription. Thanks also to the children, parents, principals, and teachers at the Ellis School and the Winchester-Thurston School for

their participation and cooperation. We acknowledge the constructive comments of David Bjorklund, Robbie Case, Leona Schauble, and two anonymous reviewers. Portions of these data were presented at the meeting of the Society for Research in Child Development, April 1997, Washington, DC, the meeting of the Cognitive Science Society, August 1997, San Diego, CA, and the meeting of the American Association of the Advancement of Science, February 1998, Philadelphia, PA.

ADDRESSES AND AFFILIATIONS

Corresponding author: Zhe Chen, Department of Human and Community Development, University of California, One Shields Ave., Davis, CA 95616; e-mail: zhechen@ucdavis.edu. Chen was at Carnegie Mellon University at the time of this study, with David Klahr.

REFERENCES

Andre, T. (1986). Problem solving and education (1986). In G. D. Phye & T. Andre (Eds.), *Cognitive classroom learning: Understanding, thinking, and problem solving* (pp. 169–204). New York: Academic Press.

Bjoklund, D. F., Coyle, T. R., & Gaultney, J. F. (1992). Developmental differences in the acquisition and maintenance of an organizational strategy: Evidence for utilization deficiency hypothesis. *Journal of Experimental Child Psychology, 54*, 434–438.

Brown, A. L. (1989). Analogical learning and transfer: What develops? In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 369–412). Cambridge, UK: Cambridge University Press.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology, 20*, 493–523.

Brown, A. L., Kane, M. J., & Echols, C. H. (1986). Young children's mental models determine analogical problems with a common goal structure. *Cognitive Development, 1*, 103–121.

Bullock, M., & Ziegler, A. (1994). Scientific Thinking. In F. E. Weinert & W. Schneider (Eds.), *The Munich Longitudinal Study on the Genesis of Individual Competencies (LOGIC): Report no. 11* (pp. 56–76). Munich: Max Plank Institute for Psychological Research.

Bullock, M., & Ziegler, A. (1999). 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. 38–54). Munich: Max Plank Institute for Psychological Research.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

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., & Daehler, M. W. (1989). Positive and negative transfer in analogical problem solving. *Cognitive Development, 4*, 327–344.

Chen, Z., & Daehler, M. W. (1992). Intention and outcome: Key components of causal structure facilitating mapping in children's analogical transfer. *Journal of Experimental Child Psychology, 53*, 237–257.

Chi, M. T. H., & Ceci, S. (1987). Content knowledge: Its role, representation, and restructuring in memory development. In L. Lipsitt (Ed.), *Advances in Child Development and Behavior* (Vol. 20, pp. 91–143). New York: Academic Press.

Dillon, R. F. (1986). Issues in cognitive psychology and instruction. In R. F. Dillon & R. L. Sternberg (Eds.), *Cognition and instruction* (pp. 1–11). Orlando, FL: Academic Press.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. Ward, S. Smith, & S. Vaid (Eds.), *Conceptual structures and processes: Emergence, discovery and change* (pp. 461–492). Washington, DC: APA Press.

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

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

Flick, L. (1991). Where concepts meet percepts: stimulating analogical thought in Children. *Science Education, 75*, 215–230.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155–170.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199–241). Cambridge, UK: Cambridge University Press.

Gentner, D., & Gentner, D. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 199–241). Hillsdale, NJ: Erlbaum.

Gentner, D., & Markman, A. B. (1997). Structure-mapping in analogy and similarity. *American Psychologist, 52*, 45–56.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10*, 277–300.

Gick M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development, 62*, 1–22.

Goswami, U. (1996). Analogical reasoning and cognitive development. In H. W. Reese (Ed.), *Advances in child development and behavior* (pp. 92–135). New York: Academic Press.

Guthrie, E. R. (1952). *The psychology of learning.* New York: Harper.

Holyoak, K. J., Junn, E. N., & Billman, D. O. (1984). Developmental analogical problem solving skills. *Child Development, 55*, 2042–2055.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought.* Cambridge, MA: MIT Press.

Jacoby, J. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of verbal learning and verbal behavior, 17*, 649–667.

Katona, G. (1940). *Organizing and memorizing.* New York: Columbia University Press.

Klahr, D. (1999). *Exploring science: The cognition and development of discovery processes.* Cambridge, MA: MIT Press.

Klahr, D., & Carver, S. M. (1988). Cognitive objectives in a LOGO debugging curriculum: Instruction, learning, and transfer. *Cognitive Psychology, 20*, 362–404.

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

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology, 25*, 111–146.

Kuhn, D., Amsel, E. D., & O'Loughlin, M. (1988). *The development of scientific thinking skills.* San Diego, CA: Academic Press.

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*(4, Serial No. 245), 1–128.

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

Linn, M. C. (1980). Free-choice experiences: How do they help children learn? *Science Education, 64*, 237–248.

Linn, M. C. (1986). Science. In R. F. Dillon & R. L. Sternberg (Eds.), *Cognition and instruction* (pp. 155–197). Orlando, FL: Academic Press.

McDaniel, M. A., & Schlager, M. S. (1990). Discovery learning and transfer of problem-solving skills. *Cognition and Instruction, 7*, 129–159.

Penner, D., & Klahr, D. (1996). The interaction of domain-specific knowledge and domain-general discovery strategies: A study with sinking objects. *Child Development, 67*, 2709–2727.

Reed, S. K., & Bolstad, C. A. (1991). Use of examples and procedures in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 753–766.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 456–468.

Ross, B. H., & Kilbane, M. C. (1997). Effects of principle explanation and superficial similarity on analogical mapping in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 427–440.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*, 31–57.

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

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

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*, 481–520.

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cognitive Psychology, 25*, 225–273.

Siegler, R. S. (1996). *Emerging mind: The process of change in children's thinking.* New York: Oxford University Press.

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct mean for studying cognitive development. *American Psychologist, 46*, 606–620.

Siegler, R. S., & Jenkins, E. (1989). *How children discover new strategies.* Hillsdale, NJ: Erlbaum.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62*, 753–766.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development, 51*, 1–10.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*, 535–585.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*, 337–375.

Wellman, H. M., & Gelman, S. A. (1998). Knowledge acquisition in foundational domains. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception & language* (pp. 523–574). New York: Wiley.