Cognitive Development 23 (2008) 488–511

Contents lists available at ScienceDirect

Cognitive Development

Developing elementary science skills: Instructional effectiveness and path independence☆

Mari Strand-Cary, David Klahr*

Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States

ARTICLE INFO

Keywords:
Science education
Direct instruction
Discovery learning
Control of Variables
Experimental design
Far transfer

ABSTRACT

We explore the immediate and longer term consequences of different types of instruction about a central topic in middle school science: the "Control of Variables Strategy" (CVS). CVS represents the procedural and conceptual basis for designing simple, unconfounded experiments, such that unambiguous causal inferences can be made. CVS appears to be what has been called a "developmentally-secondary" process, because even though infants and pre-schoolers can make simple causal inferences from data, middle school children do poorly at CVS unless they receive instruction on this important topic in the science curriculum. In this study, 72 third, fourth, and fifth-grade students were taught CVS via two instructional methods located at extreme points on the direct-to-discovery spectrum with respect to the amount of guidance, information, support, teacher control, and feedback provided during training. Our design included near- and far-transfer measures (at 1 week, 3 months and 3 years). There were two primary outcomes, both of which replicated and partially extended earlier work by Klahr and Nigam (2004) [Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*, 661–667] First, at each of the three grade levels, many more children learned CVS in the explicit condition than in the exploration condition. Second, but equally important, *what* students learned was a better predictor of far transfer than the *way* that they learned.

☆ Written tests, posters, and scripts are available directly from the authors and are posted online at http://www.psy.cmu.edu/~tedtutor/strandcary_klahr_08_p1.html.

* Corresponding author.
E-mail address: klahr@cmu.edu (D. Klahr).

0885-2014/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogdev.2008.09.005

1. Short-term and long-term effectiveness and path independence in elementary school science instruction

Scientific reasoning "includes the skills involved in inquiry, experimentation, evidence evaluation, and inference that are done in the service of conceptual change or scientific understanding" (Zimmerman, 2007, p. 172). The topic is of interest to developmentalists because "it is a fruitful area for studying conceptual formation and change, the development of reasoning and problem solving, and the trajectory of the skills required to coordinate a complex set of cognitive and metacognitive abilities" (Zimmerman, 2007, p. 172). Moreover, it presents a puzzle: why do the impressive causal reasoning processes evidenced by the "scientist in the crib" (Gopnik, Meltzoff, & Kuhl, 1999), seem to have atrophied by the time children reach their middle school science classes? On the one hand, even preschoolers appear to be able to identify causal structures and make appropriate inferences from them (Gopnik et al., 2004; Schulz & Sommerville, 2006), while on the other hand, without explicit instruction, a high proportion of third and fourth grade students are unable to create simple experimental contrasts that will enable them to unambiguously identify a causal variable (Chen & Klahr, 1999). Given that instruction plays an important role in children's acquisition of a variety of core scientific procedures and concepts, the science education community is particularly interested in "determining the best methods for improving learning and instruction in science education" (Zimmerman, 2007, p. 173).

In this article, we address these issues in the context of a key aspect of scientific reasoning: understanding the logic of experimental design. In its simplest form, this requires varying one thing at a time by using the Control of Variables Strategy (CVS). CVS is a domain-general

> 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. (Chen & Klahr, 1999, p. 1098)

Because CVS provides a strong constraint on search in the space of experiments, its acquisition contributes to the development of scientific reasoning skills (Klahr & Simon, 1999; Klahr & Dunbar; 1988).

Although some precursor skills important to CVS are evident by first and second grade (cf., Sodian, Zaitchik, & Carey, 1991), late elementary schoolchildren have only a fragile grasp of the concepts and skills underlying the logic of CVS (Bullock & Ziegler, 1999; Klahr & Nigam, 2004; Kuhn & Phelps, 1982; Kuhn, Garcia-Mila, Zohar, & Anderson, 1995; Schauble, 1996), and "there are also conditions under which adults do not show full proficiency" (Zimmerman, 2007, p. 173).

Kuhn and Franklin (2006, p. 974) note that CVS and scientific reasoning, more broadly, "cannot be counted on to routinely develop." Geary, drawing on Zimmerman, makes a similar claim:

> Without solid instruction, children do not: (a) learn many basic scientific concepts, . . . (b) effectively separate and integrate the hypothesis and experiment spaces; (c) effectively generate experiments that include all manipulations needed to fully test and especially to disconfirm hypotheses; and (d) learn all of the rules of evidence for evaluating experimental results as these relate to hypothesis testing. (Geary, 2007, p. 68)

Geary sees this distinction as a specific instance of the broader distinction between what he calls "primary" domains and those that are "secondary": "Biologically primary domains encompass evolutionary-significant content areas. . .and are composed of folk knowledge (e.g., inferential biases) and primary abilities (e.g., language, spatial)" (Geary, 2007, p. 43). The early-appearing causal reasoning processes documented by Gopnik and colleagues are clear examples of such primary processes. However, as Geary notes: "Biologically secondary domains such as mathematics, and biologically secondary abilities and knowledge, such as the ability to phonetically decode written symbols or to understand the base-10 structure of the formal mathematical number system" arise from the modification of primary abilities and folk-knowledge-based attributional biases. They are culture-specific (Geary, 2007, p. 5). He argues that "the attentional and cognitive biases that facilitate the fleshing out of primary

abilities during children's natural activities do not have evolved counterparts to facilitate the learning of secondary abilities" (Geary, 2007, p. 33).

Thus, although early cognitive development yields an array of primary scientific thinking processes, explicit instruction may be necessary for those processes that are secondary. Chen and Klahr (1999) found that CVS is one of those skills that "that are difficult for children to discover by themselves" (p. 1099). Given that CVS requires instruction, it remains to be seen what constitutes 'good' instruction. Indeed, that question is highly controversial—especially in the area of science instruction (Hake, 2005; Kirschner, Sweller, & Clark, 2006). Thus it is important to examine the effects of different types of CVS instruction on different measures of learning (number of learners and robustness of their learning), and transfer (across tasks and time).

A ubiquitous finding from the literature on instructional effectiveness is that even when one type of instruction consistently produces greater *average* gains in learning and transfer than another, there are always some children who do not learn from the more effective method and others who do learn from the less effective method. For example, Chen and Klahr (1999) compared three different types of instruction for teaching CVS. They found that while the children receiving the most explicit and teacher-controlled CVS instruction showed the largest gains in performance, a small, but non-trivial, proportion of children in their two less explicit instructional conditions also showed gains. In a similar study with the same types of materials and instruction, Klahr and Nigam (2004) found that 77% of the children receiving what they called "direct instruction" "mastered" CVS, as did 23% of the children in their "discovery" condition.

These, and similar, findings raise the question of what is learned when children do manage to master CVS in the less directive condition. Many science educators argue that discovery learning leads to richer and more nuanced understanding than does direct instruction (Hmelo-Silver, Holton, & Kolodner, 2000; Hmelo-Silver, Duncan, & Chinn, 2007; Schmidt, Loyens, van Gog, & Paas, 2007; Steffe & Gale, 1995). They predict that children who have learned through discovery will be better able to transfer their knowledge beyond the immediate learning context. However, Klahr and Nigam (2004) found no such result. Instead, children who learned under discovery conditions failed to show any far-transfer advantages over those who learned under direct instruction. Klahr and Nigam called this "path independence." An everyday example of path independence would be a situation in which one could not tell, by any reasonable assessment of driving skill, whether someone had learned to drive by being self taught or by having taken a driving course. In the case that Klahr and Nigam reported, children who had achieved a performance criterion on an immediate post-test via two different types of training conditions (discovery and direct instruction) performed equally well on a far-transfer test, regardless of which training condition they had been in. (Although many more reached the criterion in the direct condition than in the discovery condition.) Conversely, children in the two training groups who had *not* reached the criterion performed equally poorly on the far-transfer test. That is, the far-transfer performance of children who had reached criterion was independent of the "path" (training condition) leading to that performance. This finding of *path-independent transfer* challenges the common criticism that direct instruction somehow shortchanges children by teaching them only fragile and short-lived knowledge of limited generality.

However, Klahr and Nigam's (2004) evidence for path independence was based on an assessment given after a relatively short delay (1 week), used only one measure (evaluation of science fair posters), and was defined by a performance criterion of 75% correct responses (what they termed "mastery"). We wanted to determine just how robust the path-independent phenomenon is because (a) the Klahr and Nigam (2004) finding provides only modest support for a claim that could have substantial instructional implications; (b) there exists an extensive controversy about the benefits and costs of instruction located at different points along the "discovery to direct spectrum" (Adelson, 2004; Begley, 2004; Cavanagh, 2004; Kirschner et al., 2006; Tweed, 2004); and (c) results might change drastically with different operationalizations of transfer and learning. Thus, the goal of the present study is to explore the robustness of Klahr and Nigam's path-independence results by replicating their study and extending it to include a greater variety of assessments, over longer delays (3 months and 3 years), and with a more stringent criterion for learning (perfect performance on the CVS training task). For purposes of comparison and replication, the study includes the same experimental apparatus used in previous

studies, the same contrasting training conditions (here called "Explicit Instruction" and "Exploration"), and a highly similar poster evaluation assessment.

Data collection is extended by including (a) a written pretest of children's experimental design skills using a few items adapted from "high-stakes" standardized tests; (b) a greater number of opportunities for children to detect and correct flaws in the science fair posters; (c) several assessments conducted after a 3-month delay including assessment of CVS skills using the same materials as used in training, an additional poster evaluation, and a written posttest; and (d) a written posttest administered 3 years after the initial data collection. Analysis is extended by examining performance on these measures and, where appropriate, by focusing on the performance of individual children, rather than on group means. In addition, we use a very stringent performance criterion ("Expertise": perfect performance on CVS measures), rather than any of the arbitrary and less-than-flawless performance standards used in earlier work.[1] We also look at the extent to which knowledge acquired at one stage transfers to other contexts and time frames that comprise different "transfer distances." Each of these extensions we describe in detail in the Method section.

2. Method

2.1. Participants

Seventy-two third, fourth, and fifth-grade students (39 girls, 33 boys) from a Pittsburgh Catholic school participated. We recruited students by mailing parents a packet of materials that included an explanation of the study, a recruitment letter, a letter of support from the school principal, and parental consent forms. In addition to parental consent we obtained written consent from each participant. The sample included 20 third graders (13 girls and 7 boys; mean age = 8.8 years; range 8.2–9.6 years), 23 fourth graders (10 girls and 13 boys; mean age = 9.7 years; range 9.1–10.7 years), and 29 fifth graders (16 girls and 13 boys; mean age = 11.0 years; range 10.2–12.3 years), and was 68.1% Caucasian ($n = 49$), 16.7% African American ($n = 12$), 8.3% Asian ($n = 6$), and 6.9% Hispanic ($n = 5$). Within grade and gender, children were randomly assigned to the two training conditions.

2.2. Design

The overall design is depicted in Fig. 1. The study began (Phase 0) with an in-class written pretest assessing participants' knowledge about unconfounded experiments. This was followed by Session A (Phases 1 and 2), Session B (Phase 3) 1 week after Session A, and Session C (Phases 4, 5 and 6) 3 months after Session B. Sessions A, B and C each correspond to a single block of time during which children interacted one-on-one with an experimenter. Three years later (Phase 7), available children took another in-class written post-test.

Session A included the following: Phase 1a – initial exposure to the physical materials (ramps) to be used for designing experiments, followed by an assessment of initial domain knowledge and a ramps pretest of ability to design unconfounded experiments; Phase 1b – participation in either the Explicit Instruction or Exploration condition; and Phase 2 – an immediate ramps posttest and an assessment of domain knowledge.

Session B, conducted about 1 week later, consisted of a single phase (Phase 3) in which all participants were asked – via a structured one-on-one interview – to evaluate two science fair posters (ostensibly created by children in another school).

[1] Chen and Klahr (1999) defined a "good experimenter" as a child who correctly designed 7 of 8 experiments during the near transfer phases and a "good reasoner" as one who correctly answered at least 13 of 15 of the two-choice items on a remote transfer test. Klahr et al. (2001) used slightly different criteria for very high performance. They defined a "CVS expert" as a child who correctly set up at least 8 of 9 experiments, and an "evaluation expert" as a child having at least 9 correct responses on a 10-item two-choice test packet. Klahr and Nigam (2004) defined a "CVS master" as a child who correctly designed at least 3 of 4 experiments. While these different criteria were appropriate in the context of any given study, they make cross-study comparisons difficult. Indeed, this is a common problem in the expertise literature (Ericsson & Charness, 1997) because there are no universally accepted operational definitions of an expert level of performance except in a few well quantified areas, such as chess proficiency. This makes it difficult to compare different studies of, for example, expert-novice differences in physics.

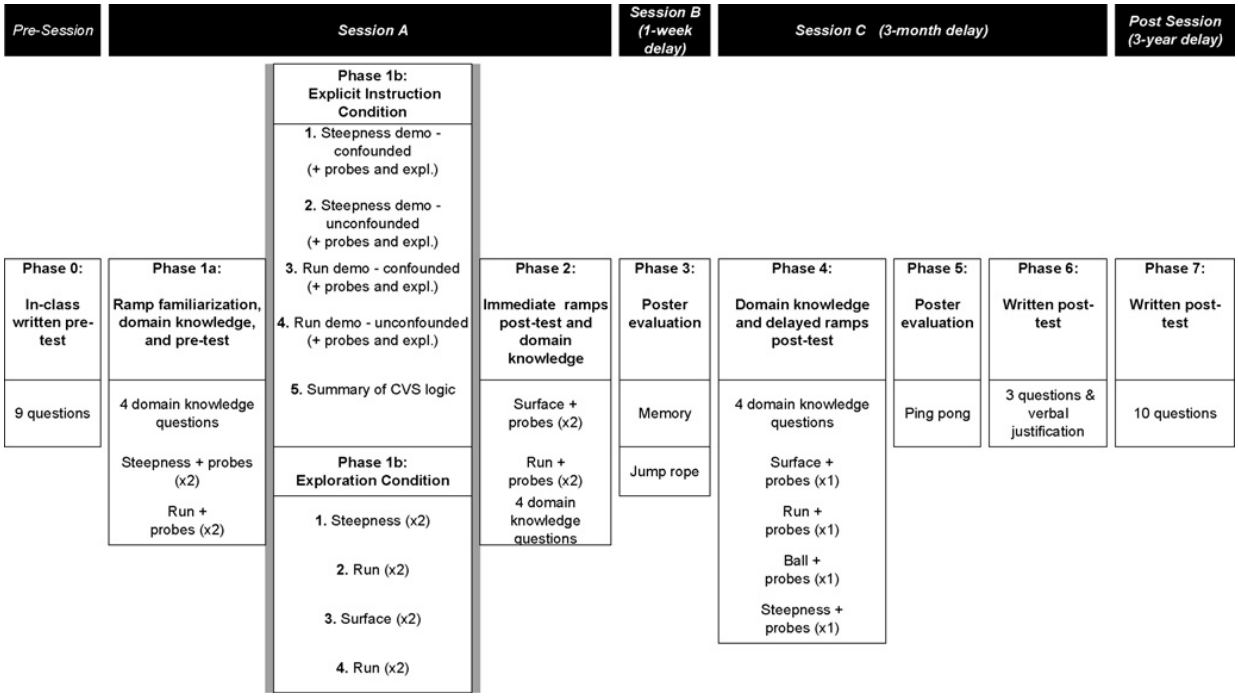

Fig. 1. *Design*. Main temporal flow is from left to right. Within a phase, temporal flow is from top to bottom. The experimental manipulation takes place in Phase 1b, during which children in the Explicit Instruction condition experienced the sequence enumerated in the upper panel, and children in the Exploration condition experienced the sequence enumerated in the lower panel.

Session C (Phases 4, 5, and 6), conducted approximately 3 months later, included each of the types of assessments used previously: domain knowledge, ramps posttest, poster evaluations, and a written posttest about experimental design.

Phases 1 through 6 were videotaped. At the end of each of the three individual sessions, participants were provided with a small gift as thanks for their participation.

2.3. Procedure and materials

2.3.1. Pre-session (Phase 0)

2.3.1.1. Materials for Phase 0. In Phase 0, children completed a written pretest in their regular classrooms. (The detailed procedure is described below. Throughout this report, for each phase, we first describe the materials and then the procedure.) We used a test booklet consisting of nine multiple-choice questions. The first five items – taken from material used by Toth, Klahr, and Chen (2000) – depicted pairs of airplanes that could vary in type of body, wings or tail (see Fig. 2). A focal variable was identified in the text (e.g., "body type") and children were asked whether the pair of airplane designs represented a "good test" or a "bad test" for the effect of the focal variable. Five different types of comparisons were presented: (a) unconfounded; (b) singly confounded; (c) multiply confounded; (d) non-contrastive; and (e) unconfounded comparisons of a non-focal variable. Students were asked to judge whether each picture pair showed a valid experiment to test the focal variable by circling "bad test" or "good test." (Only unconfounded comparisons are "good tests.") Children were also asked to correct "bad tests", but since the youngest children, in particular, misunderstood this task, the corrections were not scored.

The final four items (three multiple-choice questions and one short-answer item) were based on items taken from widely used standardized assessments, slightly reformatted so as to be more easily understood by the younger participants. These items asked children to identify unconfounded experiments in domains involving weighted carts rolling down hills, the effects of sunlight and other variables on plant growth, allergy medicines and placebos, and beetles' preferences for light.

2.3.1.2. Procedure for Phase 0. Approximately 2 weeks prior to Session A, the written pretest was administered by Experimenter 1 in the six homeroom classrooms of participants. In order to minimize the

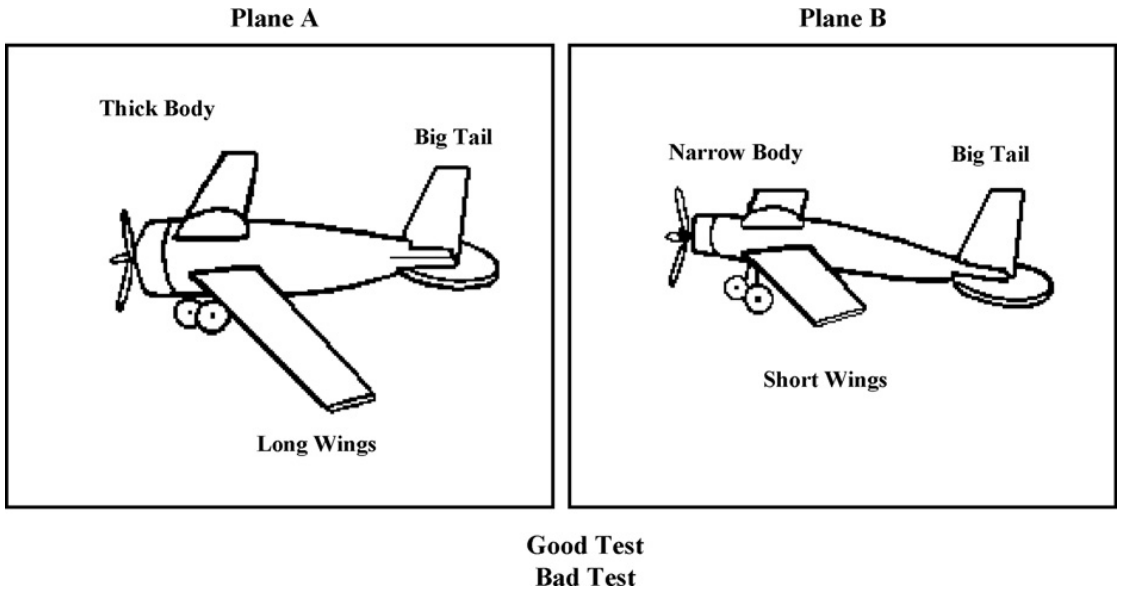

Fig. 2. Sample page from written pretest. This example has a single confound because the body type is confounded with the focal variable (wing length).

effect of different reading levels, the experimenter orally and pictorially explained the first question while children followed along on their worksheets. Children were given approximately 3 min to answer the question, and then the next question was presented. The experimenter responded to clarifying questions from individual students, but did not provide answers to test items. Classroom teachers were discouraged from assisting students in any way.

2.3.1.3. Measures for Phase 0. Test items were scored as correct or incorrect and a total score was assigned to each child. Scores for the full sample of participants who took the test ($n = 68$) ranged from 0 to 9 of a possible maximum of 9. Scores for the sample of children included in the majority of analyses ($n = 57$) ranged from 0 to 8 of a possible maximum of 9.

2.3.2. Session A (Phases 1 and 2)

2.3.2.1. Materials for Phases 1 and 2. Materials included two wooden ramps, each consisting of an adjustable downhill slide meeting a slightly uphill, stepped surface. For each ramp, the child could set the steepness (high or low), the surface (rough or smooth), the length of the downhill run (long or short), and the type of ball (rubber or golf) (see Fig. 3).

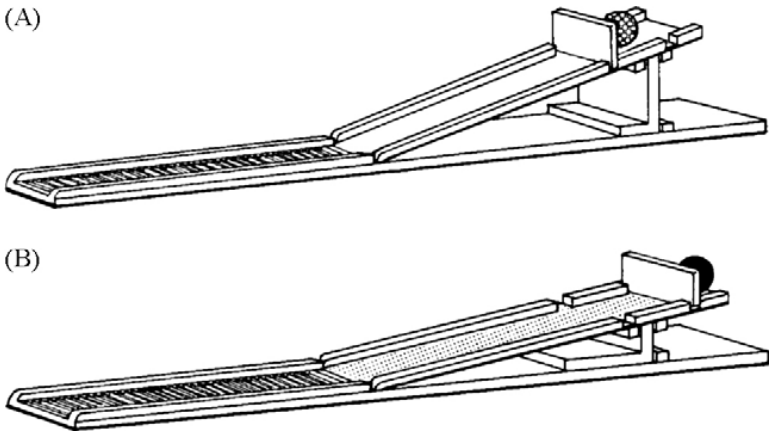

Fig. 3. The ramps used during Phases 1, 2 and 4. On each of the two ramps, children could vary the steepness, surface, and length of the ramp, as well as the type of ball. The confounded experiment depicted here contrasts (a) a golf ball on a steep, smooth, short ramp with (b) a rubber ball on a shallow, rough, long ramp.

2.3.2.2. Procedure for Phases 1 and 2. Phases 1 and 2 were conducted with individual children by Experimenter 2. In Phase 1a, the child was familiarized with the ramps and then his or her initial domain knowledge was assessed by asking which of the two settings for each of the four factors would cause a ball to roll the farthest after leaving the downhill ramp (e.g., "Do you think that a high ramp or a low ramp will make the ball roll farther?").

Following this introduction, the ramps pretest was presented. The experimenter provided a goal (e.g., "Please set up the two ramps to find out whether steepness affects how far balls roll up the steps.") and the child constructed both ramps. Once the experimenter had recorded the child's constructions, she asked the child "Why did you set up the ramps the way you did?" After answering, the child rolled the balls and then answered additional probe questions: "What did you find out?", "Can you tell for sure from this comparison whether steepness makes a difference in how far balls roll?", and "What makes you sure?" or "Why can't you be sure?". This process was repeated for a total of two steepness trials and two run length trials.

In Phase 1b, and only in this phase, children in the Explicit Instruction condition and children in the Exploration condition engaged in two different types of interaction with the experimenter and the ramps apparatus, as follows:

(a) Children in the Explicit Instruction condition received explicit goal-directed instruction that included physical examples and a series of probe questions and explanations from the experimenter. At the outset, the experimenter explained that many variables could have an effect on how far balls roll up the steps. Then she provided four examples of experiments: two focusing on ramp steepness and two focusing on ramp surface. For each focal variable, the first example was confounded, and the second was unconfounded. More specifically, the first experiment was a bad (confounded) example because the pair of ramps differed not only in steepness but also in all other variables. The child was asked whether or not the experiment was a "smart choice" for figuring out whether steepness makes a difference in how far the balls would roll. The child was then asked whether, if it turned out that one of the balls rolled farther than the other one, the child could "tell for sure" from the comparison that it was the steepness of the ramp that made that ball roll farther.

The experimenter either confirmed the child's answer and reasoning or corrected it by explaining why it was not a smart choice.

> "In fact, you *could not tell for sure* from this comparison whether it was the ⟨steepness⟩ that made a difference in how far these two balls rolled. The reason why you can't tell for sure is that these two ramps are different in other ways, not just ⟨steepness⟩. These two ramps also have different lengths of run and different surfaces, right? And the balls on them are different. So it may be that one of them rolls farther because it has a longer run or because it has a smooth surface or because it is a golf ball. As you can see, if you compare these two ramps, you can't tell whether it is the ⟨steepness⟩ or the length of the run or the surface or the different ball that makes one roll farther up the steps than the other."

The experimenter then provided a good (unconfounded) example of an experiment targeting steepness and again asked the child whether it was a "smart choice" and whether the child could "tell for sure." The experimenter again confirmed or corrected the child with an explanation of why the unconfounded comparison was a good experiment.

> In fact, you *could tell for sure* from this comparison whether ⟨steepness⟩ matters. And the reason why you can tell for sure is that the only thing different between these two ramps is the ⟨steepness⟩, right? They have the same length of run and the same surfaces, and the balls on them are the same. So, if one of them rolled farther, you'd know that it could only be the ⟨steepness⟩ of the ramp that made the difference, since it's the only thing different between these two ramps.

The experimenter repeated the cycle of demonstration, probes, and instruction with run length as the target variable and finally summarized the CVS approach and reasoning.

> Ok, so now you know how to make good comparisons with the ramps. The examples that we started with were not good, but we fixed them so that they were good comparisons, right? And now you know that if you are going to see whether something about the ramps makes a difference in how far the balls roll up the steps, you need to make ramps that are different in only one way. You want ramps that differ only in the one thing that you are trying to figure out whether it makes a difference. Like in the example we just did, we made the ⟨run⟩ different, but we made sure that the other things, like steepness and surface were the same so we could tell for sure if having a ⟨long or short run⟩ made a difference. And, you want to use the same kind of balls on both ramps. Only when you make those kinds of comparisons can you really tell for sure if that thing makes a difference.

Throughout the instruction, the balls were never rolled down the ramps. That is, the instruction provided only experimental set-ups of, not outcomes from, confounded and unconfounded experimental designs.

(b) Children in the Exploration condition continued to build pairs of ramps in the manner described in Phase 1a, but they received neither instruction about good and bad experiments nor any probe questions. Given a specific goal to find out about the effect of a particular variable, Exploration children set up the ramps and ran the experiment – rolling the balls down the ramps and observing the outcomes. (Recall that Explicit Instruction children did not run experiments during this phase.) In order to compensate for the extra time required for the child–experimenter discussions in Explicit Instruction, Exploration children completed 8 trials in the following fixed order: 2 steepness; 2 run length; 2 surface; 2 run length.

The controversy about the defining properties of instructional procedures such as "hands-on science", "direct instruction", "discovery learning", and "inquiry based science instruction" (EDC, 2006; Hmelo-Silver et al., 2007; Kirschner et al., 2006; Klahr, Triona, & Williams, 2007; Kuhn, 2007; Ruby, 2001; Schmidt et al., 2007), makes it important to articulate both the common and the distinct features of the two conditions used in the present investigation. As shown in Table 1, the conditions differed along several dimensions. Knowing which one(s) of these dimensions are responsible for differences in outcomes is not possible; indeed, it was not the point of the study. Given that our goal was to compare two educationally-realistic instructional strategies, we take the conditions in their entirety as our level of analysis.

Note that, as indicated in Table 1, children in *both* conditions were engaged in physical manipulation of the apparatus. During the ramps pre- and posttests children in both conditions set up ramps, rolled the balls, and took apart the ramps. During the training manipulation (Phase 1b), Exploration children continued to set up pairs of ramps and observe the outcome of their experiments whereas Explicit Instruction children helped take apart the experimenter-constructed ramps. Thus, both conditions involved a type of "hands-on" science instruction. In addition, in both conditions, children participated in goal-directed investigations in which the overall goal – to find out about the effect of a single causal variable – was generated by the experimenter, not the child. In *neither* condition were children unguided with respect to the purpose of the activity.

In Phase 2, children's CVS skills and domain knowledge were assessed again. For this immediate ramps posttest, the procedure was nearly identical to the pretest (Phase 1a), except that there were two trials for surface instead of steepness (and two for run length, as in Phase 1a). Finally, in Phase 2, domain knowledge was assessed *following* the test, rather than prior to it (see Fig. 1). Session A (Phases 1a, 1b, and 2) lasted approximately 45 min.

2.3.2.3. Measures for Phases 1 and 2. Each experiment designed by a child was scored according to whether it was confounded or unconfounded. Thus, total scores for each of these two phases could range from 0 to 4.

2.3.3. Session B (Phase 3)

2.3.3.1. Materials for Phase 3. Two posters were created for this session. One was a minor modification of the "Memory" poster used by Klahr and Nigam (2004) and the other was a new "Jump rope" poster

Table 1
Common and distinct features of Explicit Instruction condition and Exploration condition in Phase 1b.

	Aspect	Training condition	
		Explicit Instruction	Exploration
Common features	Materials	Pair of ramps and balls apparatus	Pair of ramps and balls apparatus
	Goal setting	By Experimenter: "can you find out whether X makes a difference in how far the ball rolls?"	By Experimenter: "can you find out whether X makes a difference in how far the ball rolls?"
Distinct features	Physical manipulation of materials by child	Child assisted in taking down ramps after each set up by Experimenter[a].	Child set up ramps, rolled ball, and took down ramps from self-designed experiments.
	Number of experiments designed	4	8
	Focal dimensions	Steepness (2 experiments) and run length (2 experiments)	Steepness (2 experiments), run length (4 experiments), and surface (2 experiments)
	Design of each experiment	By Experimenter: one "good" (unconfounded) and one "bad" (confounded) experiment for each variable under consideration	By child: child designed experiment to determine effect of focal variable chosen by experimenter
	Probe questions	Experimenter asked about whether experiment was a "smart choice" or not, and whether (hypothetical) outcome of experiment would "let you know for sure" about causal variable.	No probe questions
	Explanations	Experimenter explained why an experiment was good or bad and how it could be corrected.	No explanation
	Summary	Experimenter summarized CVS logic	No summary
	Execution of experiments	None[a]	By child
	Observation of outcomes	None: child only observed and discussed set up and a possible outcome	Child observed outcome of each experiment
	Exposure to good and bad experiments	One good and one bad experiment (identified as such by Experimenter) for each focal variable	Varied according to child (because there was no feedback from Experimenter as to good or bad design).

[a] In Phase 1b, and *only* in Phase 1b, children in the Explicit Instruction condition did not "run" their experiments. In Phases 1a, 2 and 4, children in both conditions ran every experiment that they set up.

created for this study. Posters were designed to exemplify science fair posters typical of children in this age range. Each poster described an empirical study that had a specific goal (i.e., to see if girls had better memories than boys; to see if having someone cheer for you made you jump rope better). Both posters bore titles stating the research question (i.e., "Who has a better memory? Boys or girls?"; "Do cheerleaders affect how kids do at jump-roping?") and displayed brief texts describing the hypothesis, procedure, materials, results (presented graphically) and conclusions from the study. An important feature of these posters is that they each described highly imperfect experiments, thus affording opportunities for wide-ranging evaluations.

2.3.3.2. Procedure for Phase 3. The poster evaluation session was conducted approximately 1 week after Session A and lasted approximately 45 min. Children were interviewed individually. Experimenter 1, who was blind to the training condition to which children had been assigned in Phase 1b, opened the session as follows:

> I am going to show you posters today of two experiments that were done in another school's science class. The students who did these experiments want to make their experiments good enough to enter their posters in a state-level science fair contest. I am asking you and some other students to look at these posters and tell me what's good about them and also what these students could do to make their experiments even better, since they want to enter a competition. So, I'd like to hear all your ideas about anything you know about good experiments or anything you've learned in science class that would be really helpful. They'll think about your suggestions and redo their experiments and their posters based on what you say.

The experimenter then explained the first of the two posters by carefully reading the poster text and pointing out its features, including tables, graphs, and objects used in the study. The memory poster was attributed to a girl, and the jump rope poster to a boy, and poster order presentation was counterbalanced.

Following this introduction, the experimenter began the semi-structured interview with general questions, for example, "What did Matt do well during this experiment?" and "Do you have any suggestions that would make it a better experiment or poster?" Next, questions aimed at focusing the child's attention on particularly troublesome aspects of the poster were asked in increasingly explicit ways. For example: "Is there anything about how the experiment was set up or the materials used that might have caused these results to be wrong?" then "All the kids jumped rope *with* cheerleaders before they jumped rope *without* cheerleaders. Is there anything about jumping two times with cheerleading before two times without cheerleading that could have made a difference?" The child was asked to justify her answers, except those for which she had already provided explicit reasoning or for which she had no response. For consistency's sake, for each question, this "why" prompt was only given once, regardless of whether the child took the opportunity to explain herself. When necessary, however, children were asked to clarify their answers. Near the end of the interview, the experimenter again asked general, open-ended questions targeting both the poster and the experiment. The final questions were of the form, "If Matt re-does this experiment, what would you see as the one most important thing he could do differently" and "Is there anything else you'd like to say about this poster?"

The entire process was then repeated for the second poster; questions targeted similar aspects of experimental design, implementation, analysis, inference, and presentation, thus were parallel to the first poster.

2.3.3.3. Measures for Phase 3. Poster evaluation scores were based on a coding scheme that extended and refined the coding used by Klahr and Nigam (2004). There were five broad coding categories (see Appendix A): Adequacy of the research design, Controlling for confounding variables, Measurement, Inferences, and Completeness of conclusion. Most were comprised of 1–4 subcodes. Additional codes were applied (e.g., "Communication of results," "Suggestions," "Predictions for outcomes of suggested studies") but were not theoretically relevant to this paper and thus are not included in our analyses.

Transcripts of the poster evaluations were coded by Experimenter 1 who remained blind to condition. Credit was given for each new substantive criticism or commendation; thus the same code could be given for more than one comment made by a child. Because children differed in talkativeness, explicitness, and understandability, codes were applied on a "per idea" rather than a "per utterance" basis. Regardless of where in the semi-structured interview the comment arose, credit could be given for any code. This was important because children often made comments during open-ended questions that would be elicited by later, more specific questions, and because children often provided additional responses to earlier questions throughout the interview. A second coder, blind to condition, independently coded a sample of the transcripts randomly selected from 25% of the children in each condition. Reliability – calculated by dividing the number of coder agreements by the total number of possible agreements (agreements + disagreements) – was 89.1% for the memory poster and 85.7% for the jump rope poster.

For each poster, each child received a CVS-only poster score (equal to the number of "controlling for confounding variables" codes), and a Non-CVS poster score (equal to the sum of codes from the remaining four categories). These scores were separated so that in subsequent analyses we could examine children's poster evaluation comments that were at two different transfer "distances" from the

basic CVS skills. We consider comments about confounded variables, for example, to reflect knowledge similar to that which is necessary to design an experiment in Phases 1, 2, and 4, whereas comments about measurement and inferences based on effect sizes are more distant. For each child, we calculated a Grand poster score (the sum of the CVS-only and Non-CVS poster codes).

2.3.4. Session C (Phases 4, 5, and 6)

2.3.4.1. Materials for Phases 4, 5, and 6. Phase 4 used the same ramps and balls apparatus as Phase 1. Phase 5 used a third science fair poster that had not been used in Phase 3. It was a slight modification of the Klahr and Nigam (2004) "Ping Pong" poster that described a study of whether the number of holes in ping-pong balls affected the distance that they would travel through the air. Phase 6 used a new paper-and-pencil test consisting of three multiple-choice questions about confounded and unconfounded designs in the domains of airplane design, ramps and weighted carts, and plant growth. The questions were minor variants of those used in Phase 0 and exemplify typical items relating to experimental design from "high-stakes" assessments such as NAEP and TIMMS. These items have high external validity with respect to conventional science assessments (given their types, format, and content).

2.3.4.2. Procedure for Phases 4, 5, and 6. Approximately 3 months ($M = 91$ days; $SD = 8$ days) after the completion of Session A, each child participated in Session C, which included Phase 4 (delayed ramps posttest), Phase 5 (a third science fair poster interview), and Phase 6 (written posttest) (see Fig. 1).

Phase 4 began with Experimenter 1 familiarizing children to the ramp apparatus by reminding them of its variables and settings. Current domain knowledge was then assessed by asking children about which settings for each variable would make the ball roll further after leaving the downhill ramp. To complete the phase, children participated in another ramps posttest which assessed their experimental CVS skills. This assessment was identical to the ramps pretest and immediate ramps posttest (Phase 1a and 2), except that – as shown in Fig. 1 – there was one trial for each of the four variables (steepness, surface, run length, and ball).

Phase 5 consisted of a third interview eliciting a science fair poster evaluation (regarding an experiment said to be conducted by a child with the same gender as the participant). The task was re-introduced by saying "Since the children at your school did such a great job of helping the last poster-makers improve their experiments, another child wants his/her poster critiqued before he/she re-does it for the science fair." The structure and general focus of interview questions was identical to the previous interviews in Phase 3.

Finally, children completed a 3-question written posttest about CVS (Phase 6). As with the written pretest, the experimenter read through each multiple-choice problem as the child read along. Unlike the pretest, however, children were asked to justify their answers. If a child wanted to change her answer before or after justifying it, she was allowed to do so, but had to justify the new answer as well (note that "I just guessed" was accepted as justification). After all three questions had been fully answered, the experimenter revisited any questions for which the child's justification was ambiguous (i.e., applied to more than one of the answer choices) and gave the child an opportunity to revise her answer. Fewer than 10% of the children changed any answer. Our analysis is based on children's final replies.

2.3.4.3. Measures for Phases 4, 5, and 6. Phase 4 was scored in the same way as Phases 1a and 2. Phase 5 was scored in the same way as Phase 3. Reliability was calculated for 17 (24%) of the ping pong poster evaluations (a different subset of participants than were used to calculate reliability for Phase 3) and was 89.7%. Each question on the Phase 6 written posttest was scored as correct or incorrect for a maximum score of 3.

2.3.5. Three-year follow-up (Phase 7)

2.3.5.1. Materials for Phase 7. We used a two-part paper-and-pencil test for this phase. Part 1 included four items taken directly from high-stakes elementary school science tests (e.g., TerraNova) and Part 2 included six researcher-designed items depicting experiments in three domains (lemonade stands, rockets, and baking cookies) similar in nature to the "airplane" questions from the pretest. Six different types of comparisons were presented in Part 2: (a) unconfounded; (b) singly confounded; (c)

multiply confounded; (d) non-contrastive; (e) unconfounded comparison of a non-focal variable; and (f) confounded comparison of a non-focal variable.

2.3.5.2. Procedure for Phase 7. Of the 72 children who had participated in Phase 1, 43 were available to participate in this 3-year follow-up phase (24 from the Explicit group and 19 from the Exploration group). This included 7 Explicit and 5 Exploration children from Grade 3 (now Grade 6), 7 Explicit and 10 Exploration children from Grade 4 (now Grade 7), and 10 Explicit and 4 Exploration children from Grade 5 (now Grade 8). The paper-and-pencil test was administered by an experimenter or the science teacher in the six participating classrooms. In order to be true to the nature of the standardized test items, Part 1 was completed by children on their own. For the researcher-designed items (Part 2), however, we eliminated the potential effect of different reading levels by reading the test items aloud and by giving students time to answer each question before moving to the next question. In addition, the administrator responded to clarifying questions from individual students, but did not provide hints or answers to test items.

2.3.5.3. Measures for Phase 7. Part 1 items were scored using the rubric from the standardized tests. For Part 2, children received one point for correctly identifying an experiment as a good or bad way to find out about the focal variable and another point for creating (or maintaining) an appropriate unconfounded experiment. Thus, scores on Part 1 could range from 0 to 6, and scores on Part 2 from 0 to 12.

3. Results

3.1. Pre-instructional equivalence of treatment groups

The equivalence of the treatment groups was checked by comparing their general CVS knowledge, as assessed by the Phase 0 written pretest scores, and their Phase 1a CVS scores. There were no differences on either of these measures between the Explicit Instruction and Exploration groups.

3.2. Instructional effectiveness: ramps assessments

3.2.1. Acquisition of CVS Expertise by individual students in each training condition

We first describe the extent to which Explicit Instruction and Exploration differ in instructional effectiveness as reflected by student performance in the ramps domain during Phases 2 (immediate ramps posttest) and 4 (delayed ramps posttest). Recall that in Phases 1, 2, and 4 (see Fig. 1), children had four opportunities to design an unconfounded experiment. We classified individual children as "Experts" if they designed an unconfounded experiment on all four trials in a phase (i.e., perfect performance). Although, as noted earlier, other studies have used several different criteria for assessing acquisition of CVS, in this analysis we focus on expert performance because it has high external validity. That is, a rigorous and meaningful measure of effective CVS instruction is the extent to which children are able to design unconfounded experiments on 100% of their attempts, rather than some arbitrary lesser proportion.

In Phase 1 (pretest) 11 children – 3 in the Explicit Instruction condition, and 8 in the Exploration condition – designed an unconfounded experiment on all four trials and thus were deemed "Natural Experts." Unless otherwise specified, these children are excluded from further analysis. Thus, the analyses presented in this section are based on the remaining 61 participants. Among those 61 children, there was no significant difference in Phase 1 between the Exploration and Explicit groups in the distribution of CVS scores of 0, 1, 2, and 3. In Phase 2, 59% (19 of 32) of the Explicit Instruction children, but only 10% (3 of 29) of the Exploration children were CVS Experts, $\chi^2(1, N = 61) = 15.9$, $p < .001$. This difference in Expert distribution between the two conditions in the number of CVS Experts remained after 3 months, in Phase 4, $\chi^2(1, N = 61) = 4.1$, $p = .04$, albeit with a non-significant reduction of the proportion of Experts in the Explicit Instruction condition to 53% (17 of 32) and a marginally significant increase in the number of Experts in the Exploration condition to 28% (8 of 29), $\chi^2(1, N = 29) = 2.8$, $p = .09$.

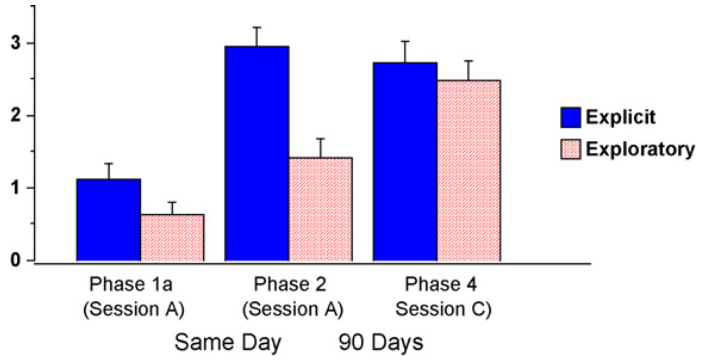

Fig. 4. Mean CVS scores for children in Explicit and Exploratory conditions in Phases 1a, 2, and 4. (Phases 1a and 2 occurred on the same day; Phase 4 occurred 90 days later). Maximum CVS score = 4.

3.2.2. Mean CVS scores in each training condition

Our analysis thus far has classified individual children according to a simple two- way classification of their performance level (Expert or not) and it suggests an immediate effect of training in Phase 2, which is maintained into Phase 4. However, when we shift from classifying individual children to a comparison of mean performance scores in each condition, we get a different and unexpected result. (Recall that in each phase, a child's score could range from 0 to 4.) The immediate effect of training type (between Phase 1 and Phase 2) was analyzed by subjecting these scores to a 2 (training type) X 3 (grade) X 3 (phase) repeated measures ANOVA, with phase as a within-subjects variable. The analysis revealed a main effect for training condition, $F(1, 55) = 7.0$, $p < .01$, for grade, $F(2, 55) = 7.8$, $p = .001$, and for phase, $F(2, 110) = 46.1$, $p < .001$. There was a phase by condition interaction, $F(2, 110) = 7.9$, $p < .001$, and no other interactions. As shown in Fig. 4, mean CVS scores for children in the Explicit Instruction condition increased dramatically from Phase 1 to Phase 2: from 1.1 to 2.9, $t(31) = 5.8$, $p < .001$. Scores for children in the Exploration condition also increased between Phases 1 and 2, but much less so: from 0.62 to 1.4, $t(28) = 3.1$, $p = .004$. Thus, the Phase 1 to Phase 2 results of the ANOVA are consistent with the Chi-square analysis based on individual children. However, Fig. 4[2] also shows that by Phase 4, 3 months after initial training, the mean CVS score for children in the Exploration condition increased to a level that is indistinguishable from the mean CVS score of children receiving Explicit Instruction, which remained essentially unchanged, $t(59) = 0.6$, $p = .56$. We further explored this finding by examining the ramps posttest scores in each phase for only the "weakest" children: those who constructed fewer than 2 of 4 unconfounded experiments in Phase 1. The results were essentially the same as for the full set. There were significant differences between instructional groups in Phase 2, and no differences in Phase 4.

The source of this unanticipated gain in the mean CVS score for the Exploration group can be identified by returning to the classification of individual children. For this analysis, in addition to the Expert classification (4 of 4 correct), we classified children as "Masters" if they were correct on 3 of the 4 trials; otherwise they were classified as "Novices." Then we computed the proportion of children who moved from one category to another between Phases 1 and 2 and between Phases 2 and 4. For example, in the Explicit Instruction condition, between Phase 1 and Phase 2, 54% of the Novices and 83% of the Masters became Experts, while in the Exploration condition, only 11% of the Novices became Experts and none of the Masters became Experts. Moreover, 78% of the Novices in the Exploration condition remained Novices, while only 31% of those in the Explicit Instruction condition did so. This, of course, is consistent with the difference in immediate (Phase 2) mean scores between conditions in Fig. 4.

The Phase 2 to Phase 4 transition percentages provide insight into the significant increase in mean scores for the Exploration children revealed by the ANOVA. Between Phases 2 and 4, 50% of the Phase 2 Exploration Novices advanced to the Master level and 9% advanced to the Expert level. At the same time, 75% of the Phase 2 Exploration Masters advanced to Expert classification. These many "uninstructed"

[2] It is important to note that Fig. 4 does not illustrate what we mean by "path independence." Fig. 4 shows that there are two different time courses to mastery level. But it does not include any assessment of whether children reaching expertise from one form of training or another will have different results on a far transfer test. In the following section we address that issue.

advances in the Exploration group, combined with a few regressions in the Explicit group from Expert to Master (11%) or Novice (26%), and from Master to Novice (20%), produced equivalent mean CVS scores between the two training groups by Phase 4 (as shown in Fig. 4).

Why did many children in the Exploration group advance to Expert or Master categories over the 3-month delay? One possible explanation is simply that CVS instruction – in one form or another – might have been included in the school science lessons received by these children at some point during the 3-month interval. We think that this is unlikely. The teachers in all of the relevant classes claimed that they did not teach anything closely related to CVS during the interval. In addition, in order to determine the extent to which this unexpected gain was grade specific, we performed paired t-tests on the Phase 2 to Phase 4 scores for each grade. The increase in mean CVS scores for children in the Exploration condition was significant for third graders, $t(7) = 3.9$, $p = .006$, and fifth graders, $t(9) = 3.3$, $p = .009$, but not for fourth graders, $t(10) = 1.8$, $p = .096$. This finding tends to rule out the "additional instruction" explanation for the increase, as it is unlikely that such instruction – teacher claims to the contrary notwithstanding – would have produced similar effects for third and fifth graders.

3.3. Performance as a function of transfer distance

3.3.1. Explicit Instruction/Exploration group differences

3.3.1.1. Poster evaluation after a 1-week interval (Phase 3). Recall that in Phase 3, conducted approximately 1 week after the CVS assessments in Phase 2, children were asked to evaluate two science fair posters during a structured interview. Their responses were scored according to the procedure described earlier and the scores for the two posters were averaged for further analysis. These mean Grand poster scores ranged from 0 to 13 ($M = 6.3$, $SD = 3.1$). Because the poster scores were based on children's verbal responses to extensive questioning, we included grade in the analysis. A 2 (training condition) × 3 (grade) ANOVA on Grand poster scores in Phase 3 revealed that training condition had no effect, but that (unsurprisingly) grade did, $F(2, 55) = 10.9$, $p = .001$, and there was no interaction between grade and condition. Post hoc Bonferonni tests produced significant pair-wise differences ($p < .05$) between the fifth grade poster scores and the other two grades (fifth graders had higher scores), but no significant difference between the third and fourth grade scores ($p = .26$).

We next examined Phase 3 poster component scores separately using the same analyses. There was no main effect of condition for either the CVS-only poster score or the Non-CVS poster score, but there was a main effect of grade for both CVS-only score, $F(2, 55) = 4.85$, $p = .011$, and Non-CVS score, $F(2, 55) = 11.77$, $p < .001$. Specifically, for the CVS-only poster score, third graders performed significantly worse than fifth graders ($p = .004$); whereas for the Non-CVS poster score, third and fourth graders both performed significantly worse than the fifth graders ($p \leq .01$). There was no condition by grade interaction.

3.3.1.2. Poster evaluation after a 3-month interval (Phase 5). In Phase 5, conducted approximately 3 months after the Phase 3 poster evaluations, children were asked to evaluate a new science fair poster. Grand poster scores, based on responses to the same type of structured interview used in Phase 3, ranged from 0 to 14 ($M = 5.9$, $SD = 3.7$). We again examined the relation between type of training in Phase 1 and Grand poster scores in Phase 5. A 2 (training condition) × 3 (grade) ANOVA on Phase 5 Grand poster scores revealed a main effect for grade, $F(2, 53) = 15.15$, $p < .001$, and a grade by condition interaction, $F(2, 53) = 3.92$, $p = .026$, but no effect of training condition. Post hoc Bonferonni tests produced significant pair-wise differences ($p < .001$) between the fifth grade poster scores and the other two grades (fifth graders had higher scores), but no significant difference between the third and fourth grade scores. Exploring the interaction through *t*-tests, it became clear that third graders in the Explicit Instruction condition far outperformed those in the Exploration condition (5.33 vs. 2.25, respectively), $t(15) = 2.23$, $p = .041$.

Again we examined Phase 5 poster component scores separately using the same analyses. There was no main effect of condition for either the CVS-only poster score or the Non-CVS poster score, but there was a main effect of grade for both CVS-only, $F(2, 53) = 7.80$, $p = .001$, and Non-CVS, $F(2, 53) = 15.75$, $p < .001$. For each, post hoc Bonferonni tests revealed that third and fourth graders both performed significantly worse than the fifth graders ($p < .05$). There was a grade by condition interaction for

the Non-CVS poster scores, $F(2, 53) = 6.19$, $p = .004$: For the third graders only, those in the Explicit Instruction condition far outperformed those in the Exploration condition (3.22 vs. 1.13, respectively), $t(15) = 2.63$, $p = .019$.

3.3.1.3. Written posttest after a 3-month interval (Phase 6). Scores on the 3-month written posttest ranged from 0 to 3 ($M = 1.7$, $SD = 1.1$). There were no significant training group differences for the written posttest scores: A 2 (training condition) × 3 (grade) ANOVA on the 3-month written posttest scores in Phase 6 revealed that neither training condition nor grade had an effect.

3.3.1.4. Written posttest after a 3-year interval (Phase 7). Scores on the 3-year written posttest ranged from 4 to 18 ($M = 11.1$, $SD = 4.2$). There were no significant training group differences: A 2 (training condition) × 3 (grade) ANOVA on the 3-year written posttest scores in Phase 7 revealed that training condition had no effect, but that grade did, $F(2, 37) = 5.01$, $p = .012$, and there was no interaction between grade and condition. Post hoc Bonferonni tests produced significant pair-wise differences ($p < .05$) between the eighth grade (formerly-fifth grade) poster scores and the other two grades (eighth graders had higher scores).

3.3.2. CVS Expertise-based analysis of task performance

Recall that there was a change between Phase 2 and Phase 4 in the population of children in the Expert category. Between phases, some children in the Explicit Instruction condition lost Expertise, whereas others gained Expertise, and many children in the Exploration condition gained Expertise. Thus, we might expect the Phase 4 Experts to, overall, possess more entrenched and robust CVS knowledge. This is not discernible in the analyses done thus far, since we have used only training condition (Explicit or Exploration) as the independent variable in our analysis of mean CVS scores, poster scores, and written posttests scores. In this next analysis we use the Expert/non-Expert classification in Phases 2 (immediate ramps posttest) and 4 (delayed ramps posttest) as post-hoc factors in order to determine the extent to which CVS Expertise – independent of training condition – is related to each of these outcome measures. In other words, children's performance during these two CVS phases, combined with the two poster evaluation phases (Phases 3 and 5) and the two phases in which written tests were administered (Phases 6 and 7), provide opportunities for several different comparisons between CVS Expertise and performance on the other measures. The comparisons represent different transfer "distances" along the transfer dimensions of task similarity and temporal interval (Chen & Klahr, 2008). Results are described below and summarized in Table 2. For

Table 2
CVS Expertise (at Phase 2 and Phase 4) as a predictor of other performance measures: Phase 3 poster scores, Phase 4 CVS scores, Phase 5 poster scores, Phase 6 written posttest, and Phase 7 written posttest.

Phase	Measure	Phase 2 Expert/non-Expert	Phase 4 Expert/non-Expert
Phase 3	Poster scores	(Delay: 1 week)	
	Grand	$p = .003$	n/a
	CVS-only	$p = .044$	n/a
	Non-CVS	$p = .002$	n/a
Phase 4	Ramps	(Delay: 3 months)	
	Mean CVS	$p = .027$	n/a
Phase 5	Poster scores	(Delay: 3 months)	(Delay: none)
	Grand	n.s.	$p < .001$
	CVS-only	$p = .007$	$p < .001$
	Non-CVS	n.s.	$p < .001$
Phase 6	Written posttest	(Delay: 3 months)	(Delay: none)
		n.s.	$p < .001$
Phase 7	Written posttest	(Delay: 3 years)	(Delay: 3 years)
		n.s.	$p < .001$

Note: Each cell indicates the delay between the Expertise measure (column heading) and the phase, and whether Expertise was a significant predictor ($p < .05$) of each performance measure (row headings) in that phase.

these analyses we compared the mean scores for Experts and non-Experts using simple unpaired t-tests.

3.3.2.1. Phase 3 poster scores. This analysis asks whether Expertise in Phase 2 predicts poster performance – 1 week later – in Phase 3. It does: Phase 2 CVS Experts performed significantly better than Phase 2 non-Experts on the Phase 3 Grand poster score, $t(59) = 3.07$, $p = .003$, as well as on its two components: the CVS-only poster score, $t(59) = 2.06$, $p = .044$, and the Non-CVS poster score, $t(59) = 3.18$, $p = .002$. In terms of transfer distances, this is an assessment of far transfer with respect to task similarity (CVS experiments vs. poster evaluation scores), and moderate transfer with respect to temporal interval (1 week). Note that the terminology we use in characterizing transfer distance is necessarily approximate, subjective, and ordinal, at best. Moreover it provides only a weak basis for comparisons between one study and another (Barnett & Ceci, 2002).

3.3.2.2. Phase 4 CVS Expertise. Phase 2 Expertise predicts Phase 4 mean CVS scores, $t(59) = 2.26$, $p = .027$. This is very near transfer with respect to task – indeed it is a nearly identical task except that two new ramp dimensions are used in Phase 4 (ball and steepness) that had not been used in Phase 2 – but it is moderate in terms of time delay (3 months).

3.3.2.3. Phase 5 poster scores. Phase 2 Expertise predicts only the CVS component of Phase 5 poster scores 3 months later, $t(33) = 2.89$, $p = .007$. Thus, although the strong relation between CVS Expertise and poster scores diminishes slightly during this interval, it remains evident in the most pertinent respect. The fact that the CVS-only poster scores are predicted by Phase 2 CVS Expertise, while the Non-CVS poster scores are not, argues against the possibility that CVS Experts are simply better students. If this were so, then CVS Expertise would predict both components of the poster score equally well. Moreover, when Expertise is defined by the Phase 4 CVS performance – which is measured in the same time period as the Phase 5 poster evaluation (see Fig. 1) – then once again Expertise is a strong predictor of all poster score components: Grand poster score, $t(57) = 5.57$, $p < .001$), CVS-only poster score, $t(57) = 5.68$, $p < .001$), and Non-CVS poster score, $t(57) = 4.04$, $p < .001$).

3.3.2.4. Phase 6 written posttest. Phase 2 Expertise was unrelated to written posttest scores in Phase 6 (3-month delay). However, Phase 4 Expertise – assessed in the same time period as the Phase 6 written posttest – was a significant predictor, $t(56) = 4.42$, $p < .001$). That is, children who were Phase 4 CVS Experts did better on the Phase 6 written posttest than non-Experts. In this case the temporal transfer distance is essentially zero, while the task transfer distance (between physical apparatus involving ramps and a written paper and pencil test on unconfounded experimental design issues) could be considered as moderate to far.

3.3.2.5. Phase 7 written posttest. Similarly, 3-year written posttest scores were unrelated to Phase 2 Expertise, but were strongly predicted – after a 3-year delay – by Phase 4 Expertise, $t(41) = 4.84$, $p < .001$. That is, children who were Phase 4 CVS Experts did better on the written posttest than non-Experts. This final analysis represents very far temporal transfer (3 years) combined with moderate to far task differences.

3.3.2.6. Summary of Expertise-based analyses. Acquisition of CVS Expertise during training was a good predictor of performance on near and moderately far-transfer tasks after a short interval (i.e., Phase 2 Expertise to Phases 3 and 4 measures, and the CVS-only component of the poster evaluation in Phase 5.) However, for longer delays, and less task similarity, only Expertise identifiable several months after initial training (i.e., Phase 4 Expertise), was predictive of far-transfer performance (i.e., the full array of Phase 5 poster scores, and the Phase 6 (no delay) and Phase 7 (3-year delay) written tests).

3.4. Path independence?

In this section we turn to a final important question: to what extent does performance on various transfer tasks depend on children's learning paths? That is, for a given level of knowledge, does *how*

children acquired that knowledge have an impact on their ability to transfer the knowledge to new tasks and domains? Klahr and Nigam's (2004) path-independence hypothesis predicts that "*if* children achieve mastery of a new procedure ... *then* the way that they reached that mastery has no effect on their ability to transfer what they have learned." In contrast, discovery learning advocates would make the opposite prediction, i.e., that knowledge acquired under minimally guided instruction leads to better transfer performance than equivalent knowledge acquired under highly directive instruction. In this section we re-visit that issue. The operational question is whether or not children who acquire CVS via different "routes" (i.e., acquire CVS in the Exploration condition or in the Explicit Instruction condition) perform differently on our transfer measures. Given the multi-phase design of the present study, there are several paths than can be examined, and the outcome of a path-independence analysis can vary among these paths. In the following sections, we first describe the procedure we use to define and assess path independence, and then we apply that assessment procedure to several different paths. Results of all the analyses are summarized in Table 3.

3.4.1. Defining learning paths

Following Klahr and Nigam (2004), we defined a learning path by crossing condition (Exploration or Explicit Instruction) with the performance classification (e.g., Master or non-Master). This produced four types of learning paths (e.g., "Exploration Master," "Exploration non-Master," "Explicit Master," and "Explicit non-Master."

3.4.2. Statistical method for assessing path independence

Discovery learning proponents would likely expect the children in our Exploration condition to outperform children in our Explicit Instruction condition on a far-transfer task. Our primary motivation for the path-independence analysis was to investigate that claim. However, path independence relies on more than instructional condition and, as such, our statistical method for testing path independence uses demanding criteria. It requires that there be a main effect of learning path (e.g., defined by crossing instructional condition x performance classification) when a one-way ANOVA is conducted and that follow-up pair-wise comparisons meet the following criteria: (a) the mean scores of the Exploration Masters should not differ significantly from those of the Explicit Masters; (b) the mean scores of the Explicit and Exploration non-Masters should not differ significantly; and (c) both of the Master mean scores should be significantly different from both of the non-Master scores. The failure of any of the 6 pair-wise comparisons involved in this analysis challenges the claim of path independence.

3.4.3. Klahr and Nigam replication: 1-week poster evaluation (Phase 3)

For our first analysis of path independence, we sought to replicate, exactly, Klahr and Nigam (2004). Because they used a "Master" criterion (at least three of four unconfounded experiments), for this analysis we defined learning path by crossing training condition (Explicit Instruction or Exploration) with the Phase 2 Master or non-Master classification. We followed Klahr and Nigam by also including a fifth category of children, those who were Natural Experts (i.e., children who created four unconfounded experiments of a possible four trials in Phase 1). We replicated their analyses and findings (see Fig. 5). Specifically, a one-way ANOVA with learning path as the independent variable and Phase 3 Grand poster score as the dependent variable yielded a main effect for learning path, $F(4, 67) = 5.7$, $p = .001$. Pre-planned LSD tests (the heart of the path-independence analyses) showed that Masters – as well as Natural Experts – outperformed non-Masters, regardless of the specific learning path. Significant pair-wise differences ($p < .05$) existed between the non-Master paths of both types and all the other categories, but not among the Natural Experts and the two types of Masters, nor between the two types of non-Masters.

To reiterate, the Grand poster scores of the three Master/Natural Expert groups (those who scored well on the immediate ramps posttest via different paths or were deemed Natural Experts) were indistinguishable. The Grand poster scores of the two non-Master groups (those who scored poorly on the immediate ramps posttest, regardless of path) were indistinguishable and were significantly lower than the scores of the Masters/Natural Experts. Thus, the path-independence finding of Klahr and Nigam was replicated with the current data.

Table 3
Assessments of path independence based on CVS Expertise in Phase 2 and Phase 4.

	Phase 3 1-week poster evaluation			Phase 5 3-month poster evaluation			Phase 6 3-month written posttest	Phase 7 3-year written posttest
Path defined by	Grand poster score	CVS-only poster score	Non-CVS poster score	Grand poster score	CVS-only poster score	Non-CVS poster score	Test score	Test score
Phase 2 CVS Expertise	(*1 pair-wise violation*)	(*Omni n.s.*)	(*1 pair-wise violation*)	(*Omni n.s.*)	(*1 pair-wise violation*)	(*Omni n.s.*)	(*Omni n.s.*)	(*Omni n.s.*)
Phase 4 CVS Expertise				P-I	P-I	P-I	(*1 pair-wise violation*)	P-I

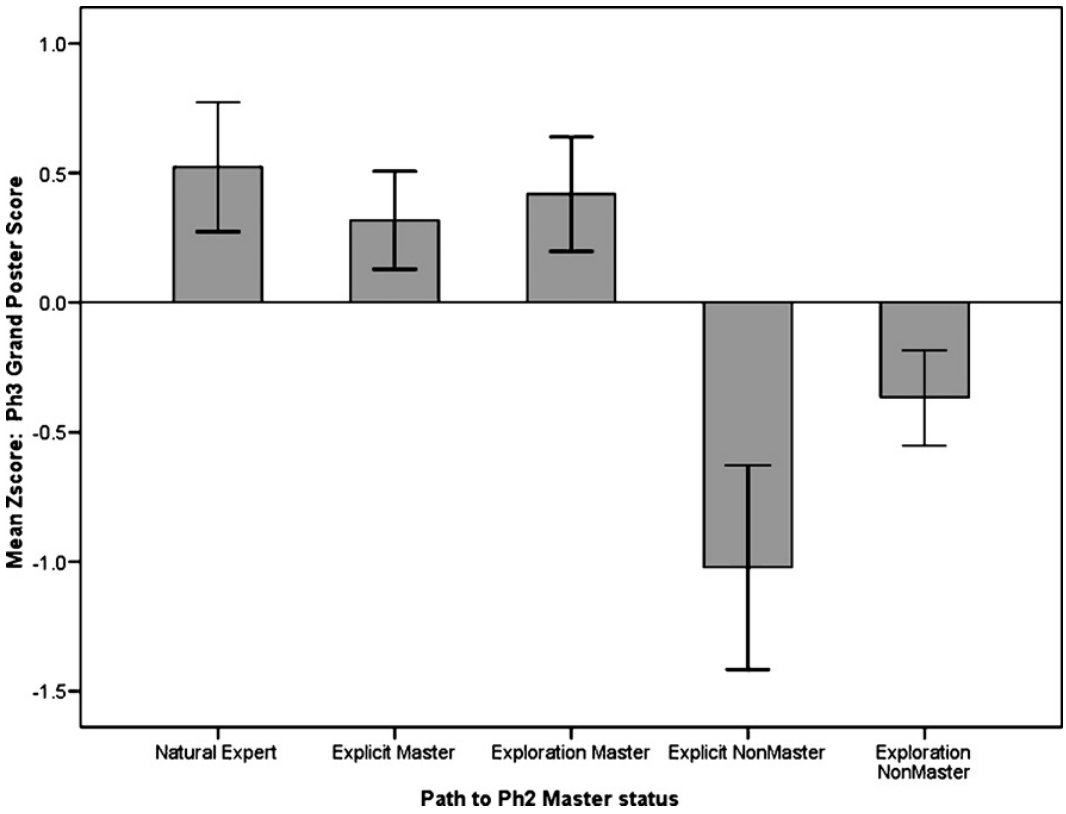

Fig. 5. Normalized Phase 3 Grand poster scores for five learning paths: Natural Experts, Explicit Masters, Exploration Masters, Explicit non-Masters, and Exploration non-Masters.

3.4.4. Path independence based on Expert level performance and different transfer distances

Our remaining analyses of path independence are based on a more stringent "Expert" criterion (four of four unconfounded experiments), for reasons described earlier. We do not include the Natural Expert category since our test of path independence hinges on an investigation of student performance following the exposure to two different well-defined instructional paths, whereas the instructional paths that led students to enter our study as Natural Experts are unknown to us.

3.4.4.1. Phase 2 Expertise and path independence. Here, we defined learning path by crossing training condition (Explicit Instruction or Exploration) with immediate ramps performance (Phase 2 Expert or non-Expert classification). This resulted in four paths: "Exploration Expert," "Exploration non-Expert," "Explicit Expert," and "Explicit non-Expert." We repeated the path-independence analysis for several transfer measures (as indicated in Table 3): poster scores in Phase 3, poster scores in Phase 5, and written posttests in Phases 6 and 7. Complete path independence based on Phase 2 Expertise was not supported in any of these comparisons, although in some cases only one of the six pair-wise comparisons failed (see row 1, Table 3).

3.4.4.2. Phase 4 Expertise and path independence. We next defined learning path by crossing training condition (Explicit Instruction or Exploration) with delayed ramps performance (Phase 4 Expert or non-Expert classification). We repeated the path-independence analysis for transfer measures (as indicated in row 2 of Table 3). This produced a very different result than just described. For Phase 5 poster scores we do find path independence. A one-way ANOVA with Phase 4 learning path as the independent variable and Phase 5 Grand poster score as the dependent variable yielded a significant main effect for learning path, $F(3, 55) = 11.03$, $p < .001$. The expected pair-wise significant differences were present ($p < .001$). This result – path independence from Phase 4 Expertise to Phase 5 poster evaluation – remains when we do the analysis on the two components of Phase 5 poster scores (CVS-only and Non-CVS). Specifically, for Non-CVS poster scores, a main effect of learning path, $F(3, 55) = 7.01$, $p < .001$, was accompanied by the expected pair-wise significant differences ($p < .05$). Similarly, for CVS-only poster

scores, a main effect of learning path, $F(3, 55) = 10.39$, $p < .001$, was accompanied by the expected pair-wise significant differences ($p \leq .001$). For the Phase 6 (3-month) written posttest, a one-way ANOVA with Phase 4 learning path as the independent variable and written posttest score as the dependent variable was significant, $F(3, 54) = 8.47$, $p < .001$. Follow-up tests revealed one pair-wise violation of the path-independence pattern. For the 3-year written posttest, we obtained "perfect evidence" of path independence, $F(3, 39) = 7.89$, $p < .001$. This was accompanied by the expected pair-wise significant differences ($p < .05$).

Overall, the path-independence analysis yielded mixed results. For the Mastery-based learning path analysis over a 1-week interval, the Klahr and Nigam findings were replicated. However, for the Phase 2 Expertise-based learning path analysis, there was no support for path independence over that same (1 week) interval, nor for any of the longer intervals. In contrast, the Phase 4 Expert-based learning path analysis *did* find path independence: for the very brief interval between Phases 4, 5 and 6, as well as for the 3-year interval to Phase 7. In simpler terms: if a child became a CVS Expert (i.e., perfect performance) by Phase 4, then his or her performance on moderate to far-transfer tasks, up to 3 years later, was, in 4 of 5 measures examined, independent of the training path that led to that Expertise. Similarly, children who did not become CVS Experts by Phase 4 performed equally poorly on the transfer tasks, regardless of which training condition they had been in.

4. Discussion

The aims of this study were to explore the immediate and longer term consequences of two different types of science instruction – Explicit Instruction and Exploration – located at different points on the direct-to-discovery spectrum, and to determine the generality of the path-independence effect reported by Klahr and Nigam (2004). This path-independence issue is of particular interest to science educators seeking evidence about the most effective instructional methods. In addition, the question has important developmental implications. For example, Adolph and Berger's (2006) elegant demonstration of the wide variety of sequences of motor acquisitions raises the question of whether the final behavior of children who follow such different paths to "expertise" will bear traces of the path by which they achieved it. A question for future research is the extent to which existence or non-existence of path independence is a function of whether the multiple pathways are imposed by instruction or naturally occurring.

One of our first challenges in considering the effects of different types of instruction was to provide clear operational definitions. In other contexts, the contrasting terms "Direct Instruction" and "Discovery Learning" have been used to characterize the collection of factors – listed in Table 1 – defined in this paper as "Explicit Instruction" on one hand and "Exploration" on the other. Some critics have claimed that our Explicit Instruction condition is, in fact, very close to what classroom teachers do when they attempt to use discovery learning approaches and that what we call "Exploration" is a parody of discovery learning that few teachers would ever use (Adelson, 2004; Crane, 2005; Tweed, 2004). Others – usually critics of constructivist approaches – see our Exploration condition as even *more* structured than much of what transpires in typical discovery learning classrooms. However, as indicated in Table 1, regardless of how our two instructional conditions are labeled or ultimately positioned on the direct-to-discovery spectrum, they are unarguably extremely different in the amount of guidance, information, support, teacher control, and feedback provided during training and thus provide strong contrasts with which to examine path independence.

Clear operational definitions are useful within studies (i.e., one can go beyond arbitrary labels to determine exactly what kinds of procedures and measures were used), but also facilitate cross-study comparisons and help determine whether studies should be compared to each other at all. The importance of this becomes clear upon careful reading of a study by Dean and Kuhn (2007) that purports to replicate important aspects of Klahr and Nigam (2004), yet finds different results. Dean and Kuhn contrast the poor performance of "Direct Instruction" students to better performance by those in the "Practice" and "Direction Instruction/Practice" conditions. However the comparison is confounded because what is labeled an "immediate posttest" is, for Direct Instruction students, a *novel* task administered *10 weeks* after their initial instruction, whereas for the other comparison groups, it is a task very similar to one they have been practicing for 10 weeks. Comparing the relative efficacy

of Direct Instruction to Explicit Instruction across the two studies becomes problematic for this and other reasons.

Dean and Kuhn (2007) find "a certain irony, if not conceptual incoherence" in a search for an empirical resolution of the most effective way to teach CVS, because it "is a component of inquiry skill and inquiry skill is broadly understood to mean skill in discovering or constructing knowledge for oneself" (p. 385). We see neither irony nor incoherence in such efforts. For example, we doubt that anyone – even Dean and Kuhn – would suggest that graduate students should discover, without substantial instruction, the cumulated foundational skills and procedures needed to pursue their own inquiries in the realm of psychological science. More specifically, they fail to acknowledge the finding, in one of our earliest investigations of CVS (Chen & Klahr, 1999), that once students had been taught CVS via direct instruction, they were able to use CVS to advance their knowledge about the relation between different values of causal variables and outcomes in the physical domains under investigation *without any explicit instruction* about that domain knowledge. To reiterate, "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" (Chen & Klahr, 1999, p. 1117).

One of the aims of the present study was to use a wider variety of transfer measures than we have used in previous studies. As Chen and Klahr (2008) argue, transfer "distance" remains an elusive and only partially defined construct, primarily because there is no theoretical basis upon which to quantify the trade-offs between, for example, two transfer "distances" that are defined by different measurement contexts (e.g., classroom vs. laboratory), different tasks, or different temporal intervals. To address this intrinsic difficulty (and critiques of Klahr & Nigam, 2004), in the present work we included transfer tasks addressing the target strategy in the same task after a delay (i.e., Delayed ramps posttest), the target strategy in different tasks and delays (i.e., CVS scores of 1-week and 3-month poster evaluations, 3-month and 3-year written posttests), and related skills in different tasks and delays (i.e., Grand and Non-CVS scores of 1-week and 3-month poster evaluations).

As in prior studies, Explicit Instruction produced many more immediate CVS Experts than did Exploration. However, at the 3-month (Phase 4) assessment, there was an unanticipated increase in the number of CVS Experts in the Exploration group, but no corresponding increase in the Explicit group. Earlier we ruled out the possibility that specific science instruction during the interval might have accounted for this discrepancy. Other possible explanations are that (a) the Phase 1b procedure for Exploration children provided sufficient variation in the focal variable to increase their sensitivity to the distinct nature of the different factors in the ramps materials, sensitivity that they were unable to consolidate in time for the Phase 2 assessments, but that they could apply by the time of the Phase 4 assessment. (Recall that, as shown in Table 1, Exploration children conducted a total of eight experiments on three different factors during training.); (b) Exploration children had a sufficient number of trials in Phase 2 to pursue a semi-systematic series of experiments and less effective strategies (e.g., "engineering approaches" to create the "best" ramps), such that by the Phase 4, all that remained to try was the CVS approach; (c) the Exploration condition stimulated the type of "preparation for future learning" identified by Schwartz and Martin (2004) in which the experience, even though it lacked any direct instruction, primed children to think about CVS related issues more broadly during the interval between Phases 2 and 4. At this point, however, such accounts remain speculative and can only be resolved by further studies.

The effects of our instructional conditions were assessed in two ways. First, we asked whether children's performance on the transfer tasks (i.e., poster assessments and written posttests) varied as a function of instruction. Since the answer was "no," this could be interpreted as a lack of transfer. Yet our second assessment, the path-independence analysis, is more fitting, since it factors in whether there was any knowledge to transfer in the first place. For learning paths defined by Phase 2 Expertise, we did not find strong evidence for path independence on any measures. However, when the paths were based on Phase 4 Expertise, we found path independence for the Phase 5 poster scores immediately following Phase 4 as well as the Phase 7, 3-year written posttest.

Recall that only part of the overall poster score was based on issues of experimental design and avoidance of confounds. Many other aspects of "good science" were included in that score, including such things as the overall adequacy of the research design, measurement issues, appropriate sample

size, legitimacy of inferences, and the extent to which the conclusion was supported by the data (see Appendix A). Indeed, that is the reason why several of our analyses break down the Grand poster score into CVS-only and Non-CVS components. Why does mastering the relatively narrow and specific concepts and procedures associated with the design of simple unconfounded ramp experiments lead to high performance on the more inclusive poster evaluation task? We believe there are several possible reasons for this particular type of far transfer from CVS Expertise to high quality poster evaluations. First, CVS requires decomposition of the overall situation, requires attention to detail, and encourages a focused search for causal paths – all three of these skills would be expected to result in more thorough critiques. Second, reasoning about causal and non-causal factors in the simple ramps domain (in either training condition) might foster and enrich a "rhetorical stance" that is fundamental to science and captures aspects of "science as argument" (Kuhn, 1993). These perspective-taking opportunities might also improve children's encoding and critiques of posters shown to them. Third, CVS reasoning conveys some essential implicit aspects of the "Nature of Science" (Aicken, 1991; Bianchini & Colburn, 2000; National Academy of Science, 1998) which might stimulate children to think beyond the confines of CVS to broader aspects of good experimental science. Similar arguments could be applied to why children's CVS Expertise transferred to the written posttest administered a full 3 years later. In that case, the arguments could be augmented, perhaps, with the idea that children who learned CVS in our study were "prepared to learn" in the following years of science education.

We sought to examine the relation between different types of instruction, learning, transfer, and assessment. This exploration might seem a straightforward undertaking, particularly for such a relatively focused topic. One compares two or more types of instruction, measures their immediate impact on students' learning, and then assesses retention and transfer at some later point. To the extent that the assessments reveal a main effect of instructional type on student performance, that type of instruction is to be preferred. Of course, the greater the number and variety of measures that show this advantage, the better, and the more students for whom the effect is demonstrated, the better. However, as evidenced by the results of the current study, the outcomes of instructional assessments are rarely straightforward, and further examination of interactions among learner characteristics, type of instruction, and transfer features are usually necessary before instruction can be further improved.

For example, one limitation of the present study is that the sample size did not allow us to fully investigate developmental differences in instructional effectiveness and path independence. Another is that the children in this study were self-selected from a single private elementary school. Similar studies with less "privileged" populations have revealed much lower overall performance, both before and after instruction (Dean & Kuhn, 2007; Li, Klahr, & Siler, 2006), and might also yield different results with respect to path independence.

We hope to further explore this rich data set by examining available responses students gave for why they set up experiments the way they did during pre- and posttests and, for students in the Explicit Instruction condition, their evaluations of the experiments presented to them. In particular, we are interested in their statements regarding CVS procedural and conceptual knowledge, misconceptions about the goal of the experiment and/or the necessity of knowing something about the ramps domain, and advanced thinking (e.g., exhibiting understanding that interactions, variability, and error are important aspects of experimentation). Additional replication studies could examine similar questions with larger, more diverse samples. More useful, however, would be to (1) adapt the instruction to reach "all" students, especially those with deeply entrenched misconceptions or alternate goals; (2) provide more training and practice in different domains so that students understand the domain independence of the CVS procedure; and (3) adapt the instruction for use by regular classroom teachers in full-class environments. A recent study by Zohar and David (2008) has addressed both the lab to classroom transition and the effectiveness of Explicit Instruction on CVS for both high- and low-achieving children and found substantial immediate and delayed impact of direct instruction for the low-achieving group. In our own work thus far, we have completed only one such "lab-script to lesson plan" project (Toth et al., 2000), but we are pursuing this general idea by creating a computer tutor – to be used in a wide range of educational settings – that will provide effective instruction on the basics of experimental design (Strand-Cary, Klahr, Siler, Magaro, & Li, 2007).

Acknowledgements

This work was supported, in part, by an APA/IES Postdoctoral Education Research Training grant to Strand-Cary (R305U030004), an IES grant (R305H060034) to Strand-Cary and Klahr, and an NSF grant 0132315 to Klahr.

Thanks go to Audrey Russo and Andrew Young for assistance at various stages of execution and data analysis and to two anonymous reviewers for their comments and suggestions. Finally, we thank the teachers, students, and their parents at Sacred Heart Elementary School for their enthusiastic cooperation and participation.

Appendix A. Poster evaluation coding categories

Adequacy of research design
Communicating study and results (NOT INCLUDED IN ANALYSES)
Controlling for confounding variables/other causes (CVS issues)
Subjects of study, pre-existing/emergent/potential differences
Variation in same treatments; poster-maker bias/motivation
Design confounds/confounds for all participants/time of day confounds
Measurement
Validity: Does it measure what it's supposed to measure?
Reliability: Consistency (same sample)
Error: Experimenter error; Cheerleader error; Equipment error
Data transformation
Inferences
Sample size/population: Generalizations require large, representative sample
Variability OR Effect size: Issues involving "spread," average masking information, differences between groups is relatively large
Completeness of conclusion
Supported by data/generalization/Need more specificity to provide accuracy
Relate back to hypothesis

References

Adelson, R. (2004). Instruction versus exploration in science learning. *Monitor on Psychology*, *35*, 34–36.
Adolph, K. E., & Berger, S. A. (2006). Motor development. In W. Damon & R. Lerner (Series Eds.) & D. Kuhn & R. S. Siegler (Vol. Eds.), *Handbook of child psychology: Vol 2: Cognition, perception, and language* (6th ed.) (pp. 161–213). New York: Wiley.
Aicken, F. (1991). *The nature of science* (2nd ed.). Portsmouth, NH: Heinemann Educational Books.
Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*, 612–637.
Begley, S. (2004, December 10). The best ways to make schoolchildren learn? We just don't know. The Wall Street Journal Online (B1). Retrieved December 10, 2004 from http://online.wsj.com/article/0,SB110263537231796249,00.html.
Bianchini, J. A., & Colburn, A. (2000). Teaching the nature of science through inquiry to prospective elementary teachers: A tale of two researchers. *Journal of Research in Science Teaching*, *37*, 177–209.
Bullock, M., & Ziegler, A. (1999). Scientific reasoning: Development and individual differences. In F. E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich Longitudinal Study* (pp. 38–54). Munich: Max Plank Institute for Psychological Research.
Cavanagh, S. (2004, November 10). NCLB could alter science teaching. *Education Week*, *24*(11), 1: 12–13.
Chen, Z., & Klahr, D. (1999). All other things being equal: Children's acquisition of the Control of Variables Strategy. *Child Development*, *70*, 1098–1120.
Chen, Z., & Klahr, D. (2008). Remote transfer of scientific reasoning and problem-solving strategies in children. In R. V. Kail (Ed.), *Advances in child development and behavior* (pp. 419–470). Amsterdam: Elsevier.
Crane, E. (2005). The science storm. *District Administration*, #3 (March).
Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, *91*, 384–397.
EDC (2006). The Inquiry Synthesis Project, Center for Science Education, Education Development Center, Inc. Technical report 2: Conceptualizing Inquiry Science Instruction. Retrieved January 2, 2007, from http://cse.edc.org/work/research/inquirysynth/technicalreport2.pdf.
Ericsson, K. A., & Charness, N. (1997). Cognitive and developmental factors in expert performance. In P. J. Feltovich, K. M. Ford, & R. R. Hoffman (Eds.), *Expertise in context: Human and machine* (pp. 3–41). Cambridge, MA: MIT Press.
Geary, D. C. (2007). Educating the evolved mind: Conceptual foundations for an evolutionary educational psychology. In J. S. Carlson & J. R. Levin (Eds.), *Educating the evolved mind: Conceptual foundations for an evolutionary educational psychology* (pp. 1–99). Charlotte, NC: Information Age.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1–31.
Gopnik, A., Meltzoff, A. N., & Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains and how children learn*. New York: Hartper Collins.
Hake, R.R. (2005). Will the No Child Left Behind Act Promote Direct Instruction of Science? *Bulletin of the American Physical Society, 50*(1): 851. APS March Meeting, Los Angles, CA. 21–25 March; retrieved online at <http://www.physics.indiana.edu/~hake/WillNCLBPromoteDSI-3.pdf>.
Hmelo-Silver, C., Holton, D. L., & Kolodner, J. (2000). Designing to learn about complex systems. *The Journal of the Learning Sciences*, *9*(3), 247–298.
Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, *42*(2), 99–107.
Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86.
Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*, 661–667.
Klahr, D., Chen, Z., & Toth, E. (2001). Cognitive development and science education: Ships passing in the night or beacons of mutual illumination? In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: 25 years of progress* (pp. 75–120). Mahwah, NJ: Erlbaum.
Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, *125*(5), 524–543.
Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical vs. virtual materials in an engineering design project by middle school children. *Journal of Research in Science Teaching*, *44*, 183–203.
Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education*, *77*, 319–337.
Kuhn, D. (2007). Is direct instruction an answer to the right question? *Educational Psychologist*, *42*(2), 109–113.
Kuhn, D., & Franklin, S. (2006). The second decade: What develops (and how)? In D. Kuhn & R. Siegler (Eds.), *Handbook of Child Psychology, Vol. 2: Cognitive perception and language* (6th ed., Vol. 36, pp. 953–993). Hoboken, NJ: Wiley.
Kuhn, D. Garcia-Mila, M., Zohar, A., & Anderson, C. (1995). Strategies of knowledge acquisition, *Society for Research in Child Development Monographs*, 60(4, Serial No. 245).
Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. *Advances in Child Development and Behavior*, *17*, 1–44.
Li, J., Klahr, D., & Siler, S. (2006). What lies beneath the science achievement gap? The challenges of aligning science instruction with standards and tests. *Science Educator*, *15*, 1–12.
National Academy of Science. (1998). *Teaching about evolution and the nature of science*. Washington, DC: National Academy Press [http://books.nap.edu/books/0309063647/html/1.html].
Ruby, A. (2001). *Hands-on science and student achievement*. Santa Monica, CA: RAND.
Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*(1), 102–119.
Schmidt, H. G., Loyens, S. M. M., van Gog, T., & Paas, F. (2007). Problem-based learning is compatible with human cognitive architecture: Commentary on Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, *42*(2), 91–97.
Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184.
Schulz, L. E., & Sommerville, J. (2006). Causal determinism and preschoolers' causal inferences. *Child Development*, *77*(2), 427–442.
Steffe, L., & Gale, J. (Eds.). (1995). *Constructivism in education*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*, 753–766.
Strand-Cary, M., Klahr, D., Siler, S., Magaro, C., & Li, J. (2007). Training in experimental design (TED): Developing scalable and adaptive computer-based science instruction. In *Paper presented at the Annual Meeting of the Cognitive Science Society in Nashville*. TN.
Toth, E. E., Klahr, D., & Chen, Z. (2000). Bridging research and practice: A cognitively-based classroom intervention for teaching experimentation skills to elementary school children. *Cognition & Instruction*, *18*(4), 423–459.
Tweed, A. (2004), December 15). Direct instruction: Is it the most effective science teaching strategy? *NSTA WebNews Digest*. Retrieved on January 3, 2005 from http://www.nsta.org/main/ news/stories/education story.php?news story ID=50045.
Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*, 172–223.
Zohar, A., & David, A. B. (2008). Explicit teaching of meta-strategic knowledge in authentic classroom situations. *Metacognition and Learning*, *3*, 59–82.