*Journal of Science Education and Technology, Vol. 14, No. 2, June 2005 (© 2005)*
DOI: 10.1007/s10956-005-4423-5

# Cognitive Research and Elementary Science Instruction: From the Laboratory, to the Classroom, and Back

**David Klahr[1,2] and Junlei Li[1]**

Can cognitive research generate usable knowledge for elementary science instruction? Can issues raised by classroom practice drive the agenda of laboratory cognitive research? Answering yes to both questions, we advocate building a reciprocal interface between basic and applied research. We discuss five studies of the teaching, learning, and transfer of the "Control of Variables Strategy" in elementary school science. Beginning with investigations motivated by basic theoretical questions, we situate subsequent inquiries within authentic educational debates—contrasting hands-on manipulation of physical and virtual materials, evaluating direct instruction and discovery learning, replicating training methods in classroom, and narrowing science achievement gaps. We urge research programs to integrate basic research in "pure" laboratories with field work in "messy" classrooms. Finally, we suggest that those engaged in discussions about implications and applications of educational research focus on clearly defined instructional methods and procedures, rather than vague labels and outmoded "-isms."

**KEY WORDS:** science instruction; direct instruction; discovery learning; hands-on science; achievement gap.

How can basic research in psychology contribute to early science instruction? Conversely, how can the challenges of classroom teaching reveal areas requiring further basic research? As social, political, and economic forces demand more "evidence-based" methods to improve science education, these questions are profoundly important and relevant to both research and practice. This paper presents a summary of our efforts to address these questions in a small corner of the rich and complex universe of science instruction. We focus on our own experience in traversing the interface between basic and applied research. While this self-examination is necessarily limited in the breadth of coverage, we hope its focus enables us to provide a clear and detailed account of the issues at stake, how we dealt with them, and what remains to be done.

[1]Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.
[2]To whom correspondence should be addressed; e-mail: klahr@cmu.edu

We use Stokes' (1997) conceptualization of three different forms of research (Table I) to describe our various projects over the past several years. Stokes described two distinct factors that motivate and characterize scientific research. One factor is the extent to which the research goal is to advance understanding of fundamental phenomena, and the other is the extent to which the research outcome has any immediate and obvious utility. Stokes convincingly argues *against* two aspects of the "purist" or "classic" view of the relation between basic and applied research: (1) the temporal flow (as well as the status system in the scientific establishment) always puts basic research first (i.e., in Bohr's quadrant in Table I); (2) the two goals—fundamental understanding and considerations of use—are somehow incompatible. Louis Pasteur serves as Stokes' counterexample to both of these outmoded views. Pasteur advanced—indeed, created—the science of microbiology while working on practical problems with immediate application. Although the work we will describe here falls far short of the monumental

**Table I.** Types of Scientific Research (from Stokes, 1997)

| Quest for fundamental understanding? | Consideration of use? No | Consideration of use? Yes |
|---|---|---|
| Yes | Pure basic research **(Bohr's quadrant)** | Use-inspired basic research **(Pasteur's quadrant)** |
| No | Not applicable. | Pure applied research **(Edison's quadrant)** |

Inside the Bohr's quadrant cell:

| | Hyp | Exp | Eval |
|---|---|---|---|
| Domain Specific | | | |
| Domain General | | | |

*Note*. We have inserted a table inside one of the cells of this Table. It is expanded in Table II.

contributions produced by the occupants of Stokes' quadrants, we use his framework to organize this paper, and to situate our series of projects.

## BOHR'S QUADRANT: HOW AND WHY DOES SCIENTIFIC REASONING DEVELOP?

Pure basic research proceeds by investigating focal phenomena in their "essence," while minimizing the potential influence of non-focal variables. Thus, much of the basic research on the psychology of scientific reasoning can be classified according to a simplified model of the overall scientific reasoning process (embedded in Bohr's quadrant in Table I and shown in expanded form in Table II). This framework distinguishes between two broad types of knowledge: domain general and domain specific.

**Table II.** A Framework for Classifying Psychological Research on Scientific Thinking

| Type of knowledge | Phase of the discovery process: Hypothesis generation | Experimental design | Evidence evaluation |
|---|---|---|---|
| Domain specific | A | B | C |
| Domain general | D | E | F |

*Note*. This framework was first introduced with a few examples by Klahr and Carver (1995), and has since been used in several of our own papers (Klahr, 2000; Klahr and Dunbar, 1989) as well as in a major review on the development of scientific reasoning processes (Zimmerman, 2000).

It also differentiates among three main phases of scientific investigation: hypothesis generation, experimentation, and evidence evaluation. The individual cells capture the ways in which psychologists typically carve up the real-world complexity of the scientific discovery process into manageable and researchable entities.

Psychological investigations that fit in Cell A—domain-specific hypothesis generation—are exemplified by McCloskey's (1983) pioneering investigation of people's naive theories of motion. McCloskey presented participants with depictions of simple physical situations (e.g., an object being dropped from an airplane, or a ball exiting a curved tube) and asked them to predict the subsequent motion of the object. He found that many college students held naive impetus theories (e.g., the belief that the curved tube imparted a curvilinear impetus to the ball, which dissipated slowly and made the ball continue in a curved trajectory). In this kind of study, participants are asked to describe their hypotheses about a specific domain. They are not allowed to run experiments, and they are not presented with any evidence to evaluate. Nor is there an attempt to assess any domain-general skills, such as designing unconfounded experiments, making valid inferences, and so on. Thus we classify McCloskey's work, as well as studies about children's understanding of heat and temperature (Carey, 1985) or evolution (Samarapungavan and Wiers, 1997), as examples of basic research in Cell A of Table II.

Studies in Cells C and F, domain-specific and domain-general evidence evaluation, focus on people's ability to match hypotheses with evidence. Typically, participants are shown tables of covariation data and asked to decide which of several hypotheses is supported or refuted by the data in the tables. In some studies, the hypotheses are about abstract and arbitrary domains (e.g., Shaklee and Paszek, 1985). Such investigations can be classified as domain general (Cell F). In other cases, the hypotheses refer to real-world domains, such as plant growth under different conditions of light and water (Amsel and Brock, 1996; Bullock *et al.*, 1992). Within these real-world domains, participants have to coordinate their prior domain knowledge with the covariation data in the tables (e.g., Ruffman *et al.*, 1993). These studies involve both domain general and domain specific knowledge (cells C *and* F).

Rather than continue to tour the cells in Table II, we refer the reader to Zimmerman's (2000) extensive review article using the same framework. The main point here is that most of the rigorous research on the development of scientific reasoning processes can be classified into one or two cells of Table II. Researchers who investigate children's causal judgment from covariation in data evidence do not require children to design the experiments and collect the data themselves. Researchers who probe children's misconceptions about the theories of motion do not present children with controlled experiments demonstrating laws of motion. The choice to limit the scope of investigation is deliberate in basic research design and necessary for researchers to focus on any particular aspect of scientific reasoning. There are a few basic research studies where researchers do attempt to integrate multiple cells in their investigations (e.g., Klahr *et al.*, 1993; Schauble, 1996; Schauble *et al.*, 1991). These studies describe the interdependence among the various types of knowledge and reasoning strategies across phases of discovery and types of knowledge. In relation to the whole body of basic research, the latter studies are a minority.

This kind of "divide and conquer" research has revealed several important features about the emergence of scientific thinking in children. First, children do develop theories, however naïve, about the natural world long before they enter school. These conceptions or misconceptions become deeply entrenched in children's mental models in various domains, including motion, force, heat, temperature, mass, density, planetary movements, animacy, and the theory of mind. Upon entering school, children organize their new learning around these prior beliefs. When contradictions between prior beliefs and new learning arise, children have difficulty integrating new knowledge with prior beliefs. Second, children are capable of thinking abstractly and symbolically in some domains, though they have difficulty in flexibly transferring such partial competence across domains. Their performance reveals varying levels of both potential and limitation in areas such as theory formation, experimental design, evidence evaluation, data interpretation, and analogical mapping. In certain areas of reasoning, children's partial competence can be improved with experience or instruction (e.g., Case, 1974; Chen and Klahr, 1999; Klahr and Chen, 2003; Kuhn and Angelev, 1976; Kuhn *et al.*, 1988; Schauble *et al.*, 1995). We refer interested readers to more comprehensive reviews of the related literature (e.g., Klahr and Simon, 1999; Metz, 1995; Zimmerman, 2000).

## EDISON'S QUADRANT: WHAT WORKS IN TEACHING SCIENCE?

Using the narrowly focused research described above laboratory scientists of psychology accumulate and value scientific knowledge in Bohr's quadrant. However, to a classroom teacher working in Edison's quadrant and trying to figure out "what works" in the classroom, the utility of such accumulated knowledge is not clear. The research and policy communities often lament that teaching practices, textbooks, and curricula are not informed by the basic research knowledge base. To that, educators respond that "research-based" findings in Bohr's quadrant are too de-contextualized to provide useful guidance in everyday practices (Hiebert *et al.*, 2002; Lagemann, 1996, 2000; Strauss, 1998).

The task of translating basic research findings into practice sometimes falls on the shoulders of applied researchers. In contrast to the basic researchers' search for answers to "why" questions, applied researchers more often take an engineering approach and ask instead, "how?" While such an engineering approach can be more relevant to the outcome goals of education, it often lacks the rigorous control and clear specification of operators and mechanisms typified by the scientific approach. For example, during the past decade, there has been great interest and investment in instructional innovations involving online technology, collaborative

learning, and the marriage of the two, online collaboration. Yet in most studies of such products (e.g., Wasson *et al.*, 2003), intervening variables tend not be controlled and often are not specified at all. In many cases, the goal is simply to demonstrate that the product works as a full package, including features of interface design, collaboration, and technology. These end products, when successful, are more readily applicable to their specific contexts than the isolated findings of basic research studies. Yet, even when successful, few such product development efforts allow for the extraction of testable design principles that would support replication outside the confines of the particular product and the knowledge domain it addresses.

Worse yet, such an engineering approach often results in "the tail wagging the dog," whereby the technological feature (tail) determines the selection of topic and design of instruction (dog). For example, in the rush to reap the benefits of "distance learning," the mere act of putting something online became the hypothesized variable that promises great return. This led to dubious implementations such as literally transferring lectures notes verbatim onto web pages or replacing live discussion with online chat in hopes that the courses would somehow become more effective (e.g., review by Bernard *et al.*, 2004). This is not to say that applied research needs to rigidly adopt the often narrow focus of basic research. One can hardly reinvent an entire curriculum and still condense the changes to a handful of clearly specifiable and testable variables. In many ways, the engineering approach is necessary for any large-scale instructional design effort. It is simply more difficult to attribute efficacy of a whole product to particular design features or principles.

A teacher's task is more like that of an engineer than a scientist. The available components are products like curricula, textbooks, teaching guides, and training workshops, instead of academic journals in which basic research findings are published. Like that of applied researchers, the teachers' primary goal is to "make it work." Seeking "why it works" usually requires additional time and energy that overburdened teachers can ill-afford. Additionally, unlike research scientists conducting experiments, teachers take pride in not executing the same lesson plan exactly the same way every time. Instead, they aim to be able to adapt and adjust based on the students' interests, or even the particular classroom dynamics of the day. Being able to go with the flow and capitalize on students' momentary shifts of interest and energy is a valued teaching skill. The engineering approach becomes part of what defines a teacher's expertise and professionalism.

Consider a simple science inquiry lesson exemplified by the following 5th-grade science unit included in the National Science Education Standards (National Research Council [NRC], 1996, pp. 146–147):

> Ms. D. wants to focus on inquiry. She wants students to develop an understanding of variables in inquiry and how and why to change one variable at a time. . . . Ms. D . . . creates an activity that awakens students' interest and encourages them to ask questions and seek answers. Ms. D. encourages students to look for applications for the science knowledge beyond the classroom. . . . Ms. D helps them understand that there are different ways to keep records of events. The activity requires mathematical knowledge and skills.

Although the goal of this unit (how and why to change one variable at a time) fit nicely into Table II's cell E (domain general experimental design), this lesson unit on pendulum motion covers every one of the cells (summarized in Table III). Children's activities included hypothesis generation in a specific domain, evidence evaluation via record keeping and data analysis. Combing through the full description of the unit, domain specific content knowledge and domain general process skills are not separable (NRC, 1996). In fact, the only thing missing from the described class activities is a focused discussion about "how and why to change one variable at a time," which was the stated goal of the unit. Instead, the intent of the lesson is perhaps to embed such abstract domain general skill within the rich context of an actual inquiry activity and hope that students

**Table III.** Fitting a Science Class onto the Research Framework

| | Hypothesis generation | Experimental design | Evidence evaluation |
|---|---|---|---|
| Domain specific | What factors determine pendulum period | Construct apparatus and procedure | Data analysis, specific measurement skills (use of stop watch) |
| Domain general | Hypothesis and prediction | Control of variable | Effect size, error and variability, size, graphing |

would acquire such skill within context and under the teacher's dynamic and customized guidance.

This example reveals the inherent conflict between the goals and constraints of basic research and the demands of practice. In order to have confidence in their conclusions, basic researchers need to isolate theoretically motivated variables and investigate them in highly constrained contexts. But a teacher's assigned task is to weave various process skills and content knowledge seamlessly into classroom lessons. The lines separating individual cells in Table II's framework offer little meaning for the teachers' lesson planning as demanded by the standards reform. In the pendulum class example, basic research could inform Ms. D. in piece-meal fashion how her students may be generally struggling with causal judgments, covariation, response to anomalous data, and experimental design. For teaching tasks like those of Ms. D., the practical helpfulness of such piece-meal insights is limited by their weaknesses in explicating the complex interdependencies among multiple reasoning processes and domain-specific knowledge. Thus the dilemma arises: fundamental understanding gained by basic research is robust, but less relevant to real-world practice because of the very constraints that make the research robust.

We have so far highlighted the mismatches between the scientific approach of basic researchers and the engineering approach of applied researchers and teachers. These disconnections partly contribute to educational institutions' tendency to swing like a pendulum amidst the cyclical reform recommendations (sometimes based on basic research)—adopting style and substance alike, mixing theory with hunch, and implementing theoretically-motivated interventions with a myriad of pragmatic alterations. Lacking clear operational and theoretical specification, many reforms come and go cyclically without ever being implemented faithfully or delivering sustainable improvements in education (Brown, 1992).

## PASTEUR'S QUADRANT: INTEGRATING USE AND UNDERSTANDING

Although we acknowledge these fundamental distinctions between pure basic and pure applied research, we argue that they can be reconciled to the mutual benefit of both approaches. Pure basic research, with all of its rigorous drive for operational specification of variables and operators, can originate its questions in the context of practical use. Likewise, it is plausible for the findings from such use-inspired basic research to directly inform the practice of education and generate insights and questions to further propel the quest for fundamental understanding. With this larger goal as the context, we describe a series of studies to illustrate a reciprocal process through which a use-inspired research program fostered a continuing ebb and flow of insights between laboratory research and classroom application. While we believe that the story we are about to tell is of some general interest and utility, we are well aware of its limitations and idiosyncrasies. We present it in the spirit of stimulating others to take similar journeys through Stokes' quadrants. In this abbreviated review of studies, we selected only study details relevant to the purpose of the present paper. We refer interested readers to the original publications for more complete descriptions.

The central topic of all of the work to be described is the Control of Variables Strategy (CVS). CVS is the method of designing unconfounded experiments by changing only one variable at a time. Competency in CVS is reflected in children's ability to generate unconfounded experiments, identify and correct confounded experiments, and make appropriate inferences from experimental outcomes. It is an essential part of the "experimental design" phase of scientific discovery in both school science and authentic scientific research. It received explicit mention in both the Standards (National Research Council [NRC], 1996) and the Benchmarks (American Association for the Advancement of Science [AAAS], 1993).

In all of our studies, we used materials in which several factors, each having only two "levels," can be varied to determine how those levels affected the experimental outcomes. One of the apparatus we used in these studies is "balls down the ramp" (Fig. 1). There are four variables to be contrasted or controlled: the surface texture (smooth or rough), the length of the run (long or short), the steepness of the ramp (steep or shallow), and the type of ball (rubber ball or golf ball). For each experiment, the participant is asked to investigate a specific question: such as, whether the surface texture makes a difference in how far a ball will roll. Ideally, the participant should conduct an unconfounded comparison by contrasting the rough surface with the smooth surface, while holding all other variables constant between the two ramps. If the participant fails to hold the remaining variables constant between the two ramps, the comparison becomes confounded.

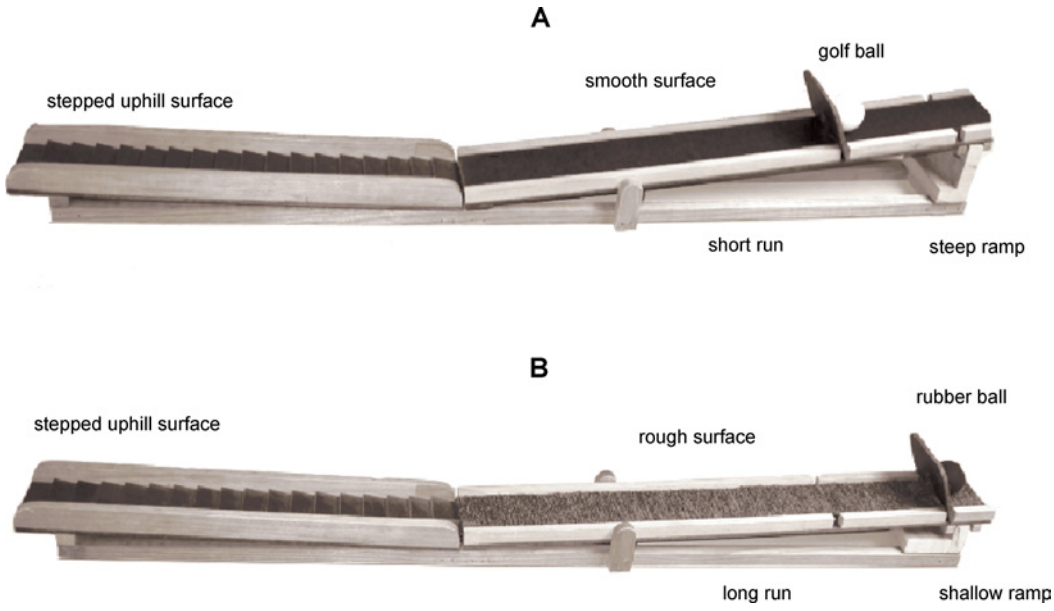


**Fig. 1.** Balls down the ramp apparatus. Illustration and photo reprinted with permission from Klahr and Nigam (2004), p. 663. Original caption reads, "The confounded experiment depicted here contrasts (a) a golf ball on a steep, smooth, short ramp with (b) a rubber ball on a shallow, rough, long ramp."

The extreme version of a confounded comparison would be to have every single variable contrasted between the two ramps (such as the experiment depicted in Fig. 1). In some studies, we used other materials having different physical properties but the same underlying logical structure (e.g., springs and weights, pendulums, sinking objects).

We present a research program that explores the teaching, learning, and transfer of CVS with elementary school children. To bridge research and practice, we worked in both the psychology laboratory and the classrooms of elementary schools. We identified research questions from pressing educational issues, translated laboratory procedures into classroom scripts, and incorporated standardized tests into our assessment instruments. In this account, we will discuss the relevance of our findings to a few pressing educational issues and we will highlight the even greater and more challenging questions spurred by the integration of cognitive research and instructional practice.

## Study 1: The Training and Analogical Transfer of CVS (Chen and Klahr, 1999)

This investigation is motivated by a basic theoretical question in Bohr's quadrant (see Table I) and Cell E (see Table II) of basic research on scientific reasoning: Are abstract reasoning skills "situated" in the domain of their acquisition and generally not usable across domains? Or, can the skills acquired in one domain be applied in another domain via the basic process of analogical transfer? Within the context of our research, we translated this enduring question about "transfer distance" into a specific query: Can CVS skills acquired in one domain transfer to other domains which bear deep structural resemblance but lack surface similarity?

To answer this question, we modeled three types of learning environments after our interpretations of common methods of instruction: *learning via discovery* (e.g., Jacoby, 1978; McDaniel and Schlager, 1990), *learning via probing*, and *learning via explicit instruction* (e.g., Klahr and Carver, 1988). In specifying these methods, we sought to maximally isolate and contrast the variables unique to each form of instruction. In the learning via discovery condition, we gave the participants opportunities to conduct hands-on experiments to answer domain specific questions. In the learning via probing condition, participants not only conducted such hands-on experiments but were also asked what they had learned from their experiments and if they "knew for sure" that a variable was causal. In the learning via explicit instruction condition, in addition to the hands-on experiments and probes, we provided

**Table IV.** Summary of Training Condition in Chen and Klahr (1999) Instructional Conditions

| | Hands-on experimentation | Probes | Explicit instruction |
|---|---|---|---|
| Learning via discovery | Yes | No | No |
| Learning via probing | Yes | Yes | No |
| Learning via explicit instruction | Yes | Yes | Yes |

participants with explicit instruction on what CVS is and how to design CVS experiments in-between their experimentations. Table IV provides a summary of these three conditions.

Eighty-seven students from second, third, and fourth grades in two private elementary schools participated. Children in each grade were randomly assigned to one of the three training conditions. Three physical apparatus were used: the balls and ramp described earlier (see Fig. 1), springs and weights, and sinking objects. All three sets of materials varied greatly both in physical appearances and in the basic physics of the domain, but shared deep structural features (e.g., each apparatus had four distinct and bi-level variables).

The measures of CVS competency were defined as follows (Chen and Klahr, 1999, p. 1099):

> *Very Near Transfer* is defined as the application of CVS to test a new aspect of the same materials used in the original learning problem. *Near Transfer* is defined as the use of CVS to solve problems using a set of different materials that are still in the same general domain as the original problem. *Remote Transfer* refers to the application of CVS to solve problems with domains, formats, and context different from the original training task after a long delay.

Participants were pretested prior to any training or probing. After they explored the first physical apparatus, participants received probing and/or training in their respective conditions. Participants' CVS skills were then assessed using the same physical materials (*very near transfer*), but different focal variables (e.g., for the ramp, if they explored surface and slope during training, they then investigated type of ball and length of run). Seven days following the training, participants were asked to investigate the effects of variables using two new sets of physical materials for the *near transfer* task (e.g., participants who had initially worked with the ramps now worked with springs or sinking objects). Participants' domain specific knowledge of the physical apparatus was also assessed. Seven *months* after the training, participants were asked to complete a paper and pencil evaluation of good and bad experiments across a variety of domains (e.g., baking cake, designing airplane), all different from any of the physical apparatus they explored earlier. We termed this task the *remote transfer* due to its domain changes, context difference (hands-on test to paper-and-pencil test), and the extended time delay. We refer interested readers to a discussion of the complexity of classifying transfer "distances" in a recent literature review (Barnett and Ceci, 2002).

Though participants were not initially competent at the use of CVS, they did learn with training. The robustness of children's CVS learning, measured by their performance on transfer tasks, varied significantly by training condition (see Figure 2). Aggregating across grade-levels, children significantly improved their CVS competency with one particular method of training—learning via explicit instruction. The learning was robust when measured at all three transfer distances. In addition, only in the learning via explicit instruction condition did the improved use of CVS significantly increase the children's ability to evaluate domain-specific evidence (see Cell C, Table II) and demonstrate a better understanding of the physical materials.

While we believe that these findings have practical implications (and we consequently expanded the research program to investigate these implications), the study itself is a typical example of basic scientific research in Bohr's quadrant. Its relevance and practical usability to a classroom teacher's integrated science teaching practice (exemplified by Table III) is limited by the study's narrow focus (primarily on Cell E of Table II, with some marginal exploration of Cell C) and contrived *psychology laboratory* setting (requiring customized physical apparatus and one-on-one instruction by trained researchers). Are such findings replicable in real classrooms? How do they inform current debates in science education (e.g., discovery learning vs. direct instruction, physical manipulation vs. computer simulation)? Motivated by these more pragmatic questions, we aimed subsequent research efforts towards Edison's and Pasteur's quadrants of research.

## Study 2: Classroom Replication of CVS Instruction (Toth *et al.*, 2000)

Can the training method for CVS that proved to be effective in the psychology laboratory be used in real classrooms? Moving from Bohr's quadrant of basic research to Edison's quadrant of applied research,

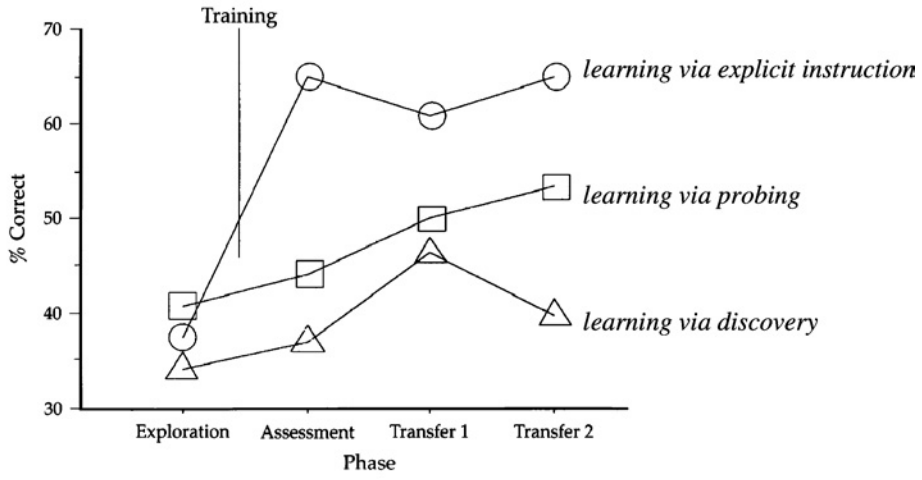


**Fig. 2.** Percentage of trials (by all participants) with correct use of CVS by training condition. Based on Chen and Klahr (1999), Figure 3, p. 1109. For children in the learning via explicit instruction condition, training was provided to them during the "exploration" phase. The "assessment" corresponds to very near transfer, or, within the same physical domain as the one used in the "exploration" phase. Both "Transfer 1" and "Transfer 2" correspond to near transfer, when participants applied CVS to investigate different physical domains from the one they used in the "exploration" phase.

we began to address this question by adapting the training procedures for learning via explicit instruction from the psychology laboratory to the elementary school science classroom. The instructional goal remained focused on CVS (see Cell E, Table II). The procedure, however, required many adjustments (Klahr *et al.*, 2001).

Classroom teachers replaced trained researchers as the providers of probing and explicit instruction. The trainer to student ratio changed from 1:1 to 1:20. The limited quantity of physical experimental apparatus (a common problem in most elementary science classrooms) made collaborative learning necessary in addition to individual processing. The participants themselves, instead of the one-on-one experimenter, kept records of each experiment and its outcome. With all these changes simultaneously put in place, Study 1's controlled comparison of three training methods—varying one thing at a time and using randomized assignment—is necessarily replaced by a much simplified "pre/post" design to replicate "what works." The latter approach of replicating one single training condition along with all of the necessary classroom adaptations can be described as "holding one thing (the training script) constant at a time," more typical of the engineering approach within Edison's quadrant.

Seventy-seven fourth graders from four classrooms, taught by two different science teachers in two different private elementary schools, participated in Study 2. The adapted-for-classroom learning via explicit instruction training condition was effective. Both the direction and the magnitude of change from pretest to various posttests were comparable to those of Study 1. The absolute change and achieved level was not only statistically significant, but also educationally meaningful: The percent of children demonstrating consistent use of CVS (correct use in at least eight out of nine trials) rose from 5% on pretest to 95% on posttest. To ensure that participants were not simply learning to execute a procedure by rote, a second measure of "robust CVS use" was used to identify children who could both physically use and *verbally justify* CVS in at least eight out of nine trials. The corresponding percentages rose from 0% to 55%. These results were evidence of successful replication and classroom implementation of the CVS training using Study 1's learning via explicit instruction method.

Two basic research questions emerged during the classroom study that required new investigations back in the psychology laboratory in Bohr's quadrant. Although the main focus of Study 2 was domain-general experimental design (see Cell E, Table II), the learning of CVS in the science classroom interacted with participants' knowledge and skill in Cell F—domain-general evidence evaluation. During our classroom study, we found that

many elementary school students were unable to develop adequate notation systems to record their experimental design and outcomes. In addition, the students had some difficulty discriminating experimental effects from random error while analyzing the results of their experiments. Consequently, students would draw incorrect inferences despite having conducted logically correct experimental comparisons. Although these Cell F questions do not directly affect laboratory investigations of CVS training, understanding the interdependency between evidence evaluation and experimental design is no less important to the science teacher than knowing how to teach CVS. However, cognitive and developmental researchers have not studied these issues extensively. Heading back into Bohr's quadrant, researchers within our immediate group began exploring children's understanding of data variability (Masnick and Klahr, 2003; Masnick *et al.*, in press; Masnick and Morris, 2002) and children's use and learning of notation (Triona and Klahr, 2002). Although these studies are outside the scope of this paper, we mention them here to illustrate how issues raised during applied research in Edison's quadrant can stimulate basic research efforts in Bohr's quadrant.

In addition to raising new basic research questions, Study 1 and Study 2 also generated several practical questions. The successful replication of the learning via explicit instruction training method shifted the research focus from comparative analysis of the relative efficacy of training methods to more in-depth analysis about why a particular training method was effective. First, to what extent is the efficacy of the learning via explicit instruction dependent upon hands-on manipulation of physical materials (i.e., an interaction between method of instruction and physical manipulation?) Hands-on manipulation of physical apparatus has been part of every training method tested in our studies thus far, effective or ineffective. We compared the manipulation of physical and non-physical materials in Study 3 (Triona and Klahr, 2003). Second, what does it mean when we say an instruction is "working?" We are aware that CVS transfer could come in more varied ways than our particular specifications of *very near transfer, near transfer*, and *remote transfer*. Elementary school science teachers are more concerned with broad applications of CVS in varied and authentic contexts (e.g., answering ill-formed questions on tests, designing science fair projects). Our measures of CVS, while theoretically valid, have not risen to the ecological validity bar of authentic tasks. Subsequently, we employed more authentic transfer tasks in both Study 4 (Nigam and Klahr, 2004) and Study 5 (Li and Klahr, manuscript in preparation). Third, we have concerns of ecological validity with regards to our participant population. Each of the studies described thus far used participants from private schools. Our findings regarding instructional effectiveness for CVS are limited to this highly selective group of participants. Are economically disadvantaged and low-achieving students equally ready to learn CVS? Would the same instructional methods prove effective with such students? These questions were addressed in Study 5.

### Study 3: Hands-On Manipulation and Computer Simulations (Triona and Klahr, 2003)

Unlike Study 1, this investigation was not primarily motivated by the need to expand basic theories, but was instead inspired by a question of "use"—how important is hands-on manipulation of physical material, compared with virtual material (computer-based simulations), in learning basic scientific skills (e.g., CVS)? Thus, we consider this study an example of use-inspired research in Pasteur's quadrant. The basic research in this area is relatively recent and scarce (Uttal *et al.*, 1999; Uttal *et al.*, 1997), though the effectiveness of hands-on manipulation in learning science has long been a hotly debated issue in applied settings (e.g., Ruby, 2001). The importance of physical manipulation in "hands-on" science was once commonly accepted by the science education community (e.g., National Science Teachers Association [NSTA], 1990, 2002). More recently, standards reforms (e.g. AAAS, 1993, NRC, 1996) cautioned against equating "hands-on" with good science instruction. For example, the first set of science education benchmarks (AAAS, 1993, p. 319) explicitly argued:

> Hands-on experience is important but does not guarantee meaningfulness. It is possible to have rooms full of students doing interesting and enjoyable hands-on work that leads nowhere conceptually. Learning also requires reflection that helps students to make sense of their activities. Hands-on activities contribute most to learning when they are part of a well-thought-out plan for how students will learn over time.

In the two studies we described so far, "hands-on" manipulation of physical materials (such as ramps, balls, and springs) *is* a "part of a well-thought-out plan." However, does it *need* to be? Theoretically, we had hypothesized (Chen and Klahr, 1999;

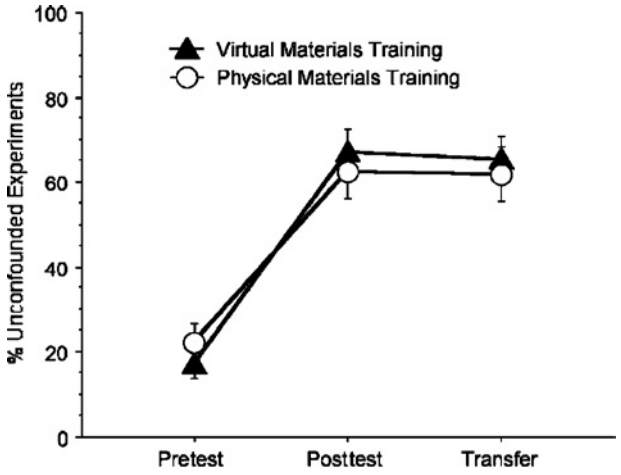


**Fig. 3.** By virtual or physical training condition, proportion of unconfounded experiments designed by children. Reprinted with permission from Triona and Klahr (2003), Figure 4, p. 162.

Klahr and Carver, 1988) that it was the "explicit" aspects of instruction that helped focus learners' attention on the CVS logic of experimentation. Thus, we proposed the hypothesis that manipulation of physical materials is not essential to the efficacy of our CVS instruction.

To test this hypothesis, we replicated the most effective form of CVS training from our prior studies (learning via explicit instruction in Study 1) using both the original springs and weights apparatus and a newly developed virtual version of the springs and weights domain implemented on a computer interface. The experimenter provided explicit and direct instruction identical to that used in Study 1. The participants learned how to design and interpret unconfounded experiments with either the physical or the virtual materials. All participants, whether trained in the physical or the virtual condition, were required to design experiments using the physical "balls down the ramp" (Figure 1) apparatus on the transfer test. We wanted to test whether learning in the virtual environment would somehow handicap the participants' ability to apply CVS using physical materials.

Ninety-two fourth and fifth graders from two private schools participated in Study 3. Across all measures of CVS at pretest, posttest, and transfer, we found no significant difference between participants trained using physical and virtual materials. Those participants trained in the virtual springs and weights condition could transfer their CVS skills to the physical ramps apparatus as well as their counterparts trained in the physical springs and weights condition (Fig. 3). The absolute magnitude of learning gains was comparable to what was achieved in prior studies.

To interpret this finding, we must first note that the virtual interface is designed with utmost care to preserve the essential experience of manipulation and experimentation in the absence of physical objects. We cannot infer from this finding that hands-on manipulation of virtual rather than physical objects will always be equally effective in science education. Our result does add confidence to the assertion that "a well-thought-out plan" of instruction and student learning is a more significant factor than the actual materials used. We suggest that the efficacy of a theoretically sound and empirically proven method of instruction would be relatively independent of the medium of delivery, so long as the materials (physical or virtual) incorporate features that align with the basic principles of instruction. In our case, such features are (1) explicit procedural instruction about the CVS procedure, (2) provision of good and bad examples of CVS accompanied by explicit conceptual explanation of CVS, and (3) combination of instruction with experimental manipulation and outcome observation.

Our contribution to the discussion of the "hands-on" question is perhaps to argue that a wrong question has been asked. In its place, we propose the following alternatives: What is essential in the design of materials for science education that encourages "minds-on" science? What are the tradeoffs in using physical and virtual materials? What design

features could effectively compensate for such trade-offs? Can a principled approach to science instruction transcend the medium through which it is delivered? These questions call for investigations across all three quadrants—basic, applied, use-inspired—of research. We added these questions in our current research agenda in both the psychology laboratory and the classrooms.

### Study 4: Authentic Transfer of CVS (Klahr and Nigam, 2004)

While the context and materials varied greatly in the three studies reviewed so far, the measures of CVS have stayed fairly consistent. One may well question whether we have underestimated the efficacy of learning via discovery in Study 1 because we lacked more authentic and challenging transfer tasks (Chinn and Malhotra, 2001). The constructivist literature has often argued that the use of knowledge constructed by the learners themselves would be more robust than that which was simply "told" to the learner, especially in challenging and complex real-world tasks (e.g., Elkind, 2001; Loveless, 1998; Piaget, 1970). The contrast between Study 1's learning via discovery and learning via explicit instruction can be theoretically framed as the contrast between self-constructed CVS knowledge and "told" CVS knowledge. Is there a more "authentic" task that would reveal the learning differences proposed by this contrast? The task we chose was the evaluation of science fair posters. Making science fair posters is a common activity in many elementary school science programs. We constructed two science fair posters based on actual posters generated by sixth-grade students. Participants were asked to make suggestions to help make the poster "good enough to enter in a state-level science fair". They were encouraged to comment on various aspects of the poster, including experimental design, data analysis, measurement, explanations and conclusions. We place Study 4 in Pasteur's quadrant as it sought understanding on a basic theoretical question within a context inspired by actual use.

We assigned 112 third and fourth graders into two training conditions, based on the scripts for learning via discovery and learning via explicit instruction in Study 1. We chose to label the two conditions *discovery learning* and *direct instruction*. In the discovery learning condition, students were provided with a goal (e.g., "see if you can set up the ramps to find out if the height of the ramp makes a difference?"). They were free to explore, in a hands-on fashion, various combinations of arrangements. They can run the experiments and observe the results. Then, under the experimenter's suggestion, they move on to the next goal. In the direct instruction condition, explicit instruction of CVS was provided by the experimenter to the students before and during the students' experimentation and observation. The two conditions were designed to contrast students who discovered CVS on their own and students who were told CVS by experimenters.

The CVS performance following the training replicated Study 1's findings, favoring direct instruction by a wide margin. Of the direct instruction participants, 77% designed at least three out of four unconfounded experiments after training, compared with just 23% in the discovery learning condition. This wide gap between conditions suggested that direct instruction was more efficient than discovery learning in the short-term, but did not warrant the conclusion that discovery learning produced no *added* long-term benefit beyond what direct instruction could deliver. To inform the latter claim, we contrasted the CVS "masters" in direct instruction (77%) and those in discovery learning (23%) to test the hypothesis that discovery learning produced more robust and flexible learning at the cost of less efficiency. Consistent with this hypothesis, we should expect that expertise reached via discovery learning to have an advantage in the more authentic and less constrained transfer task of evaluating science fair posters. We found no evidence supporting this hypothesis (see Table V). On the overall evaluation score of the posters (including all aspects of the poster), we found no statistically significant difference between the two types of CVS masters. That is, the many masters ($n = 40$) produced by direct instruction did just as well on the poster evaluation task as the few masters produced by discovery learning ($n = 12$).

**Table V.** Poster Evaluation Scores for Masters and Non-Masters Following Two Different Learning Paths (Direct Instruction or Discovery Learning)

| Learning path | *n* | Mean poster evaluation score | SE |
|---|---|---|---|
| Master (discovery) | 12 | 13.8 | 1.2 |
| Master (direct instruction) | 40 | 12.5 | 1.0 |
| Nonmaster (discovery) | 40 | 9.1 | 0.7 |
| Nonmaster (direct instruction) | 12 | 8.2 | 1.3 |

*Note*. Based on Klahr and Nigam (2004), Table 2, p. 666.

These findings prompt "a re-examination of the long-standing claim that the limitations of direct instruction, as well as the advantages of discovery methods, will manifest themselves in broad transfer to authentic contexts" (Nigam and Klahr, 2004, p. 666). It is important to note that these findings challenge the claims and implementations of a common form of discovery learning: giving students a well-specified task and the necessary physical tools, but not offering explicit instruction on the preferred strategy or corrective feedback. These findings do not equate discovery learning in this form with various forms of guided discovery. In fact, by some definitions, we could classify the learning via explicit instruction script as "guided" or "scaffolded" discovery. We shall return to the issue of "naming" later in the paper.

Far more important than comparing methods of instruction, Study 4 refreshes a fundamental discussion for both cognitive research and educational practice: To what extent and under what conditions is the type of acquired knowledge and expertise independent of the learning paths? We demonstrated that in the specific instance of CVS learning, the particular paths defined by our conditions of direct instruction and discovery learning did not result in different types of expertise, but only in the different numbers of experts produced. We do not know how generalizable our findings are across domains and across tasks. In a policy climate of seeking "what works" in educational practices and amidst heated arguments among various approaches, we cannot overemphasize the importance of revisiting this fundamental question.

## Study 5: Narrowing The Achievement Gap on CVS

Are three controlled experiments with random assignment and one successful classroom replication sufficient to qualify learning via explicit instruction as a "research-based" educational prescription (No Child Left Behind Act, 2002)? We are far from making such a claim. One consistent factor in all four studies discussed is the high economic and academic levels of our participant population. One could fault the entire series of studies by saying that our training method is helping the good students to get even better. We know nothing about how our training method might impact students in the lower economical and achievement stratum.

This raises several important questions. Can the CVS training lesson plans in Study 2 be adapted for classrooms with predominantly low-achieving and low-SES students? Are such lesson plans sufficient to close the achievement gap (within the topic of CVS) across diverse student populations? What principles can we extract from the design of CVS training to apply more broadly to elementary science instruction? Must the researchers specify lesson plans to the level of scripts for the teachers to recite verbatim, or could teachers master the essential principles of lesson planning to devise and improve their own lesson plans? In order to investigate these questions, we began a long-term fieldwork project to understand and improve science education in low-achieving urban schools. Extending beyond Study 2's limited focus on CVS training, we now search for operational definitions of the learning via explicit instruction method that is generalizable to elementary science education as a whole. Our methodological approach is typical of applied research in Edison's quadrant, holding one variable (the essential design features underlying the instructional method) constant while making several concurrent adjustments to fit into real-world classrooms. Within the context of this project, we describe an initial investigation most closely built upon the four studies thus reviewed.

Our primary goal in this study is to determine whether the method that was so effective when used with high-achieving students in our earlier studies, learning via explicit instruction, would be equally effective with low-achieving student populations. For this comparison, we recruited 42 fifth and sixth grade students from a low-achieving school (School L) as the *training group* and 181 fifth through eighth grade students from a high-achieving school (school H) as the *comparison group*. Thus, the comparison group contained both same-grade counterparts to the training group and those at one and two grade levels above. This design enabled the gathering of cross-sectional benchmarks of CVS performance by grade level in the high achieving school. The training group's CVS performance, before and after training, would be compared with these benchmarks. School L had 90% African American students and 85% of its students were on free lunch programs. School H had 10% African American students and less than 10% of its students were on free lunch programs. On Cognitive Tests of Basic Skills [CTBS] Terra Nova tests (CTB/McGraw-Hill, 2001a,b) adopted by both schools, School L's students' scores were significantly below those of School H's students in every academic

subject test (e.g., reading, language, math, science) and intelligence test (e.g., verbal and non-verbal reasoning), by an average of 12 to 25 national percentile points.

During the study, students in the training group received whole-class instruction from one researcher-teacher for a total of four class periods. Students in the comparison group did not receive instruction from the research team, but had studied CVS as it was included in both their regular science instruction and curriculum materials. Of course, in both groups, CVS performance reflected the combination of students' general intellectual abilities, the quality of their school's science program, and their training and performance in related areas, such as reading and math.

In preparation for revising and adapting the CVS lesson plans (from Study 2) for the training group, we logged more than 100 h of classroom observations in School L. While our long-term goal is to allow our methods to empower the teachers' lesson planning, we chose to use the researcher as the classroom instructor as a preparatory step towards that objective. Putting the researcher in the classroom teacher's shoes minimized issues of implementation fidelity of the newly revised lesson plan and gave the researcher a first-hand experience of the classroom environment. This design choice obviously confounded the effect of a novel instructor with the instructional method. At the onset of the study, we anticipated two possible consequences of this intentional confound: (1) the researcher might succeed because of his knowledge base in cognitive science, independent of the CVS training method; and (2) the researcher might fail because of his lack of classroom skills, in spite of the CVS training method.

Because evaluations of individual students' hands-on CVS performance were not logistically feasible during a tightly packed school schedule, we modified how we measured CVS use: Measures traditionally used as posttests were utilized as formative assessments. Specifically, *Very near transfer* was measured by students' performance on paper and pencil forms that asked them how they would design their experiments with the physical apparatus (e.g., balls down the ramp). *Near transfer* was measured by students' performance in dyads as they worked with physical materials to determine the factor that influences the period of simple pendulums. Data for both very near and near transfer measures were collected from students' worksheets used during the class sessions. These formative assessments provided information that allowed for the targeted repetition of CVS instruction if a significant portion of students appeared not to have mastered it. The "real" posttest (i.e., administered after all instruction had been completed) was the *remote transfer* incorporating both the research instruments from Study 1 and Study 2 and standardized test items. The standardized test items were a diverse assortment of multiple choice and constructed response questions (e.g., National Assessment of Educational Progress [NAEP], 1996, 1999; Third International Math and Science Study [TIMSS], 1995, 1999; CTBS, CTB/McGraw-Hills, 2001a,b). All items required some knowledge of CVS, though none explicitly mentioned that students should apply the "control of variables" strategy. All items were used in their original forms to maintain comparability to national and international benchmark data. Both the training group and the comparison group were administered these remote transfer items.

The overall results showed promise that our instruction could narrow the achievement gap. Before training, students in the training group (School L) performed significantly below the benchmark set by their same-grade counterparts in the comparison group (School H). The training group's performance improved significantly during training. In the formative assessment of very near transfer, the percent of training group students who correctly used CVS improved from less than 20% at the beginning of instruction to over 80% at the end of instruction. In the formative assessment of near transfer, the percent of training group dyads who correctly used CVS in the new domain (i.e., pendulum) and new context (i.e., hands-on, group work) improved from less than 40% at the beginning of the class to over 90% at the end of class.

In the between-group comparisons of remote transfer performance, the training group students (School L) performed significantly better than their same-grade comparison group counterparts (School H) on evaluating and correcting experiments (based on measures used in Study 1 and Study 2), with an effect size of 1.02 using pooled standard deviation from the two samples (Rosnow and Rosenthal, 1996). Only the eighth graders in the comparison group matched training group students' performance on these items (see Fig. 4.)

The most policy-relevant outcome for this study is how the training group performed on remote transfer items selected from standardized tests used for benchmarking at national and international levels.

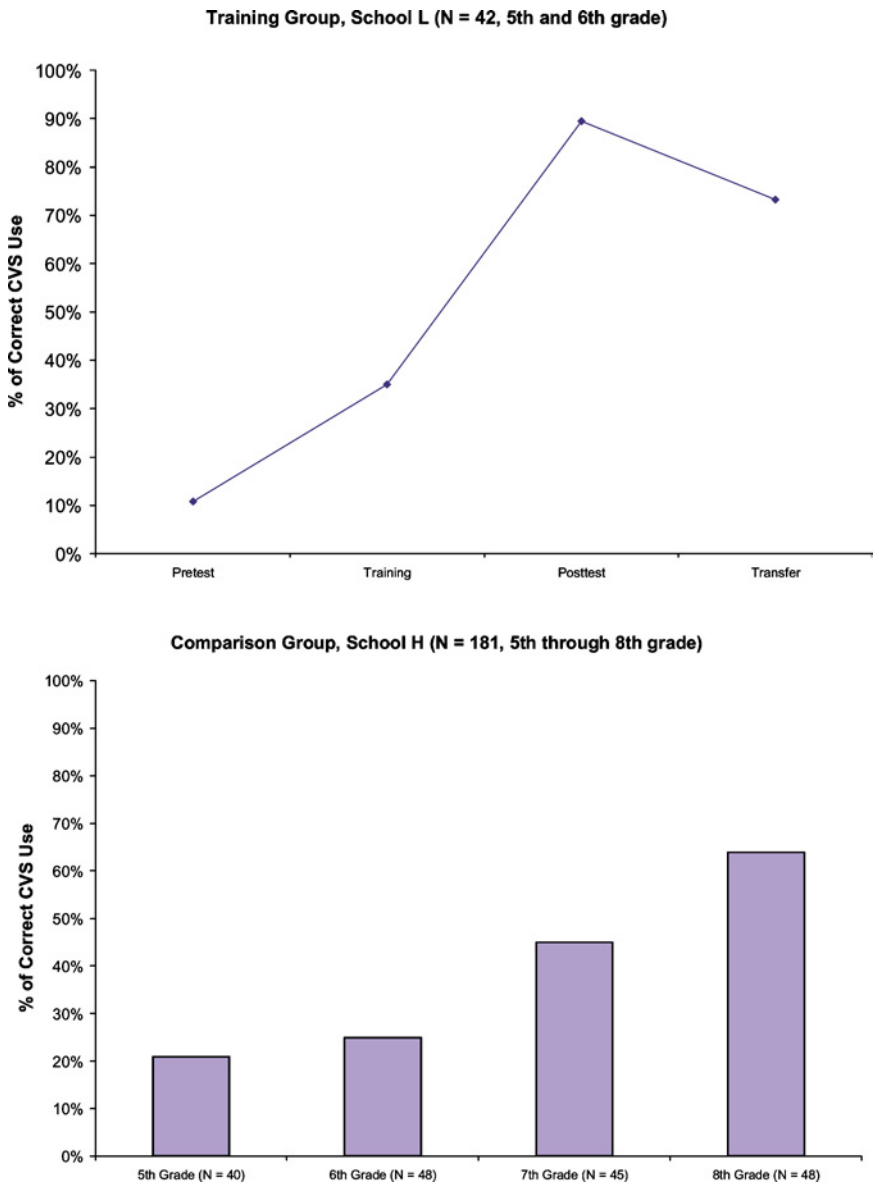


**Fig. 4.** CVS performance of *training group* across a 2-week period (including 4 days of training), compared with CVS performance of the *comparison group* (without training) across four grade-levels.

The training group met or exceeded the benchmarks on all items for which the original test makers provided a U.S. average score and/or an international percent-correct average (see example item, Fig. 5 and the corresponding performance comparison, Fig. 6). However, the training group's performance on these standardized test items did not significantly exceed that of the same-grade comparison group.

These results demonstrate that the CVS lesson plans, adapted from the learning via explicit instruction method previously found effective with high-achieving student populations, can narrow the achievement gap on CVS performance. However, the training group's performance did not entirely transfer to the selected battery of standardized test items. We briefly describe our analysis of the reasons for this less than optimal transfer.

The primary instructional goal of all of our CVS training studies is focused in a single cell of our framework of scientific thinking (see Cell E,

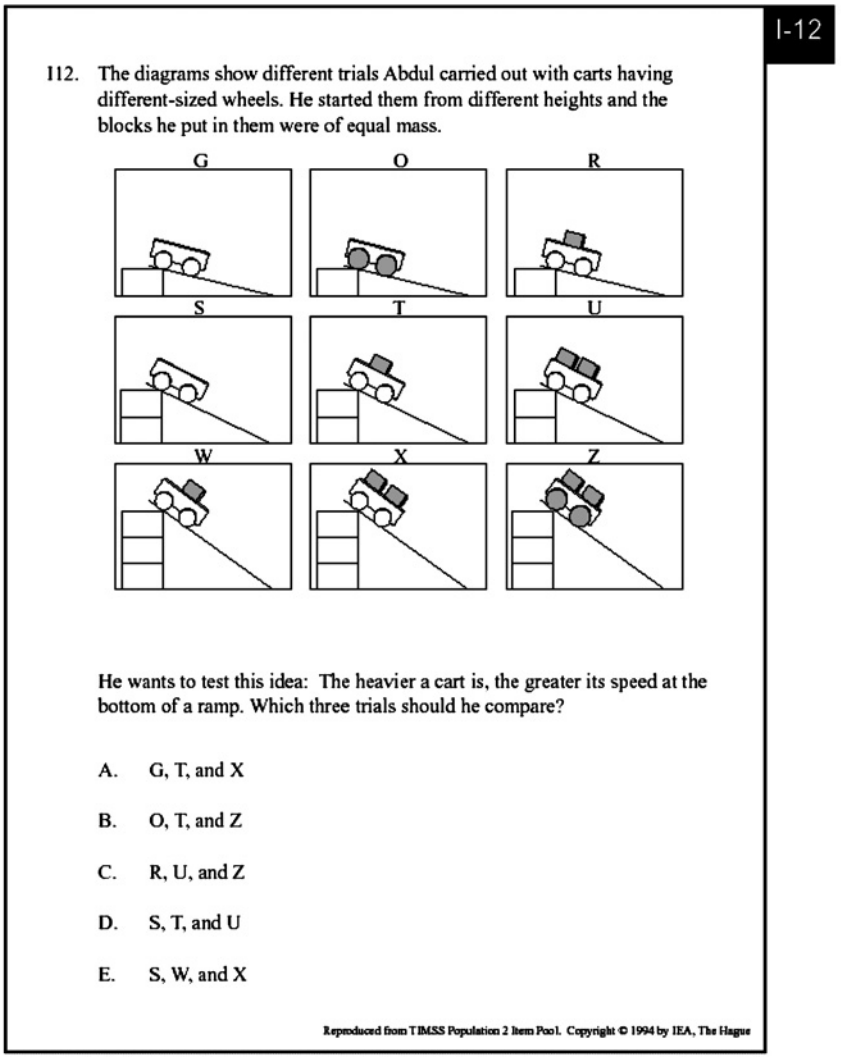


**Fig. 5.** Sample standardized test item relevant to CVS. Reprinted from TIMSS 1995 Released Set, Item I-12. Correct answer is D.

Table II)—domain general experimental design. The goals of science education, noted earlier, make more holistic demands upon the students. The application of experimental design skills is dependent on other related skills, such as the ability to identify target variables from a question statement or to identify potentially confounding variables in an experimental setup. Our training group students in School L performed the worst on those items requiring the most integration across cells (e.g., linking hypothesis with experimental design) and providing the least scaffolding for underlying logical structures (e.g., not stating explicitly what the key variables are). This raises the question of whether a training procedure developed in the Bohr's quadrant, which narrowly focuses on a *single* domain-general reasoning process, can be effective in practice where the task demands require the integration of domain knowledge with *multiple* reasoning processes.

We also find evidence that performance on standardized science test items depended on other academic skills. In a stepwise regression to account for CVS performance on standardized test items, we entered as predictors the current year achievement scores in six CTBS subject tests (CTB/McGraw-Hill, 2001a,b) and our measure of CVS performance. We find that the most significant predictor for the performance of School H's students on both multiple

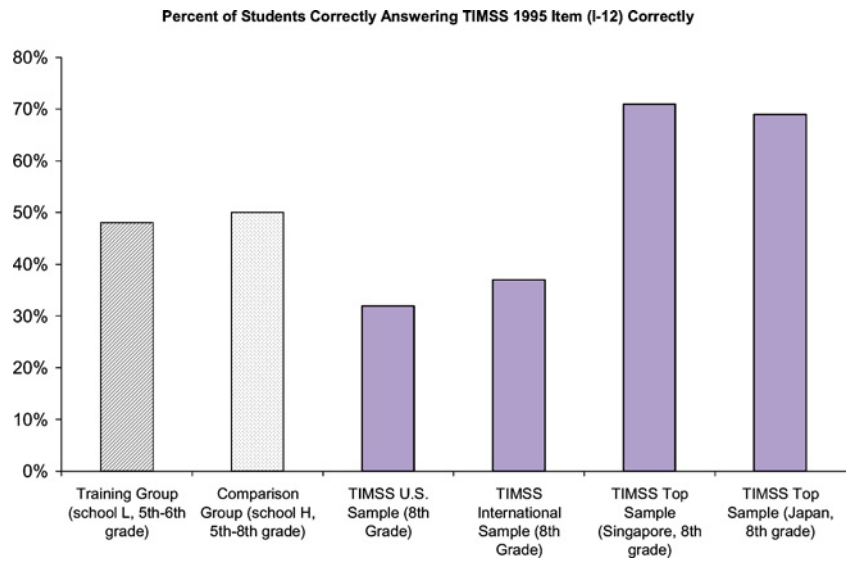


**Fig. 6.** *Training group's* and *comparison group's* performance on a sample TIMSS 1995 8th Grade Science item, compared with TIMSS benchmarks collected from U.S. sample, international sample, and international leaders in science achievement.

choice and constructed response items was their current year science achievement score. In stark contrast, the only significant predictor for the performance of School L's students on multiple choice items was their current year reading achievement score, and for constructed response items, their language arts achievement score. Thus, performance on standardized tests of accountability seem to place high demands on students' reading and writing skills requiring them to understand descriptions of experiments and explain their reasoning and justifications. Students in the training group could not effectively read and write in relation to CVS, even though they exceeded the comparison group in their understanding of CVS (as assessed by tasks whose instructions were read aloud and required only responses with check marks and drawing). Reading and writing science are not the focus of our studies or of most basic research of scientific reasoning. In fact, we minimized the demand of these skills on the participants in our past studies. Probes and cover stories were always read aloud to the participants and participants were only required to speak their answers and check "good" or "bad" experiments. Nevertheless, these skills are fundamentally important for test performance and long-term achievement in science or any academic subject. Is it within the research program's scope to address these issues? Or are we venturing into territories too messy and too far beyond our specialization to control and influence in the psychology laboratory or classroom settings? These are questions we continue to evaluate while searching for and designing "what works" in the field setting.

## CROSSING THE QUADRANTS' BOUNDARIES

By directing our research program back and forth between the laboratory and the classroom, we have traversed the boundaries between basic and applied research. Although the psychology laboratory has enabled us to explore questions with controlled methods and empirical confidence, the classroom has enriched our research agenda with a pipeline of rich and important questions. The five studies reviewed here constitute a work in progress towards making basic research findings relevant, for both high and low achieving student populations. While the stated and unstated variables in our instruction are many, we eliminated the hands-on manipulation of physical materials as a significant or interacting factor in our instruction. For the variables we could not eliminate, successive replications gave us confidence about the efficacy of our basic instructional design in spite of the possible interacting effects of materials, task domains, representations, limitations of measures, and instructor characteristics.

We are certainly not the first to have meandered our way across the boundary between basic research in the psychology laboratory and applied research in instructional development for classrooms. The long history of this type of effort is richly described in Lagemann's (2000) history of educational research. As Lagemann notes, because the dominant theories in psychology framed such efforts, for a long time they were heavily influenced by the behaviorist tradition. The Sixties produced several new efforts based on the excitement and promise of the "cognitive revolution" (e.g., Atkinson, 1968; Gagne, 1968; Glaser and Resnick, 1972; Suppes and Groen, 1967). More recent efforts, based on emerging perspectives on cognition and cognitive development, is exemplified by the work of Case and colleagues (e.g., Case, 1978; Case, 1992; Case and Okamoto, 1996), who pioneered research that sought both basic understanding of child development and effective instruction for children. Brown and colleagues (Brown, 1992; Brown and Campione, 1994; Palinscar and Brown, 1984) led a successful research program that took "reciprocal teaching" from the laboratory to the classroom across many school topics. Anderson and colleagues (Anderson *et al.*, 1995) developed an effective intelligent tutoring system for learning algebra—now used in thousands of schools nationwide—from a basic research program in computational modeling of cognition. We owe much of our approach and research motivation to these and many other forerunners.

Although we have attempted to organize our content in a logically coherent progression, we do not want to give the false impression that this sequence of studies was carefully planned and coordinated years in advance. Rather, much of what we described happened through the planned and unplanned collaboration between basic and applied researchers who were drawn to our research group at different times. What has helped most to make this research program informative for both basic research and applied research is perhaps a general openness and willingness among all researchers involved to cross traditional research barriers and tread into unfamiliar real-world territories.

## CAUTION: WHAT'S IN A NAME?

In this final section, we discuss our reaction to the way that our research program has impacted some of the popular debates on educational practice. We do this with some trepidation, because a discussion of media responses seems inappropriate in an article about scientific research. Moreover, we have only a handful of studies, each of relatively small scope, upon which some of these issues have been argued. Nevertheless, having described our venture into the real world of elementary science school classrooms, we feel a responsibility to acknowledge that we vastly underestimated the degree to which educational practice is determined by forces other than the results of research. Thus, we are (now!) mindful of the way in which our results can be used to support or attack specific aspects of science education practice and policy. This is an inescapable consequence of working in the "real world." We recognize that, as researchers, we cannot use the naive proclamation of "let the data speak for itself" as an excuse to delegate the responsibility of interpreting our findings in the real world context.

Of the five studies reported here, Study 1 (Chen and Klahr, 1999) and Study 4 (Klahr and Nigam, 2004) have received a substantial amount of media attention, especially within the educational research and science education community. Not only have they been cited in a variety of publications addressed to various educational and psychological communities and the general public (Adelson, 2004; Begley, 2004; Cavanagh, 2004; Crane, 2005; "Stand and deliver . . . or let them discover?," 2004), they also have been used to support one position or another in the "science wars." The headlines often appear far more conclusive than our results warrant. One reads "Researchers say Direct Instruction, rather than 'Discovery Learning,' is Best Way to Teach Process Skills in Science" (1998). Another reports our findings in Study 4 (Klahr and Nigam, 2004) under the headline "NCLB Could Alter Science Teaching" (Cavanagh, 2004), even as we are still in the process of replicating the study to assess the robustness of its basic findings. We hope it is clear to the readers of this article that such headlines go far beyond the empirical evidence provided in our studies. In the series of studies reported here, we investigated only a single process skill: the control of variables in experimental design. We conclude, based on initial and replicated results, that direct and explicit instruction, combined with experimentation using physical or virtual materials, is more effective for teaching CVS than simply giving children opportunities for minimally guided discovery. This conclusion is also consistent with findings from a prior study in a non-science domain—teaching children how to debug their computer programs (Klahr and Carver, 1988)—which found that children were

more effective at debugging computer programs following highly directive and explicit teaching of multistep debugging procedures than if they were left to discover these procedures on their own. Despite this accumulation of evidence, we do not have concrete hypotheses regarding the ways in which other science process skills, such as evidence evaluation and hypothesis generation, can be trained. From the studies reported here, we can say that our particular specification of learning via explicit instruction worked better than an extreme form of learning via discovery for learning CVS. We certainly do not know if our CVS instruction is the "best way" to teach CVS, or if Direct Instruction is the best way to teach *all* process skills.

Because of this kind of media reporting of our results described above, others are concerned that our findings may be used to "conclude that direct instruction is the best way to teach science" (Tweed, 2004), to promote lecture-based passive learning ("Stand and deliver . . . or let them discover?," 2004), and to equate our specification of discovery learning with the more moderate (and presumably, more often used) versions of guided or scaffolded inquiry. As we discussed above, we share the concern that our findings may be misinterpreted as evidence to promote one method over another for science education as a whole. However, within the context of teaching and learning CVS, we did show that direct instruction need not result in a passive student role, as is often suggested (e.g., Tweed, 2004). In our studies, though the students were "told" CVS, they actively applied the strategy via hands-on experiments and reasoned based on the strategy to evaluate evidence. We maintain that the combination of direct instruction (from the instructor) and active processing and application (by the learner) contributed most significantly to the learning improvements in our studies. Lastly, we do not believe our specification of "discovery learning" is so far-fetched that it does not reflect any current practices. In the aforementioned sample CVS lesson taken from the National Science Education Standards (NRC, 1996), the teacher offered much guidance and facilitation of the students' inquiry experience, but did not offer explicit instruction or feedback regarding the control of variables as an experimental strategy (at least not according to the two page description). Based on our results, we argue that, in the case of CVS strategy, students have difficulty discovering it even when they have been given clear goals and hands-on materials to conduct purposeful inquiry. We suggest that, for a domain-general science process skill such as CVS, it may be more efficient to instruct the students explicitly and directly and thus enable the students to engage more meaningfully in experimentation and discovery across domains.

Perhaps the most unanticipated aspect of the characterization of our work in these press reports and expressions of concern is the power of labels. In our very first experiment in this series, we used bland characterizations of our different training conditions. The descriptions were fairly explicit and well specified (Chen and Klahr, 1999, p. 1101, with condition names emphersised as in the original text):

> Part I included hands-on design of experiments. Children were asked to set up experimental apparatus to test the possible effects of different variables. . . .
>
> In the *Training—Probe* condition, children were given explicit instruction regarding CVS (the Control of Variables Strategy). Training occurred between the Exploration and Assessment phases. It included an explanation of the rationale behind controlling variables as well as examples of how to make unconfounded comparisons. Children in this condition also received probe questions surrounding each comparison (or test) that they made. A probe question before the test was executed asked children to explain why they designed the particular test they did. After the test was executed, children were asked if they could "tell for sure" from the test whether the variable they were testing made a difference and also why they were sure or not sure. In the *No Training—Probe* condition, children received no explicit training, but they did receive the same series of probe questions surrounding each comparison as were used in the Training—Probe condition. Children in the *No Training—No Probe* Condition received neither training nor probes."

By Study 4, we simplified the terminology by using labels that we felt were apt characterizations of the distinction between two approaches—direct instruction and discovery—never anticipating the controversy they would arouse in the science education community. We described the two conditions as follows (Klahr and Nigam, 2004, p. 663):

> Children in the direct-instruction condition observed as the experimenter designed several additional experiments—some confounded, and some unconfounded—to determine the effects of steepness and run length. For each experiment, the instructor asked the children whether or not they thought the design would allow them to "tell for sure" whether a variable had an effect on the outcome. Then the instructor explained why each of

> the unconfounded experiments uniquely identified the factor that affected the outcome, and why each confounded experiment did not. Children in the discovery condition instead continued to design their own experiments, focused on the same two variables that the direct-instruction children were focusing on, but without any instruction on CVS or any feedback from the experimenter.

And then we went on to be even more careful about a possible misinterpretation of our conditions (Klahr and Nigam, 2004, p. 663):

> It is important to note that in our operationalization, the difference between direct instruction and discovery learning does not involve a difference between "active" and "passive" learning. In both conditions, students were actively engaged in the design of their experiments and the physical manipulation of the apparatus. The main distinction is that in direct instruction, the instructor provided good and bad examples of CVS, explained what the differences were between them, and told the students how and why CVS worked, whereas in the discovery condition, there were no examples and no explanations, even though there was an equivalent amount of design and manipulation of materials.

In hindsight, we may have muddied the interpretation of our findings by incorporating popular terminology like "direct instruction" and "discovery learning" into articles and public presentations of Study 4. Only when we tuned in to the recent political debate in California about the permissible amounts of "hands-on science" vs. "direct instruction" (Alberts and Wheeler, 2004; Janulaw, 2004; Stickel, 2004; Strauss, 2004; Wheeler, 2004) did we become fully aware of how easy it is for someone to pick up a terminology, and imbue it with whatever meaning suits the purpose of an argument.

As we become better informed about the debates in educational policy related to our research, we become increasingly wary of "naming" research-based training methods using common educational terminologies. Perhaps we should do what physicists do: either invent novel terms—like "quark" and "lepton"—or use everyday words that cannot possibly be interpreted in their conventional sense, like "charm," and "flavor." When we adopt a widely used terminology (e.g., discovery learning, or direct instruction), we are burdened with ensuring that our implementation is consistent with the term's meaning in the educational community. This becomes a daunting task when the term itself is historically poorly defined but hotly contested, as most commonly used terms are in both cognitive and educational literature.

One thing is clear from all of this: it is essential for the field of education to make much more precise use of terminology before moving on to public debates and policy decisions. Indeed, it is surprising that when education researchers and science educators join in heated debates about discovery learning, direct instruction, inquiry, hands-on, or minds-on, they usually abandon one of the foundations of science—the operational definition. The field of science cannot advance without clear, unambiguous, operationally defined, and replicable procedures. Education science is no exception.

Research conducted in any of Stokes' quadrants depends on others' interpretation of its findings for generalizability and actual use. By clarifying the context, motivation, and findings from our research program, we hope this paper facilitates the interpretation of our work. Ultimately, it is up to the reader to draw implications from research, not the researcher. We hope the research paths we took across the traditional boundaries separating basic and applied research will encourage greater openness to use-inspired research questions. We hope that other researchers find it worthwhile to embark on journeys between the psychological laboratory and the real-world classrooms, even with a few necessary, if not entirely enjoyable, stumbles through the policy and media worlds.

## ACKNOWLEDGMENTS

As it takes a village to raise a child, it takes a lot of folks to support and contribute to a research program to study how children learn about science. The work reported here represents the cumulative efforts of many people and institutions. Funding came from a variety of sources, including grants from NICHD (HD 25211), NIH (MH19102), the James S. McDonnell Foundation (96–37), the National Science Foundation (BCS-0132315), and the Cognition and Student Learning program at the Institute of Education Science (R305H030229). Many colleagues have made constructive suggestions for improving the various publications upon which this paper is based. These include John Anderson, Sharon Carver, Norma Chang, Ken Koedinger, Amy Masnick, Brad Morris, Chris Schunn, Robert Siegler, Lara Triona, and Corinne Zimmerman. Our thanks to them and to Mari Strand Cary and Milena Nigam for comments on earlier versions of this paper. Over the years we have been fortunate to have some outstanding

research assistants, including Amanda Jabbour, Sharon Roque, Audrey Russo, Jennifer Schnakenberg, Anne Siegel, and Jolene Watson. Their dedication and competence contributed substantially to this work. And of course, none of this could have been done without the enthusiastic cooperation of the children, parents, principals and teachers at the many schools where we have conducted our studies, including, Ellis School, Immaculate Conception School, Sacred Heart Elementary School, Shadyside Academy, St. Teresa of Avila, Carlow College Campus School, Winchester-Thurston School, and St. James School. We thank them all for their participation.

## REFERENCES

Adelson, R. (2004). Instruction versus exploration in science learning. *Monitor on Psychology* 35: 34–36.

Alberts, B., and Wheeler, G. (2004, March 4). Letter to California state board from national academy of sciences. Retrieved on April 7, 2004 from http://science.nsta.org/nstaexpress/lettertocalffromgerry.htm.

American Association for the Advancement of Science. (1993). *Benchmarks for Science Literacy,* Oxford University Press, New York.

Amsel, E., and Brock, S. (1996). Developmental changes in children's evaluation of evidence. *Cognitive Development* 11: 523–550.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences* 4: 167–207.

Atkinson, R. (1968). Computerized instruction and the learning process. *American Psychologist* 23: 225–239.

Barnett, S. M., and 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*, p. B1. Retrieved December 10, 2004 from http://online.wsj.com/article/0,,SB110263537231796249,00.html.

Bernard, R. M., Abrami, P., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., Wallet, P. A., Fiset, M., and Huang, B. (2004). How does distance education compare to classroom instruction? A meta-analysis of the empirical literature. *Review of Educational Research* 74: 379–439.

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

Brown, A. L., and Campione, J. C. (1994). Guided discovery in a community of learners. In McGilly, K. (Ed.), *Classroom Lessons: Integrating Cognitive Theory and Classroom Practice,* MIT Press, Cambridge, MA, pp. 229–272.

Bullock, M., Ziegler, A., and Martin, S. (1992). Scientific thinking. In Weinert, F. E., and Schneider, W. (Eds.), *LOGIC Report 9: Assessment Procedures and Results of Wave 6,* Max Plank Institute for Psychological Research, Munich.

Carey, S. (1985). *Conceptual Change in Childhood,* Bradford Books, MIT Press, Cambridge, MA.

Carnegie Mellon researchers say direct instruction, rather than 'discovery learning' is the best way to teach process skills in science (1998, February 13). Retrieved April 19, 2004 from http://www.eurekalert.org/pub_releases/1998-02/CMU-CMRS-130298.php.

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

Case, R. (1978). Intellectual development from birth to adulthood: A neo-Piagetian interpretation. In Siegler, R. (Ed.), *Children*'s *Thinking: What Develops?,* Lawrence Erlbaum Associates, Hillsdale, NJ.

Case, R. (1992). *The Mind*'s *Staircase: Exploring the Conceptual Underpinnings of Children*'s *Thought and Knowledge,* Erlbaum, Hillsdale, NJ.

Case, R., and Okamoto, Y. (1996). The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development* (Serial No. 246).

Cavanagh, S. (2004), November 10). NCLB could alter science teaching. *Education Week* 24(11): pp. 1: 12–13.

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

Chinn, C. A., and Malhotra, B. (2001). Epistemologically authentic scientific reasoning. In Crowley, K., Schunn, C. D., and Okada T. (Eds.), *Designing for Science: Implications for Everyday, Classroom, and Professional Settings,* Erlbaum, Mahwah, NJ, pp. 351–392.

Crane, E. (2005). The Science Storm. *District Administration*, #3(March).

CTB/McGraw-Hill (2001a). *TerraNova CAT Complete Battery Plus Level 15, Form C,* CTB/McGraw-Hill, Monterey, CA.

CTB/McGraw-Hill (2001b). *TerraNova CAT Complete Battery Plus Level 16, Form C,* CTB/McGraw-Hill, Monterey, CA.

Elkind, D. (2001). Much too early. *Education Next* 1: 8–21.

Gagne, R. (1968). Contributions of learning to human development. *Psychological Review* 75: 177–191

Glaser, R., and Resnick, L. (1972). Instructional psychology. *Annual Review of Psychology* 23: 207–276.

Hiebert, J., Gallimore, R., and Stigler, W. (2002). A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational Researcher* 31: 3–15.

International Association for the Evaluation of Educational Achievement. (1998). *TIMSS Science Items: Released Set for Population 2 (Seventh and Eighth Grades).* Retrieved on September 16, 2004 from http://timss.bc.edu/timss1995i/TIMSSPDF/BSItems.pdf.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior* 17: 649–667.

Janulaw, S. (2004, January 9). *Letter to California curriculum commission from California science teachers association.* Retrieved on April 7, 2004 from http://science.nsta.org/nstaexpress/ltr_to_commission.htm.

Klahr, D., and Carver, S. M. (1995). Scientific Thinking about Scientific Thinking. *Monographs of the Society for Research in Child Development* #245, 60: 137–151.

Klahr, D., and 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., and Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin* 125: 524–543.

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

Klahr, D., and Chen, Z. (2003). Overcoming the "positive capture" strategy in young children: Learning about indeterminacy. *Child Development* 74: 1256–1277.

Klahr, D., Chen, Z., and Toth, E. (2001). Cognitive development and science education: Ships passing in the night or beacons of

mutual illumination? In Carver, S. M., and Klahr, D. (Eds.), *Cognition and Instruction: 25 Years of Progress,* Erlbaum, Mahwah, NJ.

Klahr, D., Fay, A. L., and Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology* 24: 111–146.

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

Kuhn, D., Amsel, E., and O'Loughlin, M. (1988). *The Development of Scientific Thinking*. Harcourt, Brace & Jovanovich, New York.

Lagemann, E. C. (2000). *An Elusive Science: The Troubling History of Educational Research*. Chicago: University of Chicago Press.

Lagemann, E. C. (1996). Contested terrain: A history of education research in the United States, 1890–1990. *Educational Researcher* 26: 5.

Li, J., and Klahr, D. (in press). The psychology of scientific thinking: Implications for science teaching and learning. In Rhoton, J., and Shane, C. A. (Eds.), Issues and Trends in science Learning for the 21st Century, NSELA/NSTA.

Loveless, T. (1998). *The Use and Misuse of Research in Educational Reform*. Brookings Papers on Educational Policy: 1998, pp. 279–317.

Masnick, A. M., Klahr, D., and Morris, B. J. (in press) Separating signal from noise: Children's understanding of error and variability in experimental outcomes. In Lovett, M., and Shah, P. (Eds.), *Thinking With Data,* Erlbaum, Mahwah, NJ.

Masnick, A. M., and Klahr, D. (2003). Error matters: An initial exploration of elementary school children's understanding of experimental error. *Journal of Cognition and Development* 4: 67–98.

Masnick, A. M., and Morris, B. J. (2002). Reasoning from data: The effect of sample size and variability on children's and adults' conclusions. In Gray, W. D., and Schunn, C. D. (Eds.), *Proceedings of The Twenty-Fourth Annual Conference of The Cognitive Science Society,* Erlbaum, Mahwah, NJ, pp. 643–648.

McCloskey, M. (1983). Naive theories of motion. In Gentner D., and Stevens A. (Eds.), *Mental Models*, Erlbaum, Hillsdale, NJ, pp. 229–324.

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

Metz, K. E. (1995). Reassessment of developmental constraints on children's science instruction. *Review of Educational Research* 65: 93–127.

National Center for Education Statistics (n.d.). *The Nation's Report Card (NAEP): 1996 Assessment Science Public Release Grade 4*. Retrieved on September 16, 2004 from http://nces.ed.gov/nationsreportcard/itmrls/sampleq/96sci4.pdf.

National Center for Education Statistics (n.d.). *The Nation's Report Card (NAEP): 1996 Assessment Science Public Release Grade 8.* Retrieved on September 16, 2004 from http://nces.ed.gov/nationsreportcard/itmrls/sampleq/96sci8.pdf.

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

National Science Teachers Association. (1990, January). *Position Statement: Laboratory Science*, Retrieved on November 16, 2004 from http://www.nsta.org/postitionstatement&psid=16.

National Science Teachers Association. (2002, July). *Position Statement: Elementary School Science.* Retrieved on November 16, 2004 from http://www.nsta.org/positionstatement&psid=8.

No Child Left Behind Act of 2001. (2002). *Public Law 107–110-January 8, 2002. 107th Congress*. Washington, DC.

Palinscar, A., and Brown, A. (1984). Reciprocal teaching of comprehension-fostering and comprehension monitoring activities. *Cognition and Instruction* 1: 117–175.

Pennsylvania Department of Education. (2002). Preparing for the STEE grade 7 assessment. In *STEE Assessment Handbook,* Retrieved March 15, 2004 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/STEEAssessHandbook.pdf.

Piaget, J. (1970). Piaget's Theory. In Mussed, P. H. (Ed.), *Carmichael*'s *Manual of Child Psychology: Vol. 1* (3rd edn.) Wiley, New York, pp 703–772.

Rosnow, R. L., and Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods* 1, 331–340.

Ruby, A. (2001). *Hands-on Science and Student Achievement*, RAND, Santa Monica, CA, Retrieved on December 1, 2004 from http://www.rand.org/publications/RGSD/RGSD159/.

Ruffman, T., Perner, J., Olson, D. R., and Doherty, M. (1993). Reflecting on scientific thinking: Children's understanding of the hypothesis-evidence relation. *Child Development* 64: 1617–1636.

Samarapungavan, A., and Wiers, R. W. (1997). Children's thoughts on the origin of species: A study of explanatory coherence. *Cognitive Science* 21: 147–177.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology* 32: 102–119.

Schauble, L., Glaser, R., Duschl, R., Schulze, S., and John, J. (1995). Students' understanding of the objectives and procedures of experimentation in the science classroom. *Journal of the Learning Sciences* 4: 131–166.

Schauble, L., Klopfer, L., and Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching* 18, 859–882.

Shaklee, H., and Paszek, D. (1985). Covariation judgment: Systematic rule use in middle childhood. *Child Development* 56: 1229–1240.

Stand and deliver . . . or let them discover? (2004, November). *District Administration*, p. 79.

Stickel, S. (2004, January 29). *Curriculum Commission: Approval of Criteria for Evaluating k-8 Science Instructional Materials for 2006 Primary Action,* Retrieved on April 7, 2004 from http://www.cde.ca.gov/be/pn/im/documents/infocibcfirfeb04item01.pdf.

Stokes, D. E. (1997). *Pasteur's Quadrant: Basic Science and Technological Innovation,* Brookings Institution Press, Washington, DC.

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

Strauss, V. (2004, February 3). Back to basics vs. hands-on instruction: California rethinks science labs. *The Washington Post*, p. A12.

Suppes, P., and Groen, G. (1967). Some counting models for first grade performance data on simple addition facts. In Scandura, J. (Ed.), *Research In Mathematics Education,* National Council of Teachers of Mathematics, Washington DC.

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

Triona, L. M., and Klahr, D. (2002, August). Children's developing ability to create external representations: Separating what information is included from how the information is represented. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, 1044.

Triona, L. M., and Klahr, D. (2003). Point and click or grab and heft: Comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition and Instruction* 21: 149–173.

Tweed, A. (2004, December 15). Direct instruction: is it the most effective science teaching strategy? *NSTA WebNews Digest*. Retreived on January 3, 2005 from http://www.nsta.org/main/news/stories/education_story.php?news_story_ID=50045.

Uttal, D. H., Liu, L. L., and DeLoache, J. S. (1999). Taking a hard look at concreteness: Do concrete objects help young children to learn symbolic relations? In Balter, L., and Tamis-LeMonda, C. (Eds.), *Child Psychology: A Handbook of Contemporary Issues,* Garland, Harlen, CT, pp. 177–192.

Uttal, D. H., Scudder, K. V., and DeLoache, J. S. (1997). Manipulatives as symbols: A new perspective on the use of concrete objects to teach mathematics. *Journal of Applied Developmental Psychology* 18: 37–54.

Wasson, B., Lundvigsen, S. and Hoppe, U. (2003). Designing for change in networked learning environments. In Wasson, B., Ludvigsen S., and Hoppe, U. (Eds.), *Proceedings of the Computer Support for Collaborative Learning 2003 conference,* Kluwer Academic Publishers, Boston, pp. 405–409.

Wheeler, G. (2004, January 15). *Letter to California Curriculum Commission From National Science Teachers Association,* Retrieved on April 7, 2004 from http://science.nsta.org/nstaexpress/california_letter.htm.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review* 20: 99–149.