Instr Sci
DOI 10.1007/s11251-012-9248-z

Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: Is it all in the timing?

Bryan J. Matlen · David Klahr

Received: 11 January 2012 / Accepted: 18 May 2012

Abstract We report the effect of different sequences of high vs low levels of instructional guidance on children's immediate learning and long-term transfer of simple experimental design procedures and concepts, often called "CVS" (Control of Variables Strategy). Third-grade children ($N = 57$) received instruction in CVS via one of four possible orderings of high or low instructional guidance: high followed by high (HH), high followed by low (HL), low followed by high (LH), and low followed by low (LL). High guidance instruction consisted of a combination of direct instruction and inquiry questions, and low guidance included only inquiry questions. Contrary to the frequent claim that a high degree of instructional guidance leads to shallow learning and transfer, across a number of assessments—including a 5-month post-test—the HH group demonstrated a stronger understanding of CVS than the LL group. Moreover, we found no advantage for preceding high guidance with low guidance. We discuss our findings in relation to perspectives advocating "invention as preparation for future learning", and the efficacy of "productive failure".

Keywords Instruction · Inquiry · Science Education · Experimentation skills · Learning · Transfer

Introduction

Although great strides have been made within the field of instructional science over the past three decades, intense debate still exists over two fundamental issues: (a) What amount of instructional guidance will maximize learning? (Kirschner et al. 2006; Klahr 2009; Koedinger and Aleven 2007; Kuhn 2007; Taber 2010; Tobias and Duffy 2009) (b) If there is an optimal level of guidance, where should it be placed during the course of instruction and assessment? (Kalyuga 2007; Kalyuga et al. 2001; Kapur 2008; Schwartz

B. J. Matlen (✉) · D. Klahr
Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave., 455B Baker Hall, Pittsburgh, PA 15213, USA
e-mail: bmatlen@cmu.edu

Published online: 04 June 2012

and Bransford 1998; Schwartz and Martin 2004). The present study addresses these questions in the context of teaching simple experimental design to third graders.

The first issue is typically framed in terms of the relative effectiveness of different amounts of "explicitness" or "directness" of the instruction. Instruction at the "Direct Instruction" end of the continuum has been, on the one hand, advocated as producing rapid and significant learning and transfer (Chen and Klahr 1999) while on the other hand, critiqued for producing only short-term learning of fragile knowledge that is unlikely to transfer to remote or "authentic" settings (Chinn and Malhotra 2001; Germann et al. 1996; Kuhn and Dean 2005; McDaniel and Schlager 1990). Such critiques of Direct Instruction argue that it assigns a passive role to learners, whereas low levels of instructional guidance facilitate learners' active construction of knowledge, which in turn leads to meaningful and long-term retention and transfer (Dean and Kuhn 2007; Schwartz et al. 2011). However, these critiques of direct instruction are challenged by the results of several training studies that provide students with a high degree of instructional guidance: direct, explicit, teacher-controlled instruction coupled with inquiry questions. These studies have consistently produced meaningful performance gains immediately after training as well as near, medium, far, and remote transfer up to three years after training (Chen and Klahr 1999; Lorch et al. 2010; Strand-Cary and Klahr 2009).

The second of the two issues extends the first by asking about the effects of different *sequential orderings* of high and low guidance (Kalyuga 2007; Koedinger and Aleven 2007; Schwartz and Martin 2004). Some studies suggest that delaying high levels of instructional guidance improves learning. For example, Schwartz and Martin (2004) found that 9th grade students who—prior to explicit instruction—were challenged to invent formulas for calculating variance benefited more from subsequent explicit instruction than students who received instruction first and then practiced applying variance equations. Similar effects were reported by Kapur (2008) in his investigation of students' "productive failures", in which high-school students who initially failed at solving ill-structured physics problems showed better near and far transfer on related tasks than did students who first attempted to solve well-structured problems. The theoretical rationale for asking students to invent solutions before receiving high guidance is that the approach induces students to struggle with various aspects of challenging problems. Although students' attempts to invent solutions will usually be unsuccessful (Kapur 2008), the process will familiarize them with some essential elements, constraints, and partial solutions. Consequently, when students ultimately *do* receive higher guidance, the underlying procedural and conceptual understanding will be salient, meaningful, better encoded, and mastered. Impressive learning gains have been reported in a number of students that use such "invention" procedures (Kapur 2008, 2009; Schwartz and Martin 2004; Schwartz et al. 2011).

However, some theoretical views suggest that rather than delaying a high level of guidance, providing it from the outset will optimize student learning. For example, Cognitive Load Theory (Sweller 1988) proposes that working memory limitations dictate high levels of instructional guidance initially for domain novices, but that such guidance becomes redundant, and even dysfunctional, as learners acquire expertise (Kalyuga 2007; Kalyuga and Sweller 2004). Kalyuga and colleagues report an "expertise reversal effect" whereby novices benefit more from viewing detailed examples of solution steps, and as they gain domain expertise, they learn more from engaging in unstructured practice problems (Kalyuga et al. 2001; Kalyuga and Sweller 2004). This basic effect has been documented in a number of different domains spanning both math and science, as well as in diverse populations including both high-schoolers and adults (see Kalyuga 2007 for review).

Given the stark disagreement in the literature surrounding not only the utility of providing high guidance in the form of direct instruction, but also the effectiveness of different sequences of high and low instructional guidance, the present study has two inter-related goals. First, we test the claim that "direct instruction appears to be neither a necessary nor sufficient condition for robust acquisition or for maintenance over time." (Dean and Kuhn 2007, p. 385) by contrasting the relative effectiveness—in both the short and long term—of (1) high guidance, in which students receive both inquiry questions and direct instruction, and (2) Low Guidance, in which they receive inquiry questions, but no direct instruction. The second goal is to examine the effect on learning and transfer (both near and far) of the four possible *sequences* of High (H) and Low (L) levels of instructional guidance: (HH), (HL), (LH), and (LL). If there are benefits to delaying guidance, as suggested by approaches advocating "preparation for future learning" and "productive failures", we expected the children in the LH condition to outperform others on both near and far transfer. However, if children benefit from receiving high guidance initially—as predicted by Cognitive Load Theory—we expected either of the HH or HL conditions to outperform other groups.

Method

We address these issues in the context of teaching third-grade children about the Control of Variables Strategy (CVS) for scientific experimentation. CVS is the procedure used to create unconfounded experiments by changing only the variable of interest while keeping the values of all other factors the same in order to determine whether or not that factor is causal with respect to the experimental outcome. The procedures and concepts associated with CVS are invariably included in high stakes assessments for middle school science (Klahr and Li 2005; National Research Council 1996). However, elementary school children's difficulties in understanding and applying CVS have been demonstrated repeatedly over the years (Chen and Klahr 1999; Kuhn et al. 1988, 1995; Zimmerman 2007), making it a highly relevant domain in which to investigate these research questions.

Participants

Fifty-seven third grade children (27 girls, 30 boys, $M = 9.12$ years, $SD = .37$ years) from two middle-class Pittsburgh elementary schools participated in the study.[1] Children were randomly assigned to one of the four experimental conditions.

Design

The overall design consisted of four experimental conditions and 10 test phases. The first eight test phases were conducted approximately weekly, and the final two test phases occurred 5 months later (see Table 1). The four experimental conditions differed only in the sequencing of the two levels of instructional guidance (High or Low) that were provided during the training sessions as described in Phases 3 and 4 below: High followed by High, High followed by Low, Low followed by High, and Low followed by Low. All other phases were identical across conditions.

[1] Five children dropped out of the study between the early and later phases, leaving 52 children who were included in the entire study: Ns in each group were HH = 14, HL = 11, LH = 14, LL = 13.

Table 1 Study design

Phase	1	2	3		4		5	6	7	8	9	10
Time[1]	*Week 1*	*Week 2*	*Week 3*		*Week 4*		*Week 5*	*Week 6*	*Week 7*	*Week 8*	*Week 24*	*Week 24*
	Pre-tests		**Training 1**	Immediate Assessment 1	**Training 2**	Immediate Assessment 2	Near Transfer Assessments				Remote transfer Assessments	
HH Condition	Story Problem pre-test	Ramps pre-test	**Ramps: High Guidance**	Ramps post-test	**Ramps: High Guidance**	Ramps post-test	Springs post-test	Car Design post-test	Ramps post-test	Story Problem post-test	Remote Story Problem post-test	Remote Ramps post-test
HL Condition			**Ramps: High Guidance**		**Ramps: Low Guidance**							
LH Condition			**Ramps: Low Guidance**		**Ramps: High Guidance**							
LL Condition			**Ramps: Low Guidance**		**Ramps: Low Guidance**							

Times are approximate: actual procedures occurred within ± a few days of the listed time in the table

Procedure

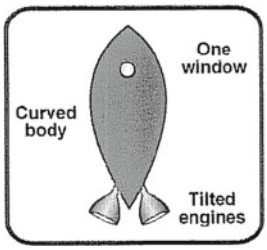
Phase 1: Story Problem Pre-test

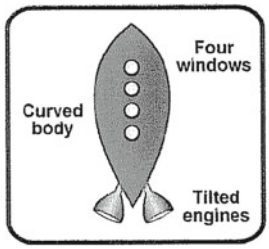
This paper and pencil test consisted of a series of six simple scenarios depicting experimental contrasts in domains that included baking cookies, selling beverages, and designing rockets (see Fig. 1). Three questions asked children to design an unconfounded experiment and three questions asked them to judge whether or not an experiment was confounded, and if so, to correct it. (Two of these three experiments were confounded.) Children completed Story Pre-tests at their own pace in their regular classrooms and were given as much time as needed. Story responses received one point if children correctly identified or designed experiments consistent with CVS, and an additional point if they correctly modified incorrect experiments. All other responses were assigned a 0. Scores ranged from 0 to 8.

Phase 2: Ramps Pre-test

During Phase 2—and also in phases 3–7, and 10—children were tested individually. In Phase 2, children were introduced to two physical ball and ramp apparatuses—of the type used by Klahr and Nigam (2004). They were told that they would be designing experiments to determine whether certain variables made a difference in how far the ball rolled, and that the outcome *might* be affected by: steepness (steep or not steep), position of the starting gate (high or low), the surface type ("fim" or "sif"), and the ball type ("bab" or "lof"; see Fig. 2). Surface and ball type were given non-sense names to minimize children's expectations about their effects. After children indicated that they could reliably identify these variables, one variable was chosen at random and the child was asked to design an experiment to test whether or not that variable made a difference in the outcome. Children were allowed to set up an experiment and observe its outcome. This procedure was repeated until all four variables had been tested in a random order. A score of 1 was assigned if children designed a "good" experiment (i.e., varied the target variable and kept all other variables the same), and a 0 otherwise. Scores ranged from 0 to 4.

The two pictures show an experiment to figure out whether or not the number of windows makes a difference in how high the rockets fly. Look carefully at the pictures. Each rocket has a certain body shape (Curved or Straight), number of windows (one or Four), and engine direction (Down or Tilted).





a) Do you think this is a good or bad way to find out whether the number of windows (One or Four) makes a difference in how high the rockets fly?

Good Way

Bad Way

b) Explain why you think this is a Good/Bad way: _____

_____

c) If you said it was a "Bad way": Change the picture(s) above to make it a Good Experiment. (For example, you might want to change the body shape, the number of windows, or engine direction for one of both of the rockets.)

Fig. 1 Example of a Story Problem pretest (Phase 1) and Story Problem Transfer Tests (Phase 8). Six problems, from three different scenarios (baking cookies, selling beverages, and designing rockets) were used

Phase 3: Training and Immediate Assessment 1

The between subjects contrast in the type of training—that is the level of instructional guidance—took place in Phases 3 and 4 (see Table 1). These two levels were similar to that of previous work (e.g., Klahr and Nigam 2004; Strand-Cary and Klahr 2009). The same materials that had been used in the Ramps Pre-test (Phase 2) were used in the two training phases. At the start of Training 1, the child was shown the ball and ramp apparatus and asked to identify the four variables (height, length, surface, and type of ball) that might make a difference in how far the ball rolled. Once children demonstrated that they could correctly identify the four variables, the child's condition determined what happened next. The differences between the High and Low Guidance conditions are summarized in Table 2 and described in the next two paragraphs.

(a) (a) *Low guidance*. In Training 1, children in the LH and LL conditions engaged in minimally guided discovery learning accompanied by inquiry questions. They were told that they would be setting up experiments to see what made a difference in how far the balls 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 questions. The first question occurred after children set up their experiments, but before they viewed the experimental outcome. The experimenter asked the children why they had set up the experiment that way. The experimenter listened to the explanation and provided neutral feedback such as, "*okay*," or "*alright*" and then asked the child to run the experiment by letting the balls roll down the ramps. Upon observing the result, the child was asked whether he/she could tell for sure from the experiment whether the target variable made a difference in the outcome (see Table 2 for the exact wording of experimental questions). This procedure was repeated for all four variables

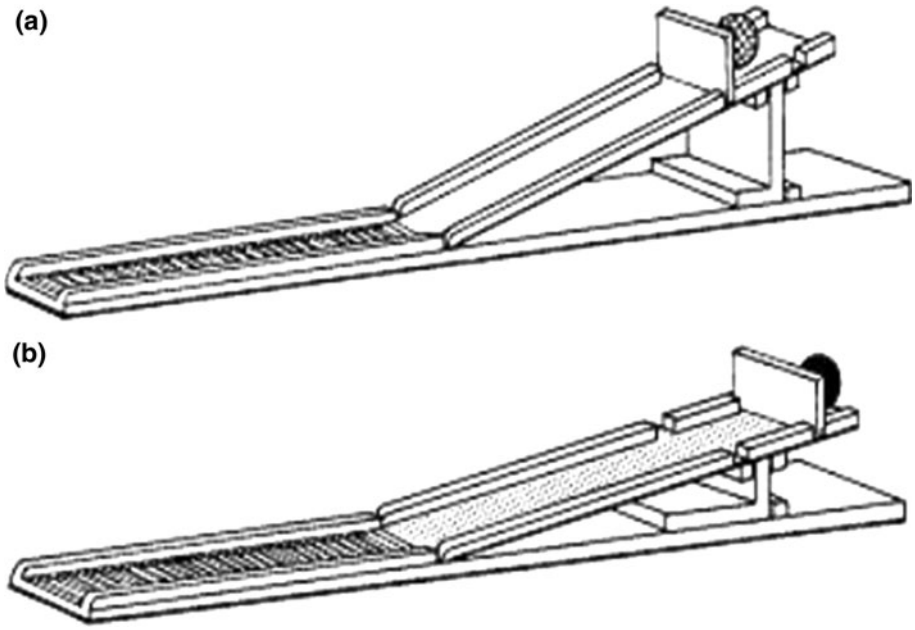


Fig. 2 Ramps apparatus used in Phases 2, 3, 4, 7 and 10. This particular illustration shows a "bad" experiment, because it is completely confounded: one ramp is high, and the other is low, one has a long run, the other a short run, one uses a "sif" surface and a "bab" ball type and the other uses a "fim" surface and a "lof" ball. If there were differences in how far the *balls* rolled, it would be impossible to identify the causal factor

in a random order, and the child had the opportunity to design up to two experiments to test each variable. In total, each child designed eight experiments.

(b) (b) *High guidance*. In the HH and HL conditions in Training 1, children were told that they would be setting up experiments—with the help of the experimenter to see what made a difference in how far the balls rolled down the ramps. The Experimenter and the child proceeded to set up four different experiments in a fixed order (multiply confounded, unconfounded, singly confounded, and unconfounded). At the outset, the child observed while the experimenter set up the multiply confounded experiment and was asked whether it was a "smart" or "not smart" way to test the target variable. Regardless of the child's response, the Experimenter explained why the experiment was not smart and explained the logical basis behind CVS. The child was also asked "Can you tell for sure from this experiment whether the target variable made a difference in how far the ball rolled?", and was provided immediate feedback followed by an explanation about why one could or could not "tell for sure" whether the target variable was causal. The child and the experimenter then proceeded to set up a "smart experiment" that changed the original multiply confounded experiment to an unconfounded one. Children were again asked to explain whether this set-up was smart, and why they could or could not tell for sure whether the target variable made a difference in the outcome and children were given immediate feedback on their responses. The Experimenter then repeated the process with a confounded and then unconfounded set up with a randomly selected variable (e.g., surface type). After the four experiments had been presented and explained by the experimenter, the logic

Table 2 Similarities and differences between the two types of instruction

	Amount of instructional guidance	
	High	Low
Experiments set up	By experimenter and child	By child
Number of experiments	4 (two CVS and two confounded)	Up to 8 (of any type)
Example 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
Summary	Experimenter summarized the logic of CVS	No summary

of CVS was summarized.[2] A central component of the experimenter's explanation about why some experiments were "smart" and others were not was telling the children that the only way to "tell for sure" if a factor was causal with respect to the outcome was to ensure that that factor was the only difference between the two set ups.

The Immediate Assessment at the end of the Training 1 phase consisted of a post-test that was administered to all children in which they were asked to design a ramps experiment to test each target variable in a procedure identical to the Ramps Pre-test. Scores could range from 0 to 4.

Phase 4: Training and Immediate Assessment 2

In this phase children in LL and HH conditions followed the identical procedures to the Phase 3 (Training 1), while LH children now received a high degree of instructional guidance and HL children now received low levels of instructional guidance.[3] The Immediate Assessment at the end Training 2 was identical to the Immediate Assessment at the end of Training 1.

Phases 5–10: Transfer Assessments

The remaining test phases consisted of a series of assessments of the extent to which children could transfer CVS to familiar and/or novel tasks. No feedback was provided in any transfer phases, and the dependent variable was always how many CVS experiments each child designed.

[2] Mean time on task four the four set-ups in the High Guidance condition was 9:15 and for the eight set-ups in the Low Guidance condition 14:33.

[3] Recall that in the Low guidance condition, there were two experiments to assess the effect of each of the four variables. In order to minimize redundancy and potential disengagement in Phase 4, the procedure was slightly truncated whenever children's responses indicated that they had already learned how to test for a specific variable. More specifically, if a child's first experimental design for a specific variable was unconfounded, then the second opportunity to test this variable was skipped—because the child had already mastered CVS. All HL children and five children LL children exhibited this pattern.

Phase 5: Springs Transfer Test Materials for this phase consisted of a set of eight springs (identical to those used by Triona and Klahr 2003) that varied across three binary factors: spring length, wire size, and spring width. For each of the factors (chosen in a random sequence), children were asked to design an experiment to determine whether it made a difference in how far a spring stretched. Children chose two springs, hung them on a wooden rack, and then "ran" the experiment by hanging weights on the springs and observing how far they stretched (see Fig. 3). The child was assigned a 1 for an unconfounded comparison and a 0 otherwise. Scores ranged from 0 to 3.

Phase 6: Car Design Transfer Test A computer program (adapted from Klahr et al. 2007) presented on a laptop computer allowed children to design simple cars to determine which of four variables affected how far a car traveled. (Two of the variables had two levels, while the other two variables had three levels.) Children were told that they would be building and testing cars to figure out what might make one car travel farther than another. The experimenter recorded which configurations children had tested, so that they could refer to the record as necessary. Children proceeded to test each variable in a fixed sequence determined by the experimenter. Children had the opportunity to design 10 cars and the total possible chances of demonstrating CVS was 6.[4] Scores ranged from 0 to 6.

Phase 7: Ramps Post Test The materials, procedure, and scoring in this phase were identical to the Ramps Pre-test (Phase 1).

Phase 8: Story Post Test The materials, procedure, and scoring in this phase were identical to the Story Pre-test (Phase 2).

Phase 9: Remote Story Post Test This phase took place approximately 5 months after the second training (Phase 4) when children were in the 4th grade. Children sat at their desks and completed a paper and pencil test—adapted from the transfer test used by Toth et al. (2000)—that showed nine different experimental comparisons similar to those in Fig. 1 but in different domains involving foot races, plant growth, cookie baking, etc. Children were asked to indicate whether each experiment was a "good test" or a "bad test", and if "bad" to change it into a good test. Three of the nine questions were good (unconfounded tests of the focal variable), and the others were bad (one or more confounds). Children were given a score of 1 if they correctly identified the test as either good or bad, and an additional score of 1 if they changed a bad experiment to a good one. Scores ranged from 0 to 15.

Phase 10: Remote Ramps Post Test The materials, procedure, and scoring in this phase were identical to the Ramps Pre-test and Post-test (Fig. 2) and took place 5 months after the second training.

Results

One-way ANOVA's on each of the two pre-test scores revealed no differences among the four groups (all $ps > .82$). A 4 (condition) $\times$ 10 (test phase) mixed ANOVA found a

[4] Children tested each variable separately. For example, children designed two cars to test a binary variable (score of 1 possible), three cars to test a ternary variable (score of 2 possible), etc.

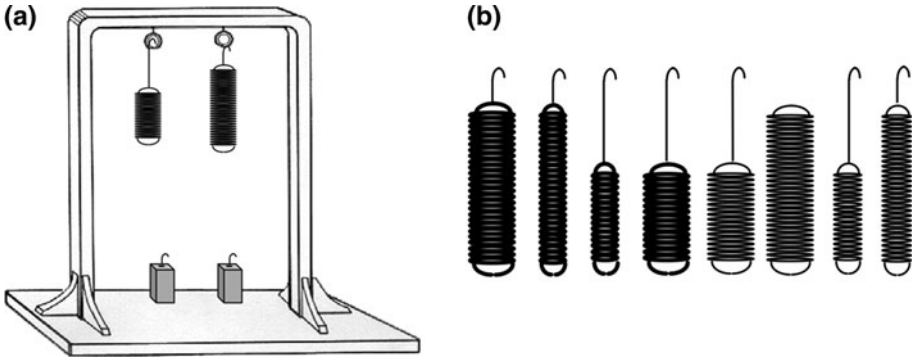


Fig. 3 Springs materials used in Springs Transfer Test (Phase 5) (from Triona and Klahr 2003). **a** A "good" experiment to assess the effect of spring *length* on how far a spring stretches: comparing a *short*, wide spring with thick wire to a *long*, wide spring with thick wire. **b** Set of eight springs that span all possible combinations of two spring lengths, two widths and two wire sizes

significant effect of condition $F(3, 48) = 4.88$, $p = .005$, $\eta^2 = .234$, test phase $F(9, 432) = 53.40$, $p < .001$, $\eta^2 = .527$, and a significant interaction $F(27, 432) = 1.89$, $p = .005$, $\eta^2 = .106$ (see Fig. 4).

Some of the phase-to-phase variability depicted in Fig. 4 is due to the interspersing of the repeated ramps assessments (Phases 2, 3, 4, 7, and 10), with several qualitatively different types of assessments (Phase 1—story problems; Phase 5—springs; Phase 6—cars; Phases 8 and 9—story problems). In order to remove this source of variability, we conducted a 4 (condition) × 5 (test phase) mixed ANOVA on only the phases that involved ramps (Phases 2, 3, 4, 7, and 10). This analysis revealed a main effect of condition $F(3, 48) = 4.31$, $p < .01$, $\eta^2 = .212$, test phase $F(4, 192) = 84.27$, $p < .001$, $\eta^2 = .637$, and a significant interaction $F(12, 192) = 2.60$, $p < .005$, $\eta^2 = .14$. Post-hoc tests revealed that at the end of Training 1, all groups improved from the Ramps Pre-test (all $ps < .01$).[5] However, HH and HL children performed at significantly higher levels than did the LL and LH children (all $ps < .01$). The LH group improved significantly from Training 1 to Training 2 ($p < .05$), while the LL group did not ($p > .53$). The HH and HL groups continued to perform at ceiling (all one-tailed $ps > .33$) at Training 2. Additionally, the performance of both the HH and HL groups was significantly superior to the LL group at all phases (all $ps < .05$).

To test whether there was an effect of instructional sequence on just the mixed groups (HL and LH), we conducted a 2 (condition: HL vs. LH) × 4 (phase: Ramps Pre-test, Training 2, Ramps Post-test, and remote ramps) mixed ANOVA. This analysis revealed a main effect of test phase $F(3, 69) = 67.22$, $p < .001$, $\eta^2 = .745$, but no effect of condition or condition × phase interaction (all $ps > .17$), suggesting that while both groups learned from the training sessions, there were no differences in performance between groups.

To examine performance on the Story Pre-tests (Phase 1) and post-tests (Phase 8), we conducted a 4 (condition) × 2 (test phase) mixed ANOVA. Results indicated a main effect of condition $F(3, 48) = 3.59$, $p < .05$, $\eta^2 = .183$, test phase $F(1, 48) = 93.07$, $p < .001$, $\eta^2 = .66$, and a significant interaction $F(3, 48) = 3.50$, $p < .05$, $\eta^2 = .179$. All groups evidenced significant gains from pre- to post-test (all $ps < .05$), and post hoc tests revealed a significant difference only between the HH and the LL groups ($p = .01$).

[5] All post hoc between-subject comparisons are Tukey-adjusted.

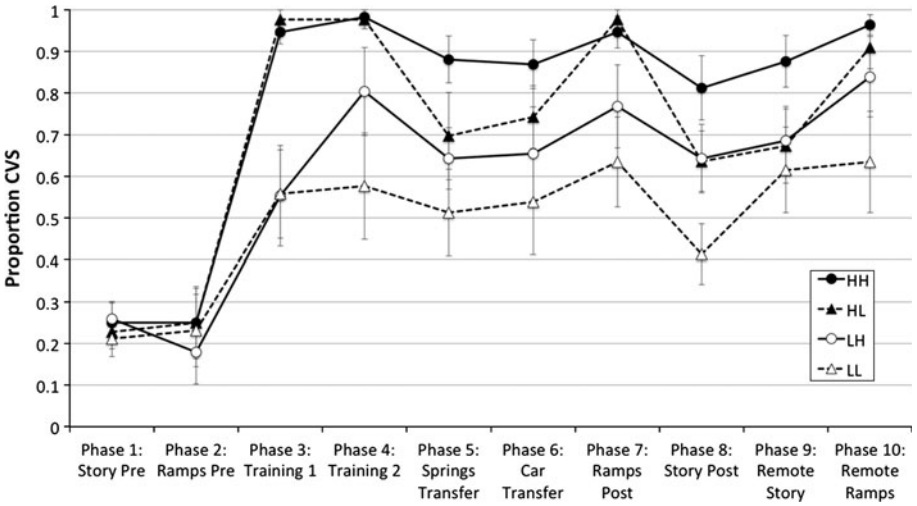


Fig. 4 Mean proportion of CVS experiments for each condition across all phases. Error bars represent standard errors of the mean

To examine performance on post-tests that involved transfer to a novel material over a relatively short period of time, we analyzed children's performance on the Springs and Car Transfer tests (Phases 5 and 6) by conducting a 4 (condition) × 2 (phase) mixed ANOVA. This analysis revealed a significant difference only between groups $F(3, 48) = 4.22$, $p < .05$, $\eta^2 = .209$. Post hoc analyses showed significant differences only between the HH and the LL groups ($p < .01$).

In order to determine how training conditions impacted remote transfer (after a long delay), we analyzed children's performance on the remote story problems (Phase 9) and remote ramps (Phase 10) 5 months after training. One-way ANOVA's at each of the assessments revealed no significant differences between groups on the remote story problems ($p > .14$), but significant differences on the remote ramps $F(3, 51) = 2.96$, $p < .05$. Post hoc tests revealed that the HH group performed significantly better than the LL group on the remote ramps ($p < .05$).

To further examine the remote transfer effects of different training conditions on individual children, we categorized each child as either an "expert" or a "non-expert" on the remote ramps and on the remote story problems. Children who correctly answered at least 12 of the 15 remote story problem items (80 %) were classified as story problem experts, and children who designed at least three of four unconfounded experiments (75 %) on the remote ramps assessment were classified as remote ramps experts. A χ^2-test of independence revealed a significant difference between groups on the remote ramps assessment ($\chi^2(3, 52) = 11.1$, $p = .01$) and a marginally significant difference between groups at the remote story problems assessment ($\chi^2(3, 52) = 6.23$, $p = .10$; see Table 3). Follow up tests on the remote ramps assessments indicated a significant difference between the HH and LL groups ($\chi^2(1, 52) = 5.85$, $p < .05$). No other tests on the remote ramps assessment were significant. Follow up tests on the remote story-problem assessments indicated marginally significant differences between the HH and HL groups ($\chi^2(1, 25) = 2.92$, $p = .08$), the HH and LH groups ($\chi^2(1, 28) = 2.62$, $p = .10$), and the HH and LL groups ($\chi^2(1, 27) = 3.13$, $p = .07$).

Table 3 Percentage of children (and raw numbers in parentheses) across each condition who were classified as remote ramps experts (Phase 10) and remote story problem experts (Phase 9)

Condition	Remote ramps		Remote story problems	
	Non-experts	Experts	Non-experts	Experts
HH	0 (0)	100 (14)	14 (2)	86 (12)
HL	9 (1)	91 (10)	45 (5)	55 (6)
LH	14 (2)	86 (12)	50 (7)	50 (7)
LL	46 (6)	54 (7)	54 (7)	46 (6)

Discussion

The primary aim of this study was to determine the effect of different sequences of high vs low guidance on learning and transfer. More specifically, we evaluated whether providing learners with low guidance prior to high guidance—similar to approaches that advocate "preparation for future learning" (e.g., Schwartz and Martin 2004) and "productive failures" (Kapur 2008, 2009)—would be more effective than high guidance followed by low guidance in helping children learn about experimental design. In contrast to predictions derived from "invention" approaches, we found no differences between the HL and the LH conditions on any of our measures of learning and transfer. Instead, children learned and transferred relatively well as long as they received high guidance at some point during instruction.

If invention approaches such as "preparation for future learning" and "productive failures" are robust, why didn't students in the present study benefit from high guidance after being allowed to invent partial steps toward effective procedures? We believe that one important difference between the present study and such invention studies is in the extent to which learners are aware that their failures are indeed failures. For instance, Kapur (2009) showed that learners given ill-structured problems often exhibited low confidence in their solutions, suggesting they were aware of their shortcomings before instruction. In contrast, our materials provide no feedback on whether an experimental set-up is "smart" or not. That is, if a child thinks that a confounded experiment is "good", there is nothing in the experimental set up or its outcome to indicate that it is not. Thus, in domains where learners have difficulty assessing the correctness of their solutions, there may be no advantage for delaying the provision of explicit guidance. In addition, invention studies of the type reported by Schwartz et al. typically implement instructional scaffolds that were omitted in the present study, such as the use of contrasting cases, which have been shown to produce robust learning gains on their own (e.g., Gentner et al. 2009). We conclude that there is likely a benefit for having students engage in invention activities before the provision of more explicit guidance, however, the present study suggests that the simple timing of when guidance is provided is not sufficient. Future research should more clearly specify the boundaries of when such instructional interventions will be effective (e.g., Roll et al. 2009).

Another aim of this study was to examine the immediate and longer-term effects of a second exposure of either high guidance or low guidance. We found that children in the HH group outperformed children in the LL group at every test phase after training. These results suggest that repeated exposure to instruction that consists of inquiry questions coupled with explicit instruction (i.e., our high guidance condition) can be a powerful way to promote robust learning and transfer of scientific experimentation procedures and concepts (Chen and Klahr 1999; Klahr and Nigam 2004; Lorch et al. 2010; Strand-Cary

and Klahr 2009; Klahr et al. 2007). These results are generally consistent with the predictions derived from Cognitive Load Theory in that a strong degree of guidance is predicted to optimize novices' learning. In addition, children's performance on the remote transfer tests challenges the assertion that direct instruction produces shallow learning and/or brief retention (Dean and Kuhn 2007). Instead, the results of these tests revealed not only that children who received high guidance at any point in the early training phases (HH, HL, and LH) retained what they learned about CVS over a five-month period, but also that the HH group outperformed the LL group on the remote ramps test. There was also some marginally significant evidence that the HH group outperformed all other groups on the remote story problems, suggesting that the extra exposure to high-guidance—received only by the HH children—was not redundant, but instead had effects over long duration and across domains. In other words, only children who had received a "double dose" of high guidance demonstrated expert levels of performance after a 24-week delay both on an assessment that was very similar to their initial training (Phase 10: ramps) and quite different in format, domain, and response type (Phase 9: remote story problems).

In conclusion, the results of this study seriously challenge the view that direct instruction is insufficient for promoting robust learning and that shallow transfer is one of its hallmarks. Instead, they suggest that, for novices struggling with domains providing few indications of mistakes or misconceptions, minimally guided instruction that eschews direct instruction may fail to optimize learning. We recognize that a variety of different instructional strategies may be optimal at different points in learning (Koedinger and Aleven 2007). Moreover, we caution against the overextension of our results to assume that high guidance is always the most effective form of instruction for all types of students, in all domains. At the same time, instructors should not be discouraged from providing high guidance to students under the premise that it is likely only to lead to fragile learning. The results of the present study support the view that providing high amounts of explicit, direct guidance—at least in early stages of learning—can be particularly effective in promoting robust understanding of scientific procedures and concepts in children.

Acknowledgments The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grants R305B040063 and R305A100404 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Thanks to Howard Seltman, Anna Fisher, Marsha Lovett, Stephanie Siler, and Miriam Novack for their comments and suggestions on earlier version of this paper. We also thank the teachers and students of The Campus School at Carlow University and Sacred Heart Elementary School for their enthusiastic cooperation throughout the project.

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.

Chen, Z., & Klahr, D. (1999). Remote transfer of scientific reasoning and problem-solving strategies in children. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol. 36, pp. 419–470). Amsterdam: Elsevier.

Chinn, C. A., & Malhotra, B. A. (2001). Epistemologically authentic scientific reasoning. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 351–392). Mahwah, NJ: Erlbaum.

Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education, 91*, 384–397.

Gentner, D., Levine, S., Dhillon, S., & Poltermann, A. (2009). Using structural alignment to facilitate learning of spatial concepts in an informal setting. In B. Kokinov, K. Holyoak, & D. Gentner (Eds.), *Proceedings of the second international conference on analogy*. Sofia: NBU Press.

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. (2009). Productive failure in mathematical problem solving. *Instructional Science, 38*, 523–550.

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 & T. M. Duffy (Eds.), *Constructivist theory applied to instruction: Success or failure?*. New York: 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, 44*, 183–203.

Koedinger, K. R., & Aleven, V. (2007). Addressing the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*, 239–264.

Kuhn, D. (2007). Is direct instruction an answer to the right question? *Educational Psychologist, 42*, 109–113.

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

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about controlling variables? *Psychological Science, 16*, 866–870.

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), 137–151.

Lorch, R. F., Jr, Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). Learning the control of variables strategy in higher- and lower-achieving classrooms: Contributions of explicit instruction and experimentation. *Journal of Educational Psychology, 1*, 90–101.

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.

Roll, I., Aleven, V., & Koedinger, K. R. (2009). Helping students know 'further'—Increasing the flexibility of students' knowledge using symbolic invention tasks. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 1169–1174). Austin, TX: Cognitive Science Society.

Schwartz, D. L., & Bransford, J. (1998). A time for telling. *Cognition and Instruction, 16*, 475–522.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology, 103*, 759–775.

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.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.

Taber, K. S. (2010) *Constructivism and direct instruction as competing instructional paradigms: An essay review of Tobias and Duffy's constructivist instruction: Success or failure?* (Vol. 13, No. 8) New York: Routledge.

Tobias, S., & Duffy, T. M. (2009). *Constructivist instruction: Success or failure?* New York: Routledge.

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.

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.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172–223.