When the Rubber Meets the Road: Putting Research-based Methods to Test in Urban Classrooms

Junlei Li, David Klahr, and Amanda Jabbour
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213
Email: junlei@andrew.cmu.edu, klahr@cmu.edu

Abstract: We created a hypothetically "optimal" instructional scenario in which a knowledgeable researcher, under the guidance of an experienced classroom teacher, carried out a set of research-based science instruction in a low-SES urban school. The training group's performance was assessed by standardized test items and compared with that of a high-SES no training comparison group. The results demonstrate that instructional methods based on experimental psychological research have great potential for addressing the achievement gap problem. However, the analyses also reveal a significant discrepancy between low-SES students' performance on standardized test items and on alternative assessments with lower reading and writing demands. We discuss our methodological choices in making basic research more relevant to real world issues. We highlight the critical challenges of test validity, research relevancy, and reform feasibility for researchers and policymakers.

Introduction

Current federal legislation on education (No Child Left Behind Act, 2002) listed the following as necessary conditions for closing the achievement gap: 1) adopting research-based teaching practices, 2) having high quality teachers, and 3) using standardized tests as accountability measures. We tested this legislative assumption by putting instructional resources presumably meeting these three conditions into low-SES urban school science classrooms. Specifically, we adapted for classroom use instruction and materials previously tested in random-assignment experimental studies. We ensured teacher quality and implementation fidelity by having a researcher-teacher conduct the classroom instruction under the supervision of an experienced classroom teacher. We held ourselves accountable by measures consisting of both researcher-designed assessment instruments and original standardized test items selected from publicly or commercially available standardized tests, such as the Third International Mathematics and Science Study [TIMSS], the National Assessment for Educational Progress [NAEP], and the Terra Nova Comprehensive Test of Basic Skills [CTBS]. We tested whether this instructional scenario, designed specifically to meet the above NCLB criteria, would close the achievement gap in one single topic area in our local setting. Through this evaluation, we assessed the potential for basic psychological research to inform practice, explored the adaptations researchers need to make in transferring research-based instruction to classroom settings, and encountered new challenges for basic experimental studies to become practical and relevant for real-world classrooms.

We set as our instructional goal the mastery of an important component skill of scientific inquiry – how to design unconfounded scientific experiments. It is a skill explicitly included in nearly all national and state science inquiry standards (e.g., American Association for the Advancement of Science, 1993; National Research Council [NRC], 1996). Standardized science tests at international, national, and state levels have consistently assessed this particular inquiry component (e.g., International Association for the Evaluation of Educational Achievement – TIMSS 1995 released items; National Center for Education Statistics – NAEP 1996 released items). In addition to its prominence in K-12 science education, experimental design skill has also been well studied in cognitive and psychological research in terms of its acquisition (with or without training), development, and transfer (e.g., Ross, 1988; Chen & Klahr, 1999; Klahr & Nigam, 2004; Klahr & Li, 2005; Kuhn, Amsel, & O'Loughlin, 1988). Like only a few handful domains in science (e.g., forces and motions in physics), designing experiments is at a point of convergence among science standards, standardized assessments, and basic research. Such convergence makes the mastery of this skill an ideal instructional goal under which to explore issues of research, practice, and assessment.

Background

Children up to late elementary school age have only a partial grasp of the logic and procedure of experimental design (Kuhn, Garcia-Mila, Zohar, & Anderson, 1995; Schauble, 1996). However, these deficiencies do not imply a lack of developmental readiness to learn. Training studies have shown that various methods, ranging from mere exposure to experimental tasks to explicit instruction, can improve children's understanding and use of variable control (Case, 1974; Kuhn & Angelev, 1976; Ross, 1988; Schauble, 1996; Chen & Klahr, 1999).

Children's ability to transfer their experimental skill was significantly improved when training combined explicit instruction with hands-on experience rather than simply relying on self-directed hands-on exploration in carefully structured task domains (Chen & Klahr, 1999). The efficacy of the specific training method and materials used by Chen and Klahr has since been replicated in two additional experimental studies (Nigam & Klahr, 2004; Triona & Klahr, 2003) and adapted for a classroom validation study (Toth, Klahr, & Chen, 2000). Klahr and colleagues referred to experimental design skill as the control of variables [CVS]. Hereafter we will use "CVS training" to refer specifically to the particular method developed by Klahr and colleagues.

In what ways is CVS training used in these laboratory studies differ from conventional classroom practice? Klahr, Chen, and Toth (2001) contrasted CVS training with how experimental design was taught in observed science classrooms, select science textbooks, and even exemplary lessons described by the national standards (NRC, 1996). CVS training focused on helping students design unconfounded experimental comparison and differentiate it from confounded experimental comparisons. It did not overburden the learner with the cognitive and procedural demands of a complex scientific experimentation, which included hypothesis generation and complex experimental procedures among other things. Instead of relying solely on student-directed exploration with lab materials, CVS training required the instructor to explicitly provide instruction, evaluative probes, and corrective feedbacks. In the below script excerpted from the CVS training procedure, the instructional objective was to help children master CVS while investigating the relationship between spring lengths and hanging weight. The investigative question assigned to the learner was whether the springs' width, length, and wire size affect how long they stretched when pulled by different weights (*italics* represent planned emphasis in speech and gesture).

> (Experimenter sets up a confounded experimental comparison: Spring A is long, wide, thick, with a heavy weight and Spring B is short, narrow, thin, with a light weight.)
> Remember, I am trying to find out about whether one of these would stretch farther just because of its length. Do you think this is a smart choice to find out about *length*? Why? (Or: Why not?)
> What if you found out that one of these springs stretches more than the other one, could you tell for sure from this comparison that it was the *length* of the spring that made it stretch more? Why? (If they haven't already pointed out all the differences)
> What is different between these two springs? (If they haven't mentioned all the differences) Is there any other way they are different?
> Actually, you could *not* tell for sure from this comparison whether it was the *length* that made a difference in how far these two springs stretched. And the reason why you cannot tell for sure is that these two springs are different in other ways, not just length. These two springs also have different width and different wire size, right? And the weights on them are different. So it may be that one of them stretches more because it is wider or because the wire is thicker or because of the kind of weight on it. As you can see, if you compare these two springs, you can't tell whether it is the length or the width or the wire size or the different weight that makes one stretch farther than the other.

As the script above illustrates, the training procedure embodied many aspects of instructional strategies, including learning by doing (i.e., students actively manipulated variables and designed experiments before, during, and after instruction), metacognitive evaluation (i.e., students were asked to determine whether experiments were valid and whether they were sure of the conclusions), and explanation and justification (i.e., students were prompted to explain their reasoning). Explicit feedback and correction were given only after the learner had the opportunity to explore the task domain.

Using this training procedure, students from second to fifth grade can reach varying levels of mastery of CVS skills (summarized in Klahr & Li, 2005). Students had been able to transfer these skills over long time delay (over six months), or from hands-on task to paper-and-pencil task, or from one physical apparatus to another, or from constrained tasks to authentic unstructured tasks (e.g., evaluating a science fair poster). Most of the participants from fourth and fifth grade achieved near-ceiling mastery on immediate or delayed paper-and-pencil and hands-on assessment tasks.

Despite the success of CVS training in these studies, we had thus far skirted the most pressing issue of NCLB – the achievement gap. All of the previous studies of CVS training used participants from high-achieving and high-SES schools. In either one-on-one or whole-classroom settings, these students' attentiveness, cooperativeness, and ability to comprehend instruction and follow procedure were generally impeccable. We believe that our reliance (and that of many other basic researchers) on easy-to-access participant pools skews our results towards the higher performance end of the student population and leaves us with little empirical basis to evaluate usability and efficacy outside such settings. In addition, we have defined success criterion based on

questions of theoretical interest, such as long-term retention and far transfer. In a policy climate where achievement, proficiency, and gap are all measured by standardized test items, it seemed a bare minimum for researchers with applied interests to incorporate some standardized test items as performance measures. To address these limitations, the present study put CVS training to test in low-SES urban classrooms and measured outcomes using publicly and commercially available standardized test items in conjunction with researcher-designed measures.

Method

Design

Instead of a traditional treatment and control design within one single population group, we opted for a design that better reflected the achievement gap context for which NCLB sought research-based interventions. We adopted a pre and post design within a training group consisting of two classrooms of low-SES fifth and sixth grade students. For the posttest-only comparison group, we used fifth through eighth grade students from a high-SES school.

The training group received whole-class instruction from a researcher-teacher under the supervision of their regular science teacher. The comparison group received no training from the researcher-teacher. We presumed that the latter group of students would have learned experimental design skills from their regular science instruction. The efficacy of CVS training in closing the achievement gap was evaluated by between-group comparison of posttest performance.

We preferred this design over its more traditional alternative (Table 1). Our selected design dealt with the absolute difference between low-SES and high-SES student populations, which is at the core of the achievement gap question. The traditional design of control and treatment within a population would simply be making a "straw-man" comparison. We had already known that in the high-SES schools, training group outperformed no-training group. We had no reason to expect that relative difference to be different in a low-SES school. More importantly, achievement gap was always measured between high and low-SES, not amongst low-SES groups. For this study, we preferred going straight to the heart of the achievement gap issue. We did not believe the training and no-training comparison in our selected design was trivial. We regarded the training as an intervention strategy to help the lower group to "catch up" with the higher group, which was already far ahead in test scores across all subject areas, including science. Just to be sure that we did not create a "straw-man", we included in the high-SES comparison group both fifth and sixth grade students (same age as the training group) and seventh and eighth grade students.

Table 1: Selected Design vs. Alternate Designs

Selected Design	Traditional Design
low-SES: X O X	low-SES: X O X
high-SES: X	low-SES: X X

Participants

Training group participants were 42 fifth and sixth grade students in an urban low-SES parochial school in southwestern Pennsylvania. More than 90% of the training group students were African American and more than 80% of them were eligible for free and reduced lunch programs. Nonprofit foundations heavily subsidized the tuition payments to make the private school affordable to low-income families. Comparison group participants were 190 fifth through eighth grade students in a high-SES parochial school. Less than 10% of the comparison group were eligible for free and reduced lunches. There was no subsidy for the high-SES school. Training group mean national percentiles on 2004 district-mandated Terra Nova CTBS tests (CTB/McGraw-Hill) were significantly below ($p < 0.001$) those of the comparison group in every academic subject tested, including reading, language, math, science, verbal and non-verbal reasoning (mean difference from 12 to 25 percentile points).

Procedure

The CVS training lesson plans from a prior classroom study (Toth, Klahr, & Chen, 2000) was revised and adapted for the urban classroom after the researchers had spent more than 100 hours observing science teaching in urban schools. The adapted instruction incorporated formative assessments to flexibly adjust the length of instruction and used transfer tasks not simply as a way to measure learning, but also as a way to re-teach. The instruction had three stages.

Stage One: Initial CVS Acquisition

Students were first given a researcher-designed pretest consisting of evaluations of five different experimental setups (Figure 1). Only one of five setups represented an unconfounded experiment. Instructions were read aloud to the students. Students could use drawing and/or words to indicate their answers, as shown in the example below.

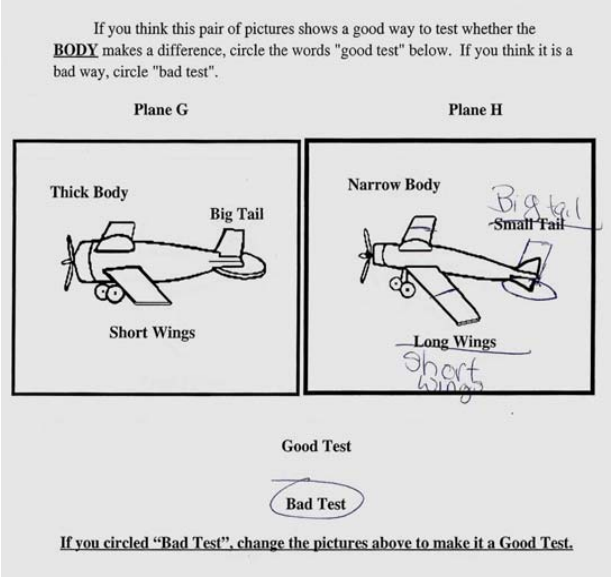

Figure 1. Sample item from the researcher-designed CVS pretest/posttest.

The researcher-teacher led the whole-class through a series of activities using a single set of physical apparatus for demonstration in the classroom (Figure 2). The use of a single set of apparatus reflected the actual condition in many urban schools where materials were scarce. The apparatus included two ramps with four bi-level variables. In the particular setup shown below, the nearer ramp had a rubber ball, a low steepness, a rough surface, and a shorter run and the farther ramp had a golf ball, a high steepness, a smooth surface, and a longer run. Thus, it represented a confounded comparison no matter what the hypothesis was (e.g., if one wanted to find out if the type of ball made a difference, this setup confounded the ball variable with three other variables). Though this apparatus was custom built for previous studies, it was relatively easy to rebuild.

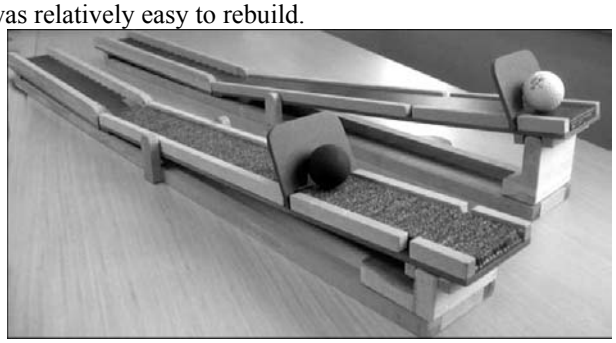

Figure 2. A confounded experimental comparison using the Ball and Ramp apparatus.

The instructor first familiarized all students with the apparatus. Then, he posed a question (e.g., "Does the type of the ball matter in how far it rolls?") and asked the students to write down their experimental design on paper. Then, one student (purposefully selected because her worksheet had showed a confounded experiment) would be asked to physically set up her experimental comparison while the rest of the class quietly observed. The instructor then asked the whole-class to evaluate whether the setup was a "good way" to answer the posed question. Students would offer various critiques. The instructor would stop the discussion when one or more valid critiques were given. The instructor confirmed the correctness of the valid critiques and added explicit explanations or restatements. Then, the instructor asked all the students for suggestions to revise the experiment. Once the experiment was revised and set up properly, it would be run three times. Students would record the data and answer the originally posed question.

This process was iterated (by changing the question, but not the probing and evaluating process) as more and more students began to design unconfounded experiments on paper and were able to offer valid critiques of

others' confounded experiments. The process terminated only when a vast majority of the students (over 80%) have mastered the use of CVS in this domain. In our study, it took four iterations of the above described procedure over three 40 minute class periods.

Stage Two: Transfer as Opportunity for Re-Training

Once the vast majority of the students achieved mastery in the first task domain, the instructor set up the transfer context that differed in the following aspects: 1) task domain was changed to a pendulum; 2) task demand was changed from designing experiment on paper to building and timing pendulum using string and paper clips; 3) social demand was changed from whole-class teacher-facilitated discussion to students working in dyads with teacher moving around the room to check and coach. The intent for this transfer phase was both to assess the robustness of initial CVS acquisition and also to provide instructional opportunities in a new domain. The instruction provided by the researcher-teacher was comparable to the probes used in the whole-class setting with some additional reminders for students to connect the present task with the previous task. This process iterated until a vast majority of the student dyads could independently design unconfounded experiments.

Stage Three: Transfer as Assessment

With approximately a 2-week delay after stage two, a posttest was given to the students. The posttest included a researcher-designed instrument (similar to pretest in form but different in domain) and a battery of original test items taken from publicly and commercially available standardized tests. It is important to note that the researcher-designed portion of posttest and pretest were read aloud to the students and could be answered with just checkmarks and drawings; in contrast, the standardized test items were administered in traditional manners where students read the items on their own and selected choices (for multiple choice items) or gave written paragraph-length responses (for constructed response items).

Results

Across all assessment instruments, the training group closed the achievement gap with the same-age students in the comparison group. However, only when the assessment instruments were researcher-designed (i.e., the instructions were read aloud and the responses required only drawings and checkmarks, see Figure 1) did the training group significantly exceed their same-age counterparts.

Researcher-designed CVS Assessment

There were five items on the researcher-designed portion of the posttest. An item is scored as correct if and only if the student was able to correctly classify the experiment as good/bad *and* to show how a "bad" experiment can be made "good". Table 2 summarized the scores for grade levels in training and comparison groups. The training group as a whole, consisting of low-SES fifth and sixth graders, performed significantly better than their same-age high-SES comparison group. The effect size was 1.02, $F(1, 129)=33.4$, $p<0.001$, using pooled standard deviation from the two samples (Rosnow & Rosenthal, 1996). Training group's performance even significantly exceeded that of comparison group's seventh graders, $F(1, 86)=8.5$, $p<0.01$ and was only matched by the comparison group's eighth graders.

Table 2: Performance on Researcher-designed CVS Assessment

Group (Grade Level)	Mean Score (Standard Deviation)
low-SES training (5th and 6th grade, n = 42)	3.25 (.70)
high-SES comparison (5th and 6th grade, n = 88)	2.38 (.35)
high-SES comparison (7th grade, n = 45)	2.88 (.43)
high-SES comparison (8th grade, n = 48)	3.07 (.44)

CVS Assessment Using Standardized Test Items

In addition to researcher-designed items, all participants received a common set of originally standardized test items. The items included multiple choice and constructed response questions intended to assess students' knowledge of the relationships among research question, variables, and experimental design. The items represented TIMSS (N-1, I-12, 8th grade, 1995), NAEP (K-033501, K-033502, K-033503, 4th grade, and, K-045101, 8th grade, 1996), Pennsylvania state test sample item, and Terra Nova CTBS items from grade five through eight.

Overall, the performance of the training group met or exceeded national and international benchmarks, where available. The scores on the two TIMSS items matched or exceeded 1995 U.S. and International 8th grade benchmarks (45% compared with 47% U.S. and 45% International for item N-1, and 48% compared with 32% U.S.

and 37% International for item I-12). The average scores for constructed response items from NAEP 1996 also matched the 4th and 8th grade U.S. national benchmarks. These results are promising considering the enormous test gap between racial and income groups reported for the very same TIMSS and NAEP tests from which these test items were selected. On less challenging test items (those from the PA state sample test and Terra Nova CTBS), the training group had an average of over 80% correct rate.

In between-group comparisons, the low-SES training group matched their same-age high-SES comparison group counterparts on standardized test items. However, they scored significantly lower than their higher-grade high-SES comparison group counterparts, ($F(1, 89)=6.47$, $p<0.05$). This result was at odds with between-group comparison on researcher-designed portion of posttest shown in Table 2.

To reconcile this difference, we tested the post hoc explanation that, in addition to underachievement in science, the training group also lacked necessary reading and verbal skills for comprehending standardized test items and constructing written responses. In a stepwise regression for both the researcher-designed and the standardized test portions of the posttest, we entered as predictors the current year achievement scores in six Terra Nova CTBS sub-tests (reading, language arts, math, science, verbal reasoning, and non-verbal reasoning). The strongest predictor for high-SES comparison group's CVS posttest performance was the students' overall Terra Nova CTBS science score, accounting for 25% of the variance in the researcher-designed portion and 41% of the variance in the standardized test item portion. This supported our design assumption that, in the high-SES setting, the students' mastery of experimental skills grew as part of their general science education. In stark contrast, the strongest and only predictor for the low-SES training group's CVS posttest performance was their Terra Nova CTBS language arts score, accounting for 26% of the variance in the researcher-designed portion and 53% of variance on the standardized test item portion. To further substantiate this explanation, we entered a measure of CVS performance collected immediately after initial acquisition (Stage One) into the regression for the training group (this analysis in inapplicable for the comparison group because they did not receive training). Table 3 summarized the resulted models and the accounted variance. Performance immediately following training only predicted the researcher-designed portion of posttest (i.e., read-aloud and figural response permitted).

Table 3: Reading and Language Arts Skills Confounding CVS Performance (significant predictors and variance)

	CVS Performance (initial acquisition)	Reading Score (Terra Nova, CTBS)	Language Arts Score (Terra Nova, CTBS)
Researcher-designed posttest items	20% (having accounted for reading)	26%	not significant
Standardized test posttest items (multiple choice)	not significant	45%	not significant
Standardized test posttest items (constructed response)	not significant	not significant	51%

Discussion

This study presented both a success story and a host of challenges for researcher and policymakers. It is a success story in that we were able to transfer a set of experimentally tested instruction, developed and validated using only high-SES students, to a low-SES school setting and achieve results that were practically meaningful in terms of closing the achievement gap. However, it highlighted, above all other issues, a critical challenge for science assessment. Our post hoc analyses to explain the performance discrepancy between two portions of the posttest suggest that standardized test items, including those used in the influential TIMSS, NAEP, PA state, and Terra Nova CTBS tests, make substantial demands on the students' ability to read and write. If our explanation based on selected items is generalizable to the tests as a whole, then these tests could have severely underestimated what children with lower reading and writing abilities know in science. Likewise, the efficacy of reform efforts to improve children's inquiry and reasoning abilities would be greatly underestimated if the sole measure for success is standardized test items like the ones we have incorporated. This issue becomes particularly relevant now that we are barely a year away from the NCLB's stipulation for every state to include science in mandatory testing. The question for the research community is – How do we make scalable science assessments that are not reading/writing tests by another name? The question for policy makers is – Until assessment technique improves, what do we learn by mandatory standardized testing in science?

The other challenge is actually implied by our "success". The instructional environment created in this study is not to be mistaken as a model for actual practice. In the real world, not every lesson can be taught through such intense collaboration of researchers and teachers. Our decision to teach experimental design rather than other aspects of science inquiry standards was based on the availability of substantial prior research. Most topics in

science (particularly those involving content knowledge) do not have such a rich history of *directly* relevant cognitive and instructional research. In the absence of a long history of relevant research and close researcher/teacher collaboration, it seemed tenuous to expect comparable results from classroom teachers.

Lastly, researchers who intend to conduct educationally relevant research, like us, need to methodologically and practically answer the question, "When is research relevant enough?" Can we be satisfied if we achieved a statistically significant learning gain in the laboratory (e.g., Chen & Klahr, 1999), or if we replicated a method from laboratory to classroom within a homogeneous high-SES population (e.g., Toth, Klahr, & Chen, 2000), or if we closed the achievement gap between heterogeneous groups (e.g., the present study)? We believe the envelope needs to be pushed continuously.

References

American Association for the Advancement of Science (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

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

Chen, Z. & Klahr, D. (1999). All other things being equal: Children's acquisition of the control of variables strategy. *Child Development, 70 (5),* 1098-1120.

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

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*. Mahwah, NJ : Erlbaum.

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

Klahr, D. & Li, J. (2005*).* Cognitive Research and Elementary Science Instruction: From the Laboratory, to the Classroom, and Back. *Journal of Science Education and Technology, 14-2*.

Klahr, D. & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15 (10).*

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

Kuhn, D., Garcia-Mila, M., Zohar, A. & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, 60* (4, Serial No. 245), 1-128

National Center for Education Statistics (n.d.). *The nation's report card (NAEP): 1996 assessment science public release grade 4 or 8.* Retrieved on September 16, 2004 from http://nces.ed.gov/nationsreportcard/itmrls/sampleq/96sci4(or 8).pdf

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

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

Rosnow, R. L. & 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.

Ross, A. J. (1988). Controlling variables: A meta-analysis of training studies. *Review of Educational Research*, *58*, 405-437 Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32,* 102-119.

Toth, E., Klahr, D. & Chen, Z. (2000). Bridging research and practice: A cognitively-based classroom intervention for teaching experimentation skills to elementary school children. *Cognition & Instruction, 18 (4),* 423-459.

Triona, L. M. & 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 & Instruction, 21,* 149-173.

Acknowledgement

This research is funded by the Cognition and Student Learning Program at the Institute of Education Sciences, U.S. Department of Education. We thank the students, principals, and teachers at both St. James School and Sacred Heart Elementary in Pittsburgh, PA for their assistance in this research. We thank our colleagues Stephanie Siler, Norma Chang, Elida Laski, Mari Strand Cary, and Audrey Russo for their feedback and help.