

Sequential Effects of High and Low Guidance on Children's Early Science Learning

Bryan J. Matlen & David Klahr
 Department of Psychology, Carnegie Mellon University
bmatten@cmu.edu, klahr@cmu.edu

Abstract: We describe a microgenetic approach aimed at examining the effect of different sequences of high vs low levels of instructional guidance on students' learning about concepts and procedures associated with simple experimental design – often called the "Control of Variables Strategy (CVS)". Third-grade children were randomly assigned to one of four conditions in which CVS was taught via one of the four possible orderings of high or low instructional guidance: (High + High, High + Low, Low + High, and Low + Low). High guidance consisted of a combination of direct instruction and inquiry questions, whereas low guidance was comprised only of inquiry questions. Contrary to commonly held beliefs that high levels of guidance, and in particular, direct instruction, lead only to shallow learning, results show that repeated instances of high instructional guidance were advantageous for both learning and transfer of CVS. Moreover, the High + High group continued to demonstrate strong conceptual understanding of CVS relative to other groups five months after training. Overall the study suggests that strong amounts of instructional guidance are capable of exhibiting powerful effects on children's early science learning.

Introduction

Effective instructional design is crucial to promoting learning and transfer. While great strides have been made within the field of instructional science over the past three decades, intense debate still exists over some fundamental issues (Kirschner, Sweller, & Clark, 2006; Klahr, 2009; Kuhn, 2007). For example, what, if any, amount of instructional guidance will maximize student learning? If there is an optimal level of guidance, where should it be placed during the course of instruction? The present study is aimed at addressing these fundamental issues facing the learning and instructional sciences in the context of elementary science education.

Of particular importance to science education is whether high levels of instructional guidance in the form of detailed, explicit instruction can produce robust learning (i.e. learning that transfers across extensive periods of time and to disparate tasks). Instruction at the extreme end of the explicitness spectrum – commonly known as "Direct Instruction" – has classically been construed as providing only short-term learning gains, leading to knowledge that is fragile and incapable of transferring to remote or "authentic" settings (Germann, Aram, & Burke, 1996; McDaniel & Schlager, 1990; Koedinger & Aleven, 2007; Kuhn & Dean, 2005). Direct instruction has also been criticized on the grounds that it allocates a passive role to learners, rather than promoting active student learning, whereas low levels of instructional guidance are more effective at inducing learners to construct knowledge on their own, thereby leading to more meaningful understanding (Kuhn & Dean, 2005). Indeed, Dean and Kuhn (2007) conclude from their investigation that, "... *direct instruction appears to be neither a necessary nor sufficient condition for robust acquisition or for maintenance over time.*" (p. 394).

However, these claims are inconsistent with the results of several of our prior investigations of the effectiveness of direct instruction, in which we have repeatedly found positive learning gains in situations where we provided students with a high degree of instructional guidance: direct instruction coupled with inquiry questions. This pedagogical approach has consistently produced meaningful and significant performance gains immediately after training (Chen & Klahr, 1999; Toth, Klahr, & Chen, 2000) that have been sustained for at least a period of three years (Strand-Cary & Klahr, 2009).

In this paper we address a more nuanced version of the question about the optimal level of instructional guidance. Instead of simply asking "which is best?" we also consider the possibility that there may be an optimal temporal sequence of different levels of guidance. As Koedinger and Aleven (2007) have demonstrated in the context of cognitive tutors, each end of the high-to-low instructional guidance spectrum has its unique strengths and weaknesses. Accordingly, rather than pit one type of instruction against another, several recent studies have reframed the question in terms of how different *sequences* of high and low guidance lead to different levels of student learning. However, this research has yielded some contradictory results. For example, Schwartz and Martin (2004) have suggested that high amounts of guidance are most effective after students attempt to invent solutions in minimally guided settings, an instructional method known as *Invention as Preparation for Future Learning* (IPL). In their study, students who invented original formulas for calculating variance benefited more from subsequent direct instruction than students who initially received instruction and then practiced applying equations of variance. The notion that students may initially fail to produce correct

solutions in minimally guided settings, but that this failure can serve as a ‘productive’ learning experience when followed by instructional support (i.e. termed *productive failures*), has also been reported by Kapur (2008; In press). These studies, however, seem in contrast to research that have observed an *Expertise Reversal* effect as learners acquire knowledge (Kalyuga, 2007). According to this latter perspective, since working memory is limited, novices need high levels of guidance in order to ameliorate the strong demands of cognitive processing required of them: As learners gain expertise, however, low guidance is thought to be most beneficial because it reduces the amount of redundant information presented. In support of this, Kalyuga and colleagues have provided empirical evidence that novices benefit more from viewing modeled solution steps and that, as they gain domain expertise, they learn more from engaging in unstructured practice problems (Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Kalyuga & Sweller, 2004). Overall, the implications these approaches suggest for sequencing instruction appear to be incongruous.

Given the stark disagreement in the literature surrounding not only instructional sequence, but also the utility of providing high guidance in the form of direct instruction, the present study has two inter-related goals. First, we test the claim that strongly guided instruction is insufficient for promoting robust understanding (Dean & Kuhn, 2007). To this end, we contrast the relative effectiveness of 1) High Guidance, in which students were provided inquiry questions and direct instruction, and 2) Low Guidance, in which we encouraged learning through the provision of inquiry questions, while purposefully withholding direct instruction. Our second goal is to examine the efficacy of varying sequences of these forms of instruction. Thus, while simple contrasts of high vs low guidance have been used in earlier work, in the present study, we contrast all possible orderings of these two instructional approaches across two separate training sessions, leading to four experimental conditions (i.e. High + High, High + Low, Low + High, and Low + Low). By contrasting these four conditions we are able to assess whether there indeed exist certain progressions of high and low guidance that are more effective in promoting student learning than multiple sessions of high or low guidance by themselves.

We address these issues in the context of teaching children about the Control of Variables Strategy (CVS) for scientific experimentation. CVS is the procedure used to create unconfounded experiments by keeping the values of all factors the same while changing only the variable of interest in order to determine whether or not that factor is causal with respect to the experimental outcome. The procedures and concepts associated with CVS have become increasingly influential in state standardized tests in elementary school and appear on almost all state standards as well as the National Science Standards for elementary school students (Klahr & Li, 2005; National Research Council, 1996). Despite CVS's importance, however, children throughout the elementary school years often demonstrate difficulty in applying it to relevant situations (Chen & Klahr, 1999; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995), making it an ideal domain – for both theoretical and practical reasons – in which to explore these research questions.

Method

Participants

Forty-two third grade children (22 girls, 20 boys, $M = 9.03$ years, $SD = .33$ years) from three middle-class Pittsburgh elementary schools participated in the study. Children were randomly assigned to one of the four experimental conditions.

Design

The overall design was microgenetic in nature (Siegler & Crowley, 1991), and consisted of four experimental conditions and 10 test phases. The first eight test phases were conducted approximately one week apart from each other, whereas the final two test phases occurred five months later. All phases were identical across each condition, with the exception of phases 3 and 4, where the experimental treatment varied by condition across two training sessions: (i) High Guidance followed by High Guidance (High + High), (ii) High Guidance followed by Low Guidance (High + Low), (iii) Low Guidance followed by High Guidance (Low + High), and (iv) Low Guidance followed by Low Guidance (Low + Low). Each test phase is described in detail, below.

Procedure

Phase 1: Story Pre-test

The first phase consisted of a paper and pencil pre-test with six questions aimed at assessing children's understanding of CVS. Three questions assessed children's ability to create an unconfounded experiment (design questions), and three asked them to evaluate whether an existing experiment was unconfounded (evaluate questions). Children completed the Story pre-tests at their desks in their regular classrooms and were given as much time as they needed to finish the test. Story responses were scored as 1 if children correctly identified or designed experiments consistent with CVS, and all other responses were assigned a 0. Scores ranged from 0 to 6.

Phase 2: Ramps Pre-test

Children were introduced to two physical ball and ramp apparatuses and were told that they would be designing experiments to test whether certain variables made a difference in how far the ball rolled. Children were told the four variables that could affect the outcome: the surface type (either fim or sif), ball type (bab or lof), steepness (steep or not steep) and starting gate (high or low) (Surface and ball type were given nonsense names to minimize children's expectations about their effect on the outcomes of their experiments). After children could identify these variables without the help of the experimenter, one of the variables was chosen at random and the child was asked to design an experiment to test whether that variable made a difference in the outcome. Children were allowed to set up an experiment and observed its outcome. No feedback was given. The procedure was repeated until all 4 of the variables were tested in a random order, and children were assigned a score of 1 if their set up varied the target variable and kept all other variables the same (CVS), and a score of 0 if the set up was of any other combination. Scores ranged from 0 to 4.

Phase 3: Training 1

The same materials as the Ramps Pre-test (Phase 2) were used. At the start of Training 1, the child was shown the ball and ramp apparatus and asked to identify the four variables that might make a difference in how far the ball rolled. Once children had correctly identified the four variables, the child's condition determined what happened next.

(a) Children who experienced Low Guidance in Training 1 (i.e., those in the Low + High and Low + Low conditions) engaged in minimally guided discovery learning while being provided inquiry questions in order to scaffold their learning. Children were told that they would be setting up experiments to see what made a difference in how far the ball rolled down the ramps and, subsequently, they were asked to set up an experiment to test one of the four variables, chosen at random. Children were also asked two scaffolding questions to prompt their self-explanations and (minimally) guide them in the discovery process. The first question occurred after children set up their experiments, but before they viewed the experimental outcome. The experimenter asked the child why he/she had set up the experiment that way. The experimenter listened to the child's explanations and provided neutral feedback such as, "okay," or "alright" and then asked the child to roll the balls down the ramp. Upon observing the result, the child was asked whether they could tell for sure from the experiment whether the target variable made a difference in the outcome. This procedure was repeated for all four variables in a random order, and the child designed two experiments to test each variable, resulting in eight child-designed experiments.

(b) Children who experienced High Guidance in Training 1 (i.e., those in the High + High and High + Low conditions), were told that they would be evaluating experiments conducted by the experimenter and that they would be asked to evaluate the experiment on whether it was a 'smart' or 'not smart' way to test the target variable. The child observed while the experimenter set up a series of confounded and unconfounded experiments (two of each kind), and the child was asked whether the experiment was a 'smart' or 'not smart' way to test the target variable. Regardless of the child's response, the Experimenter then explained why each experiment was either smart or not smart, and provided the logical basis behind CVS. The child was then asked whether they could tell for sure from the experiment whether the target variable made a difference in how far the ball rolled, and were provided immediate feedback followed by an explanation about why one could or could not tell for sure whether the target variable was causal.

At the end of the Training 1 phase, a post-test was administered to children in all conditions. Children were asked to design an experiment to test each target variable in a procedure identical to the Ramps Pre-test. Scores ranged from 0 to 4.

Phase 4: Training 2

This phase followed the identical procedures to the first training phase, the only difference being that children in the Low + High conditions now received a high degree of instructional guidance with the ramps apparatus and the children in the High + Low conditions now received low levels of instructional guidance. A post-test at the end Training 2 was scored identically to the post-test at the end of Training 1. The differences between the High and Low Guidance conditions are summarized in Table 1.

The remaining test phases (phases 5-10) consisted of a series of transfer tests that were meant to assess the robustness of children's ability to apply CVS to novel situations. In all transfer phases, no feedback was ever provided to children and the dependent variable was always how many experimental set ups children designed that were consistent with the logic of CVS.

Table 1: Similarities and differences between the two types of instruction

Amount of Instructional Guidance		
	High	Low
Goal Setting	“I’m going to design an experiment to test whether X makes a difference in how far the ball rolls.”	“Can you design an experiment to test whether X makes a difference in how far the ball rolls?”
Number of Experiments	4	8
Design of Experiments	By Experimenter	By Child
Inquiry Questions	“Is this a smart or not smart experiment?” “Can we tell for sure from this experiment whether X made a difference?”	“Why did you set up your experiment that way?” “Can we tell for sure from this experiment whether X made a difference?”
Explanations	Experimenter explained why an experiment was smart or not smart, and why the child could or could not tell for sure whether X made a difference in the outcome	No Explanations

Phase 5: Springs Transfer Test

The materials for this phase consisted of a set of eight springs that varied across three different dimensions: length (long or short), spring width (wide or narrow), and coil width (thick or thin). A wooden centerpiece that children could hang the springs from was displayed on a table and two weights were present that could be attached to the springs in order to observe their stretching. Children were asked to design an experiment to test whether each of the variables (in a random sequence) made a difference in how far the spring stretched. Children tested all three variables and the child was assigned a 1 for CVS setups and a 0 for all other combinations. Scores ranged from 0 to 3.

Phase 6: Car Design Transfer Test

A computer program (adapted from Klahr, Triona, & Williams, 2007) presented on a laptop computer allowed children to design cars to test what variables affected how far they traveled. Four variables – the body (long or short), the back axle (thick or thin), the front wheels (large thick, large thin, or small), and the back wheels (large thick, large thin, or small) – could affect how far the cars traveled. Children were told that they would be testing cars to figure out what made them travel farther, and that the experimenter would record the cars they made on a notepad to which they could refer back to if they forgot what cars they had constructed. Children proceeded to test each variable, which was indicated by the experimenter in a fixed sequence. In total, children had the opportunity to design 10 cars and the total possible chances of demonstrating CVS was 6. Therefore, scores ranged from 0 to 6.

Phase 7: Ramps Transfer Test

The materials, procedure, and scoring in this phase were identical to the Ramps Pre-test.

Phase 8: Story Transfer Test

The materials, procedure, and scoring in this phase were identical to the Story Pre-test.

Phase 9: Remote Paper and Pencil Test

This phase took place roughly 5 months from training, when children were in the 4th grade. Children sat at their regular classroom desks and completed a nine-question paper and pencil test that showed various experimental comparisons. Children were asked to identify which experiments were good and bad comparisons, and if the child identified an experiment as a bad test, they were asked to change the experiment to make it a good test. Three of the nine questions were good tests (CVS), whereas six questions were bad tests. Children were given a score of 1 if they correctly identified the test as either good or bad, and a score of 1 if they correctly changed the experiment. Scores ranged from 0 to 15.

Phase 10: Remote Ramps Transfer test.

The materials, procedure, and scoring in this phase were identical to the Ramps Pre-test and Post-test, but took place 5 months after training.

Results

Means and standard errors of the mean for each test phase by condition are displayed in Figure 1. A 4 (training condition) x 8 (test phase) mixed ANOVA on the first 8 test phases revealed a main effect of condition ($F(3, 38)$

= 6.21, $p < .005$), a main effect of test phase ($F(7, 266) = 43.64$, $p < .001$), and a significant interaction between them ($F(21, 266) = 2.05$, $p = .005$) (for the remainder of this report, statistics at each test phase will be reported for each group in the following order: 1) High + High, 2) High + Low, 3) Low + High, and 4) Low + Low) (story pre-test: $M1 = .23$, $SE1 = .06$, $M2 = .20$, $SE2 = .04$, $M3 = .28$, $SE3 = .05$, $M4 = .18$, $SE4 = .03$. For the ramps pre-test: $M1 = .25$, $SE1 = .08$, $M2 = .20$, $SE2 = .08$, $M3 = .10$, $SE3 = .06$, $M4 = .16$, $SE4 = .08$).

Post-hoc analyses (1) revealed no differences between any of the groups at the story pre-test (all p 's > .40) or the ramps pre-test (all p 's > .54). At the end of Training 1, however, the mean scores of the two groups receiving high levels of guidance improved dramatically from the ramps pre-test (all p 's < .001) while the groups receiving low amounts of guidance (i.e. Low + High and Low + Low) did not evidence reliable improvements (all p 's > .32). The groups receiving high guidance also performed at significantly superior levels when compared to the low guidance groups (all p 's < .01) (Training 1: $M1 = .98$, $SE1 = .03$, $M2 = .98$, $SE2 = .02$, $M3 = .50$, $SE3 = .14$, $M4 = .48$, $SE4 = .13$).

After Training 2, the Low + Low and the Low + High groups were no different from their performance after Training 1 or from their performance during the Ramps Pre-test (all p 's > .16), while the High + High group and the High + Low group continued to perform at ceiling levels (all one-tailed p 's > .21). Performance of both the High + High group and the High + Low group was significantly superior to the Low + Low group (all p 's < .05) (Training 2: $M1 = 1$, $SE1 = .00$, $M2 = .98$, $SE2 = .02$, $M3 = .73$, $SE3 = .14$, $M4 = .57$, $SE4 = .14$).

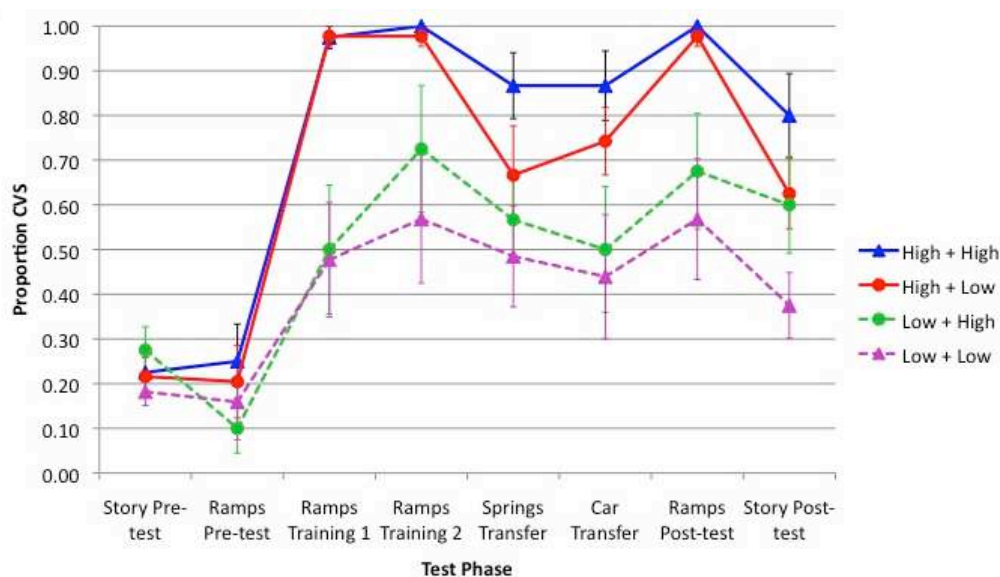


Figure 1. Means for each condition by the first 8 test phases: Error bars represent standard errors of the mean.

Of particular interest is the performance of children in the Low + High condition at the end of Training 2. Even though these children had just completed a training phase involving direct instruction, their performance was well below the ceiling levels of the High + Low group (one-tailed $p < .05$). To further examine the differences in learning gains between the High + Low and Low + High groups, a 2 (condition: High + Low vs Low + High) \times 2 (phase: Ramps Pre-test vs Training 2) mixed ANOVA was conducted and this resulted in a significant interaction $F(1, 19) = 7.35$, $p = .01$, with the learning gains greater for the High + Low group. Thus, even though both groups were provided high *and* low amounts of instructional guidance during training, the order of these two types of instruction had substantial effects on learning, with the learning gains favoring the students in the High + Low group. However, since this finding might be due to low sample size, we would expect the effects to persist across a variety of transfer tests. This hypothesis was thus put to the test in the remaining transfer phases of the study.

The Springs Test was the first opportunity to demonstrate this transfer. Post-hoc analyses for this phase revealed statistically significant differences between the High + High and the Low + Low groups ($p < .05$) favoring the High + High group. All other comparisons were not statistically significant (all p 's > .17) (springs transfer: $M1 = .87$, $SE1 = .07$, $M2 = .67$, $SE2 = .11$, $M3 = .57$, $SE3 = .09$, $M4 = .48$, $SE4 = .11$). At the car design transfer test, significant differences were found between the High + High and the Low + Low groups ($p = .05$) with no other comparisons showing reliable differences (all p 's > .13) (car design transfer: $M1 = .87$, $SE1 = .08$, $M2 = .74$, $SE2 = .08$, $M3 = .50$, $SE3 = .14$, $M4 = .44$, $SE4 = .14$).

The High + High and Low + Low groups showed a significant difference at the Ramps Post-test ($p < .05$), as well as the High + Low and the Low + Low groups ($p < .05$). A marginally significant difference was found between the High + High and the Low + High groups ($p = .10$), with no other comparisons evidencing

significance. Of interest was whether children performed at levels greater than Ramps Pre-test: here we found that all groups except the Low + Low groups showed a reliable improvement from the Ramps Pre-test to the Ramps Post-test (all p 's < .05) (ramps post-test: $M1 = 1$, $SE1 = .00$, $M2 = .98$, $SE2 = .02$, $M3 = .68$, $SE3 = .13$, $M4 = .57$, $SE1 = .14$).

The only reliable difference at phase 8 (Story Post-test) was between the High + High and the Low + Low groups ($p < .05$). No other comparisons were significant (all p 's > .32). We also compared children's post-test performance to their pre-test performance on the story problems and found that the High + High group evidenced a statistically significant improvement in performance ($p = .05$), the High + Low group showed a marginally significant improvement ($p = .07$), and the Low + High and Low + Low groups did not evidence any reliable improvements (all p 's > .16) (story post-test: $M1 = .80$, $SE1 = .09$, $M2 = .63$, $SE2 = .08$, $M3 = .60$, $SE3 = .11$, $M4 = .38$, $SE4 = .07$).

Five months after training, we assessed children's ability to apply CVS. Thirty-nine of 42 children were retained for these phases. ANOVA's for the remote paper and pencil tests evidenced no significant differences at either test phase (all p 's > .13) (5 month paper and pencil test: $M1 = .91$, $SE1 = .06$, $M2 = .66$, $SE2 = .09$, $M3 = .65$, $SE3 = .10$, $M4 = .61$, $SE4 = .12$. For the 5 month ramps test: $M1 = .97$, $SE1 = .03$, $M2 = .90$, $SE2 = .06$, $M3 = .77$, $SE3 = .13$, $M4 = .64$, $SE4 = .15$). However, because the null differences could be due to low sample sizes, we categorized students into CVS experts and non-experts. Experts on the remote paper and pencil test were defined as children who scored at least 13 of 15, and on the ramps post-test they were defined as children who designed at least 3 of 4 experiments as CVS. Distributions of experts and non-experts for each test are displayed in Figure 2. Chi-squared tests of independence revealed marginally significant differences in the distribution of experts at both the paper and pencil ($\chi^2(3, 39) = 7.12$, $p = .07$) and ramps tests ($\chi^2(3, 39) = 6.94$, $p = .07$). At the paper and pencil test, there was a greater proportion of experts in the High + High group compared to all other groups (all p 's < .05), while at the ramps test, all groups had greater proportions of experts than the Low + Low group (all p 's < .05).

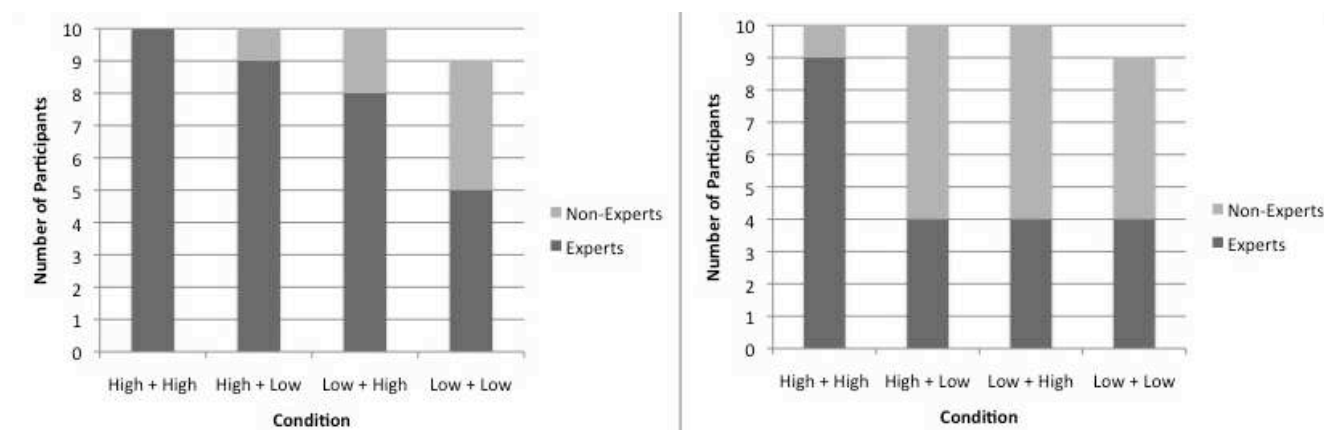


Figure 2. Left) Number of participants who were experts on the 5-month ramps transfer test by condition. Right) Number of participants who were experts at the 5-month paper and pencil transfer test by condition.

Discussion

The most consistent finding of the present study was the superior performance of the High + High group when compared to the Low + Low group at every test phase after training. This suggests that the incorporation of direct instruction can be a powerful strategy that, coupled with inquiry teaching, is capable of promoting robust learning and transfer of scientific experimentation skill in children. Our findings also extend prior research on children's CVS learning, which has shown that a) minimally guided instruction often produces only nominal learning gains in children (Chen & Klahr, 1999; Klahr & Nigam, 2004) and b) the benefits of high amounts of guidance are not only immediate, but often continue to persist even after delayed periods of time (Chen & Klahr, 2004; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2009; Klahr, Triona, & Williams, 2007).

A second goal of the present study was to examine what order of instructional guidance might prove most beneficial to student learning. While students' performance immediately after training suggested that the High + Low group may have benefited more from instruction than the Low + High group, we failed to find any differences between these groups at later points in transfer, suggesting that this initial effect after training may have been unique to that point in assessment. Moreover, in a subsequent follow-up study aimed to assess whether the differences between the High + Low and Low + High group would replicate, we divided a group of 30 third-graders ($M = 8.76$ years, $SD = .27$ years) evenly between the High + Low and Low + High groups and found no differences in learning across three separate transfer tests (averaged across all three tests, means were .69 and .63 for the High + Low and Low + High groups, respectively) independent-samples $t(28) = .15$, ns .

However, this follow-up study did replicate the finding that when students received high amounts of instructional guidance with ramps, they improved significantly from a ramps pre-test (for both groups, p 's < .05), while the Low + High group did not improve from the ramps pre-test after receiving low instructional guidance in the ramps domain ($p = .35$) (means for students ramps pre-test, ramps training 1 post-test, and ramps training 2 post-test were .14, .66, and .82 for the High + Low group, and .09, .13, and .72 for the Low + High group). Thus, it appears that, in the context of experimental design, young children learn just as well from different sequences of instructional guidance, as long as they are provided with high amounts of guidance at some point during learning.

The five-month transfer tests in the present study allowed us to assess how flexibly children could apply their CVS knowledge after extensive periods of time in 1) an identical domain to which they were trained (i.e. ramps), and 2) a domain that was removed from their original training (i.e. paper and pencil tests). Here, a large proportion of children in the High + High group evidenced flexible application of CVS to the paper and pencil post-test, whereas, only about half of the children demonstrated this proficiency in the other three groups. This result is striking when compared to the ramps transfer test, where a majority of children who had received direct instruction continued to perform at ceiling levels, and only about half of the children in the Low + Low group demonstrated this proficiency. It seems that, when transfer pertains only to time, children who receive high instructional guidance perform well on the materials on which they were trained, even after 5 months, and more so than children who do not receive high guidance. However, when transfer requires the application of knowledge over remote periods of both time *and* task, only students who received extensive amounts of direct instruction consistently applied CVS to novel situations beyond that of their training (i.e. the High + High group). Thus, results from this 5-month post-test strongly underscore the powerful effect that high levels of guidance can have on student learning.

Taken together, the present findings support the conclusion that students who are relative domain novices do not benefit more from any particular sequence of instruction, but that they do benefit – especially in the long run – from multiple lessons in which high amounts of guidance are provided. While this conclusion may, on the surface, seem conflicting with IPL and the concept of productive failures (2), it is important to note that studies reporting instructional sequence effects have often included other elements that seem key to acquiring differentiated knowledge, such as the use of contrasting cases, which we have left out of our instruction (Schwartz & Bransford, 1998). Students are also older in these studies, the youngest being 8th grade students, whereas we are investigating learning in 3rd grade students: There is little doubt that younger children differ from older children in many ways, including their working memory capacity, domain knowledge, and recruitment of metacognitive strategies. The disparity between our findings, however, prompts discussion of the circumstances that may be crucial to IPL and productive failures. One fundamental prerequisite for a failure to serve as a productive learning experience seems to be that the learner must first realize that failed solutions are indeed, failures. Without such knowledge, students may continue to practice and strengthen the use of incorrect strategies, therefore failing to acknowledge the need for more efficient solutions offered through instruction. Alternatively, if the student is aware that his/her strategies are suboptimal, the student may be more prone to recognize the value of instruction. Indeed, students in Schwartz and Martin (2004) could readily observe whether their formulas for variance accounted for the data they were trying to describe. Kapur (in press) also showed that students in less structured settings often exhibited lower confidence in their solutions, suggesting they were aware of the shortcomings of these solutions. In contrast, many of the students who designed experiments in our study gave little indication that they were aware that their strategies were incorrect, frequently responding that they *could* tell for sure whether the target variable made a difference in the outcome, even when their experiment was confounded by multiple factors. While the interactions between characteristics of individual learners, the learning domain, and instructional support are no doubt a complex issue and beyond the scope of the present study, we propose that attempting to understand the relative contributions of each of these factors, as well as how they interact, would be a productive area for future research to move towards.

What this study does contribute to our thinking about instructional design is that it refutes widely held beliefs that direct instruction is insufficient for promoting robust learning (Dean & Kuhn, 2007) and that shallow transfer is one of the hallmarks of this form of instruction. On the contrary, the present study suggests that, for domain novices, minimally guided instruction that neglects direct instruction may miss opportunities to optimize student learning. We recognize that over the course of a students' life, he/she may experience a great number of instructional events, and that this study addresses only the very earliest of these instructional experiences. We thus agree with, and emphasize that a combination of various instructional approaches may indeed prove of large value to a learner throughout the course of their education (Koedinger & Aleven, 2007). At least in early stages of learning, however, the present study suggests that providing high amounts of guidance, particularly in the form of multiple phases of direct instruction, can be a useful strategy in promoting robust understanding of scientific experimentation skill.

Endnotes

- (1) All between subject post-hoc tests are Tukey adjusted, and within subject post-hoc tests are Bonferonni adjusted.
(2) We purposefully exclude discussion of the Expertise Reversal effect here, as this view could explain the present findings by positing that, despite receiving training, children were relative domain novices at this stage in learning.

References

- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, 70, 1098-1120.
- Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, 91, 384-397.
- Germann, P.J., Aram, R., & Burke, G. (1996). Identifying patterns and relationships among the responses of seventh-grade students to the science process skill of designing experiments. *Journal of Research in Science Teaching*, 33, 79-99.
- Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19, 509-539.
- Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology*, 93, 579-588.
- Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology*, 96, 558-568.
- Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, 26, 379-424.
- Kapur, M. (In press). Productive failure in mathematical problem solving. *Instructional Science*
- Kirschner, P.A., Sweller, J., & Clark, R.E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41, 75-86.
- Klahr, D. (2009). "To every thing there is a season, and a time to every purpose under the heavens": What about Direct Instruction? In S. Tobias and T. M. Duffy (Eds.) *Constructivist Theory Applied to Instruction: Success or Failure?* Taylor and Francis.
- 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*, 4, 217-238.
- Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, 15, 661-667.
- Klahr, D., Triona, L.M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science Teaching*,
- Koedinger, K.R. & Aleven, V. (2007). Addressing the Assistance Dilemma in Experiments with Cognitive Tutors. *Educational Psychology Review*, 19, 239-264.
- Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about controlling variables? *Psychological Science*, 16, 866-870.
- Kuhn, D. (2007). Is direct instruction an answer to the right question? *Educational Psychologist*, 42, 109-113.
- Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development*
- McDaniel, M.A., & Schlager, M.S. (1990). Discovery learning and transfer of problem-solving skills. *Cognition and Instruction*, 1990, 7, 129-159.
- National Research Council (1996). The National Science Education Standards. Washington DC: National Academy Press.
- Schwartz, D.L., & Bransford, J. (2006). A Time for Telling. *Cognition and Instruction*, 16, 475-522.
- Schwartz, D.L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22, 129-184.
- Strand-Cary, M., & Klahr, D. (2009). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development*, 23, 488-511.
- Siegler, R., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, 46, 606 - 620.
- 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 and Instruction*, 18, 423-459.

Acknowledgments

We thank Howard Seltman, Anna Fisher, Marsha Lovett, Bob Siegler, Stephanie Siler, Jamie Jirout, Kevin Willows, Audrey Russo and Cressida Maguro for their various intellectual contributions. This work was supported by the Institute of Education Sciences under grant #R305B040063 to Carnegie Mellon University.