APA Reference: Martella, A. M., Klahr, D., & Li, W. (2020). The relative effectiveness of different active learning implementations in teaching elementary school students how to design simple experiments. *Journal of Educational Psychology, 112*(8), 1582-1596. https://doi.org/10.1037/edu0000449

The Relative Effectiveness of Different Active Learning Implementations in Teaching Elementary School Students How to Design Simple Experiments

Abstract

"Active learning" has been used to describe classrooms that have varied widely with respect to instructional topics, age of learners, and the particular procedures used to operationalize the general notion of the term. In most cases, the specific variant of active learning under investigation has been more effective than the particular control used for comparison. The goal of the current study was to unambiguously describe, implement, and assess four different active learning implementations that varied based on the particular instructional technique employed by the teacher. The specific topic taught was the procedure for constructing experiments in which a single causal factor is clearly identified and there are no confounds. The procedure is commonly known in the literature on early scientific thinking as the "Control of Variables Strategy" (CVS). The sample consisted of 145 third- and fourth-grade students from three schools. Students in each grade at each school were randomly assigned to one of four active learning conditions. Learning of CVS was measured through a hands-on, active learning activity and a written pre- and posttest. Results indicated that compared to Minimal Guidance/Minimal Guidance/Activity, Modeling/Direct Guidance/Activity resulted in significantly higher levels of CVS knowledge on the hands-on activity. When examining student learning from pre- to posttest, students in all conditions had significant learning gains. However, the largest effect sizes were for Modeling/Direct Guidance/Activity followed by Modeling/Modeling/Activity and the weakest effect size was for Minimal Guidance/Minimal Guidance/Activity. Thus, more direct/explicit forms of active learning promoted higher learning of CVS than more inquiry-based forms.

Keywords: active learning, science instruction, control-of-variables strategy, inquiry-based learning, direct/explicit instruction

Educational Impact and Implications Statement

Our article presents a scientific investigation of different implementations of active learning. Active learning is a popular instructional method that has not been the subject of well-controlled experimental studies that investigate which features of active learning lead to greater student learning than with other forms of active learning or with more passive forms of instruction. We implemented four forms of active learning in the context of teaching elementary students how to design simple experiments and found each form of active learning to differ in its effectiveness for student learning. Students experienced higher levels of learning when active learning was implemented with more direct/explicit forms of instruction rather than with more inquiry-based forms of instruction. Thus, active learning is an approach whose features need to be systematically isolated and studied to identify how and why active learning can be effective; moving beyond the typical "active learning versus lecture" contrast most often studied in the research literature is an important and needed endeavor.

The Relative Effectiveness of Different Active Learning Implementations in Teaching Elementary School Students How to Design Simple Experiments

A broad spectrum of teaching approaches, ranging from highly teacher-directed to highly student-directed, is found in K-16 schools. There is an equally broad spectrum of descriptive labels for these approaches, ranging from “inquiry- or discovery-based learning” (Marx et al., 2004) to “direct and explicit instruction” (Stockard, Wood, Coughlin, & Khoury, 2018). One particular term, “active learning,” has been widely used in the research literature on instructional effectiveness. In most of those studies, the typical contrast to “active learning” is “passive learning” where students do not participate in any component of the instructional process other than listening, reading, or watching videos (Chi & Wylie, 2014).

Defining Active Learning

When examining the research literature, the common feature of active learning instructional approaches is that students are engaged with the instructional materials in some fashion and are playing a larger role in the learning process. Learning is classified as “active” to the extent that during instruction, students are engaged in overt, constructive, and/or interactive behaviors (a classification we utilize in our study). However, most of the active learning literature lacks a deep theoretical model for why and how active learning is effective. In this paper, we draw on some aspects of Chi and Wylie’s (2014) Interactive Constructive Active Passive (ICAP) theoretical framework. In this framework, active learning can be distinguished from passive learning in three important ways: First, learner engagement can be objectively measured and assessed—that is, some form of overt action is observed (e.g., copying steps, rotating objects). Second, constructive behaviors can be displayed through the generation of outputs (e.g., self-explanations, reflections out-loud). Finally, relevant interactions can be observed where each member in a group contributes actively and constructively (as observed by students' overt actions and the outputs they generate); the interaction can occur between students or between students and the instructor.

According to the ICAP framework, passive learning leads to limited understanding and is least effective when students need to integrate prior knowledge to go beyond rote memorization and transfer their knowledge to new situations (Chi & Wylie, 2014). Consistent with this prediction, studies on active learning have provided evidence of its effectiveness in a wide range of student grade levels such as middle school (e.g., Akinoglu & Tandogan, 2007) and college (e.g., Wieman, 2014) and disciplines such as environmental engineering (e.g., Kinoshita, Knight, & Gibbes, 2017), chemistry (e.g., Eichler & Peeples, 2016), and biology (Taraban, Box, Myers, Pollard, & Bowen, 2007). Examples of ways in which students can transfer their knowledge to new situations include students learning chemistry concepts and applying them to problem solving questions (Eichler & Peeples, 2016) or students learning biology concepts and applying this knowledge during lab experiments/experiences and on exam questions (Taraban et al., 2007). Additionally, active learning is commonly viewed as an essential feature of problem-, project-, inquiry-, case-, experiential- and discovery-based learning (e.g., Akinoglu & Tandogan, 2007; Cattaneo, 2017; Kolb & Kolb, 2005). In explicit instruction, active learning is also seen as the engagement component (Archer & Hughes, 2011).

Implementing Active Learning

A seminal definition of active learning is "anything that involves students in doing things and thinking about the things they are doing" (Bonwell & Eison, 1991, p. 2). More recent definitions tend to be just as open-ended and, as such, lead to variations in how active learning is implemented in classrooms. For example, Freeman et al. (2014) noted in their meta-analysis on active learning that “the active learning interventions varied widely in intensity and implementation” (p. 8410). Our examination of the active learning literature suggests there are four primary sources of variation in active learning implementations. The first source of variation is the focus of our study, the second source of variation relates back to the ICAP Framework, and the third and fourth sources of variation are additional elements we identified. First, the instructional techniques employed by the instructor can differ across classrooms. For example, instructors may choose to (a) lecture on a topic before engaging students in activities (e.g., Eddy & Hogan, 2014), (b) provide students with these activities before they provide students with lecture (e.g., Kapur, 2014), (c) intersperse these activities throughout a lecture (e.g., Webb, 2017), or (d) omit lecture altogether and guide students through these activities (e.g., Adams, Garcia, & Traustadottir, 2016). Second, the types of student behaviors may differ in different active learning implementations (e.g., students may be engaging in active, constructive, or interactive behaviors) (e.g., Chi & Wylie, 2014). Third, the types of active learning activities can differ between/among classrooms (e.g., Freeman et al., 2014). These activities may include clicker questions, worksheets, or case studies, for example, and be assigned to individuals or groups. Fourth, the intensity (i.e., duration, number, variety) of activities can differ across classrooms (e.g., Rau, Kennedy, Oxtoby, Bollom, & Moore, 2017). It should be noted that a potential fifth source of variation could be differences across students such socioeconomic status, skill level, race/ethnicity, and gender.

Given these implementation variations, it is often difficult to determine which pedagogical practices of any specific version of active learning are causally related to differences in their outcomes (Rosenshine, 2008). For example, certain forms of active learning could be categorized as more direct-instruction approaches or more inquiry-based approaches although these categorizations are also not useful as there are implementation variations in these as well. One common misconception is that active learning is antagonistic to direct instruction, under the assumption that direct instruction applies only to the lecture method and is entirely teacher based. However, many studies of direct instruction also include teacher guidance, immediate feedback, and student-directed, independent activities (c.f., Lorch et al., 2010). Inquiry-based learning is another teaching approach where disagreement exists on its core features (Hmelo-Silver, Duncan, & Chinn, 2007; Kirschner, Sweller, & Clark, 2006). Inquiry-based learning may include no guidance, minimal guidance, and minimal guidance combined with teacher feedback (Lazonder & Harmsen, 2016), among other levels of support.

This lack of operationally defined instructional approaches—approaches that are unambiguous and lead to clear guidelines for implementation—is particularly troublesome in STEM education (Klahr, 2010; Klahr, 2013). When an instructional approach can be implemented in several ways due to a broad definition, it is difficult to determine which aspects of the instructional approach are causally related to differences in learning outcomes and which implementations are the most effective in encouraging student learning. The active learning literature, to date, provides no unambiguous guidance to instructional designers who would like to ground their teaching methods in consistent prior empirical results pertaining to active learning.

Relative Effectiveness of Different Types of Active Learning Implementations

To begin to address these ambiguities, we focused our efforts on the first source of variation active learning implementations—the instructional techniques employed by the instructor. We compared four active learning implementations that differed on the instructor's role (and, consequently, the students' role) in the learning process. We designed our four active learning conditions to be consistent with the ICAP framework in that students were engaged in the learning process by engaging in active, constructive, and/or interactive behaviors. We chose the ICAP Framework as it is one of the few theoretical frameworks in the research literature that addresses elements of active learning and their effectiveness and has been the subject of much research (e.g., Chi et al., 2018). We also designed our conditions to allow us to determine which specific elements of the active learning implementations were or were not effective for student learning. Students in the four conditions were taught the same scientific topic and received the same active and constructive activity at the end of the lesson (the type and intensity of the activity were kept consistent). However, the instructional techniques used in each active learning condition differentiated one condition from another; that is, active learning was implemented in four different ways with the only source of variation coming from the instructor's instructional procedure. Thus, our study addressed Freeman et al.'s (2014) call to move beyond comparing active learning to lecture and instead investigate (a) "*which* aspects of instructor behavior are most important for achieving the greatest gains with active learning" and (b) "*which* type of active learning is most appropriate and efficient for certain topics or student populations" (p. 8413). Next, we describe our instructional goal and information about the aim of our study.

Instructional Goal

Our goal was to teach third- and fourth-grade children how to use the "Control of Variables Strategy" (CVS) to design simple, unconfounded experiments. Acquisition of CVS is an important component of students' scientific-reasoning skills because it provides them with a method to ensure their experiments will be unconfounded and will enable them to identify causal variables based on experimental outcomes. To design unconfounded experiments, students must understand that only when all variables except the one being studied are held constant is it possible to make valid causal inferences (Chen & Klahr, 1999). One common way to teach CVS is to use two physical ramps where students can attempt to identify causal factors in how fast a ball rolls down a ramp by choosing to vary or to hold constant a small set of potentially causal factors, such as the heights of the ramps or their surfaces (Lorch et al., 2010). Comparing how ramps are set up can allow students to make unambiguous causal inferences.

Acquisition and mastery of CVS skills typically requires instruction and practice (Schwichow, Croker, Zimmerman, Hoffler, & Hartig, 2016). The efficacy of different methods for teaching CVS has been investigated for over 40 years (Case, 1974). Studies have ranged from examining different pedagogical approaches on students' CVS proficiency (e.g., Chen & Klahr, 1999) to investigating the impact of metacognitive strategies on the transfer of learning (e.g., Lin & Lehman, 1999). Studies of knowledge acquisition have also often addressed different temporal durations of instructional treatments ranging from short lessons (e.g., 30-45 min) (e.g., Chen & Klahr, 1999) to longer lessons (e.g., over the course of days, weeks, or months) (e.g., Kuhn et al., 1995) However, to date, the nuances of different active learning implementations have not been directly investigated in these highly focused and brief instructional interventions.

Study Aim

The aim of the present study was primarily to contribute to the active learning literature. First and foremost, given the research supporting active learning's *general* effectiveness, our goal was to examine some finer-grained aspects of active learning to determine which forms were most effective in promoting learning. The research literature on active learning is lacking in these finer-grained analyses. Second, relating to CVS, although CVS has been studied extensively by researchers in psychology (Chen & Klahr, 1999; Case, 1974) and education (Lin & Lehman, 1999; Lorch et al., 2010), it has not been the focus of active learning studies. In other words, active learning studies have not focused on teaching students CVS. Third, the instructional "grain size" in active learning studies with college aged adults is typically several weeks to months, given that a high proportion of them are conducted over the course of a college semester (see Freeman et. al., 2014). In contrast, the instructional duration in our study was under an hour and was focused on one specific topic (the design of simple unconfounded experiments). Relatedly, most of the comparisons between active learning and other instructional methods have been based on students' grades (or other aggregate performance measures) over semester-length instructional periods, whereas our assessment of instructional efficacy focuses on elementary-school children's mastery of the specific concepts and procedures associated with simple experimental design. Fourth, the instructional goal of active learning studies conducted at the college level is for students to acquire sufficient knowledge about the topic (e.g., basic physics course) that will enable them do well on the final exam, whereas our learning assessments are highly focused—they measure the acquisition and learning of the procedural and conceptual aspects of CVS.

Research Questions

Our finer-grained analysis of active learning focuses on three questions: (a) are the active learning implementations used in this study equivalent in their effects on student learning of CVS? (b) if there are differences in learning, which implementations are most effective in promoting students' acquisition and transfer of CVS? and (c) if there are differences, what about the implementation led it to be more effective than another? We did not include a "non-active-learning" control group in this study because our aim was not to assess active versus non-active learning; there is already a substantial literature base on that question, and a strong argument can be made (c.f., Freeman et al., 2014) for second generation research that moves beyond comparing active learning to lecture. Instead, our goal was to determine whether or not our four distinct active learning implementations had differential effects on acquisition and transfer of CVS, and to the extent that they did, to identify the specific pedagogical elements of each implementation that led to those differences.

Method

Participants

Participants were 145 third- and fourth-grade students (mean age = 9.2) from one local private and two local charter schools in a mid-sized urban city in the eastern U.S (see Table 1 for demographics). Students were recruited through the cooperation of school and district directors of research and/or learning, who then contacted science teachers and other third- and fourth-grade teachers. Teachers were asked if they were interested in voluntarily giving their students the option to be included in the study. Six teachers and eight classrooms of students participated in the study. Research ethics approval was obtained for our study through the Institutional Review Board; consent fell under the "opt-out method."

The study used an experimental design that randomized students to one of the four conditions (Modeling/Modeling/Activity, Modeling/Direct Guidance/Activity, Minimal Guidance/Direct Guidance/Activity, and Minimal Guidance/Minimal Guidance/Activity). There were 167 students in 8 classes of 3 schools originally recruited and randomized within classes to four conditions. After randomization, 45 students were in Minimal Guidance/Minimal Guidance/Activity; 40 students were in Minimal Guidance/Direct Guidance/Activity; 39 students were in Modeling/Modeling/Activity; and 43 students were in Modeling/Direct Guidance/Activity. Our analytic sample (students with both pre- and post- tests) had 145 students including 37 students in Minimal Guidance/Minimal Guidance/Activity; 37 students in Minimal

Guidance/Direct Guidance/Activity; 35 students in Modeling/Modeling/Activity; and 36 students in Modeling/Direct Guidance/Activity. The overall attrition rate was 13.17%, and maximum differential attrition rate was 10.28%.

The attrition rates were unrelated to the study, as the intervention happened in one single class period. Our attrition came from the absence of students in our schools on the pre- and/or posttest or lesson. Teachers notified us that some students in their classrooms did not attend school every day and thus may or may not be present for each part of our study. The combination of overall (13.17%) and differential (10.28% maximum) rates of attrition shows that the threat of attrition bias of this study is a tolerable threat under optimistic assumptions (What Works Clearinghouse Standards Handbook 4.0, 2017).

Materials

The CVS lesson was conducted using physical ramps very similar to those used in earlier CVS studies (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004; Lorch et al., 2010) (see Table 2 for a description of the problem domain). Each student received two ramps during the lesson portion of the study (either in Part 2, Part 3, or in both Parts 2 and 3). There were three variables associated with the ramps that could assume either of two values: *ramp height* (short or tall), *ramp surface* (smooth or rough texture), and the *starting position of the ball* (top of the ramp or near the middle of the ramp) [1]. Students could adjust each of the three variables using physical height, surface, and starting position pieces. They were asked to focus on one of the three variables and to design an experiment that would conclusively determine whether or not the chosen variable had a causal effect on how far the ball on each ramp rolled. In this study, as in all of the research on teaching CVS (Chen & Klahr, 1999; Klahr & Nigam, 2004; Dean & Kuhn,

[1] Ball type was not varied in this study. The same type of ball was used on each ramp.

2006; Lorch et al., 2010; Schwichow et al., 2016), the fundamental procedural and conceptual knowledge to be acquired is that the experimental set up must vary only the focal variable and hold all others constant in order to create an unconfounded experiment from which a legitimate causal inference can be made. For example, students might be instructed to "see how the ramp height affects how far the balls will roll."

Procedure

We implemented a four-level, single factor (treatment condition) design. Each condition consisted of a pretest, a lesson that included four parts, and a posttest (see Figure 1). The study began with a written pretest that was administered to the whole class. At least 1 day after the pretest, students participated in groups of three to six students in a 40-minute lesson on CVS (four parts to the lesson) (see Figure 1). For the lesson phase of the study, students in each class were randomly assigned to be in a group of three to six students; these groups were randomly assigned to one of the four conditions. Each lesson consisted of an overview (Part 1), two instructional treatments (Parts 2 and 3), and a hands-on activity involving CVS and the ramps (Part 4) (see Figure 1). Students remained in their small groups for Parts 2 and 3 but were separated in the room for Part 4. The four experimental conditions differed only in what happened in Parts 2 and 3 and were labeled according to the main parts of the lesson: (a) Minimal Guidance followed by Minimal Guidance followed by the hands-on activity (Minimal Guidance/Minimal Guidance/Activity) (n=37); (b) Minimal Guidance followed by Direct Guidance followed by the hands-on activity (Minimal Guidance/Direct Guidance/Activity) (n=37); (c) Modeling followed by Direct Guidance followed by the hands-on activity (Modeling/Direct Guidance/Activity) (n=36); and (d) Modeling followed by Modeling followed by the hands-on activity (Modeling/Modeling/Activity) (n=35). Between 2 and 8 days after receiving the lesson, students completed a written posttest (see Figure 1) that was administered to the whole class. After the posttest, students received a debriefing on the study that consisted of an explanation of what we were testing (i.e., seeing which teaching method best helped students learn about experiments).

Story problem pretest. Students began the study with a group-administered (but individually taken) written pretest containing 10 scenarios showing experimental contrasts in domains that included building planes, baking cookies, and designing rockets (see Figure 2 for example question). The test questions were selected from questions previously asked in several closely related studies (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004). One question asked students to select (from A, B, C, or D) the unconfounded experiment. The other nine questions asked students to judge whether or not an experiment was confounded by circling *good way* or *bad way* to design an experiment. Students were also asked to explain their answer choices. All students received the same pretest and were given 40 minutes to take the test.

Lesson: Four parts. Students received instruction in groups of three to six for 40 minutes. The group size was dependent on a number of pragmatic factors, unrelated to the overall experimental design, such as accretion, attrition, and class size. Of the nine groups of students in each condition, at least three of the groups were small (i.e., three or four students) and at least four were large (i.e., five or six students). Students received varying levels of teacher guidance to learn CVS and were given the opportunity to actively and constructively put CVS into practice through a hands-on activity involving the ramps. The lesson was taught in a quiet location in the students' school; the instructor and students sat around a large table. The same instructor (first author) taught each lesson. Each lesson was carefully scripted and rehearsed.

Part 1. As noted above, Part 1 was the same for all of the conditions. It consisted of a brief overview of the lesson in which the instructor described the meaning of a “good experiment” (i.e., one where all variables are held constant except the variable of interest) as well as the goal of the lesson (i.e., to see if we can determine which factor caused one ball to roll farther than the other ball). The overview lasted approximately 5 minutes.

Overview of instruction in parts 2 and 3. As previously stated, our instructional treatments occurred during Parts 2 and 3 of the lesson. The goal of the instruction was for students to learn CVS. The four experimental conditions were comprised of different sequential combinations of three types of instruction that we called “Minimal Guidance,” “Direct Guidance,” and “Modeling.” First, we will describe the *content* of each type of instruction, and then we will describe Parts 2 and 3 of the lesson in detail (as depicted in Figure 1). We will also provide justification for the different sequential combinations we chose.

In *Minimal Guidance,* students were given their own set of ramp materials and provided with a CVS task that they were told to explore using their materials. They received neither a teacher explanation nor specific feedback while they explored CVS; however, they did receive questions such as, “Why did you change two things at once?” and “Can you tell from your set up which of the materials caused one ball to roll farther than the other?” as they explored the ramp domain. Thus, the term *minimal guidance* refers to the instructor providing only inquiry-based questions to the students as they explore and inquire during the lesson.

In *Direct Guidance,* students received their own set of ramp materials and worked through a CVS task with the instructor, receiving specific feedback as they set up their ramps concurrently with the instructor. Thus, the term *direct guidance* refers to the instructor providing specific student feedback and guidance. For example, if students changed two variables at once, the instructor explained why only one of the variables should be changed at one time and had students correct their ramp setup to be correct/unconfounded.

In *Modeling*, students did not receive their own set of ramp materials nor did they receive specific feedback. Instead, the instructor used the ramp materials to provide an explanation and demonstration of the CVS task with one unconfounded and one confounded experiment. Thus, the term *modeling* refers to the instructor-led demonstration of CVS.

Part 2 details. Part 2 varied according to condition and lasted about 10 minutes. The CVS procedure that was minimally guided or modeled in this part was, “Set up the ramps to see if the height of the ramps makes a difference in how far the balls roll.” The instructional treatment was *minimal guidance* for conditions Minimal Guidance/Minimal Guidance/Activity and Minimal Guidance/Direct Guidance/Activity and *modeling* for conditions Modeling/Direct Guidance/Activity and Modeling/Modeling/Activity.

Part 3 details. Part 3 also varied according to condition and lasted about 10 minutes. The CVS procedure that was minimally guided, directly guided, or modeled in this part was, “Set up the ramps to see if the starting position of the balls makes a difference in how far the balls roll.” The instructional treatment was *minimal guidance* for condition Minimal Guidance/Minimal Guidance/Activity, *direct guidance* for conditions Minimal Guidance/Direct Guidance/Activity and Modeling/Direct Guidance/Activity, and *modeling* for condition Modeling/Modeling/Activity.

Recap of parts 2 and 3. After an overview of CVS (Part 1 of the lesson), students had the opportunity to be involved in two CVS tasks that took place during Parts 2 and 3 of the lesson. The level of teacher guidance during these tasks varied among conditions. Thus, students in Minimal Guidance/Minimal Guidance/Activity were given only inquiry-based questions during the first and second tasks; students in Minimal Guidance/Direct Guidance/Activity were given inquiry-based questions during the first task and specific feedback during the second task; students in Modeling/Direct Guidance/Activity were given a teacher-led demonstration during the first task and specific feedback during the second task; and students in Modeling/Modeling/Activity were given a teacher-led demonstration during the first and second tasks.

Justification of parts 2 and 3. The purpose of having two phases of instruction (Parts 2 and 3 of the lesson) stemmed from common instructional approaches in the research literature such as direct instruction (modeling, direct guidance) and productive failure (minimal guidance, direct guidance) that typically involve two different phases. To keep the conditions the same with regard to time on task and exposure to the CVS material, we made all four of the conditions have two instructional parts.

Although there were 9 possible combinations of the three types of instruction (i.e., modeling, direct guidance, and minimal guidance), we chose our four combinations (i.e., minimal guidance/minimal guidance[2], minimal guidance/direct guidance, modeling/direct guidance, and modeling/modeling) for the following reasons. First, we wanted to examine the relative effectiveness of active learning implementations that ranged from highly teacher-directed to highly student-directed (where the source of variation came from the instructional techniques employed by the instructor). At one end of the spectrum is a condition that includes only minimal guidance, and at the other end of the spectrum is one that includes only teacher modeling. The other two conditions, near the middle of the instructional spectrum, include one of these instructional endpoints combined with constructive feedback. Thus, although we had two

[2] We have omitted "Activity" from each condition label in this section to help focus attention on the instructional phases that occurred in Parts 2 and 3.

phases of instruction (Parts 2 and 3 of the lesson), we did not design the conditions for us to determine which form of instruction should occur in Part 2 and which should occur in Part 3. Rather, we designed our conditions to vary with regard to the levels of support provided by the teacher in an active learning CVS lesson and to assess how much support (ranging from weak support [minimal guidance/minimal guidance] to strong support [modeling/modeling]) is most effective in active learning.

Second, we wanted to choose the combinations that were highly representative of instructional approaches in the research literature. Modeling/Modeling is highly representative of a lecture-based approach (e.g., Freeman et al., 2014). Modeling/Direct Guidance is representative of a direct instruction approach with teacher modeling first and additional instruction/feedback/explanation second (e.g., Rosenshine, 2008). Minimal Guidance/Direct Guidance is representative of the productive failure approach with exploration first and instruction/feedback/explanation second (e.g., Kapur, 2016). Finally, Minimal Guidance/Minimal Guidance is representative of an inquiry-based approach with student-driven exploration occurring throughout the learning process (e.g., Clark, Kirschner, & Sweller, 2012).

Third, certain combinations could not be implemented appropriately (i.e., direct guidance/modeling, direct guidance/minimal guidance, direct guidance/direct guidance) because direct guidance requires students to have prior knowledge (whether gained through modeling or minimal guidance) with which to use when working simultaneously with the instructor. This form of instruction is not intended to teach the students the content but rather help them refine their understanding through feedback. If students had not first received a teacher demonstration (i.e., modeling) or guiding questioning strategies (i.e., minimal guidance), they would have difficulty accessing the opportunity to work in unison with the instructor due to their lack of CVS knowledge. Thus, direct guidance had to exist in Part 3 of the lesson rather than in Part 2.

Fourth, minimal guidance/modeling would have contradicted a constructivist approach by moving from minimal guidance to full support and skipping the logical progression of support from minimal to moderate to full. Thus, we did not include this combination due to our earlier discussion of choosing approaches that are common in the research literature. Fifth and relatedly, we did not include modeling/minimal guidance because the purpose of minimal guidance is for students to discover the information with minimal support; however, if they received modeling first, they would have been provided with the information and thus not been able to discover it (i.e., it is difficult to discover what you have already been shown).

Part 4. Part 4 was the same for all conditions. In this part, the instructor asked students to independently set up (active behavior) an unconfounded experiment using the ramps. Specifically, students were asked to set up their ramps to see whether or not “the surface of the ramps made a difference in how far the balls roll.” The individual students comprising each small group were positioned in separate parts of the room (so that they could not hear or observe their peers), presented with all of the ramp materials, and told they had 10 minutes to complete the task. Upon task completion, the instructor approached each student, scored the setup of the ramps, and asked the student to (a) “explain why you set up the ramps in that way” (constructive behavior) and (b) “roll the balls down the ramps and tell me what you found out and how you can be sure” (active and constructive behaviors). This activity served as an opportunity for all of the students to consolidate and demonstrate their knowledge from Parts 2 and 3; it also served as one of our dependent measures.

Additionally, this part followed the “active” and “constructive” elements of ICAP, aligning these conditions to an active learning theoretical framework. The “active” portion was the opportunity for students to manipulate the physical ramps and the “constructive” portion was the opportunity for students to reflect out loud and generate responses using their prior knowledge from Parts 2 and 3 from the lesson. The “interactive” part of the ICAP Framework was not included in Part 4 as this was an individual task and students did not talk with their peers during this portion of the lesson. It should be noted that Part 4 was specifically part of the CVS lesson and was intended to be an active learning activity in which students participate. We also used Part 4 as one of our dependent measures although its primary purpose was as the main (and consistent across conditions) active learning element of each condition. Further, our focus was on comparing four active learning implementations that differed on the instructor’s role (and, consequently, the students’ role) in the learning process. Although Part 4 of the lesson was not part of the instructional comparison, it was still an integral element of each active learning condition and the lesson as-a-whole engaged students in active and constructive processes.

Part 4 included approximately 10 minutes to set up the ramps and 5 minutes to explain ramp setup. Performance on the hands-on activity served as a dependent measure of acquisition of CVS knowledge. To reiterate, the source of variation in these active learning conditions came from the instructional techniques employed by the instructor; thus, the active learning activity at the end of the lesson was kept the same with regard to form and intensity.

Story problem posttest. The study ended with a group-administered (but individually taken) written posttest given within 8 days of completing the lesson. The posttest followed the same format as the pretest and included 10 scenarios showing experimental contrasts in domains that included boiling water, selling drinks, and designing planes. These test questions were very similar to questions previously asked in several closely related studies (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004) and were matched in wording and complexity to the questions asked on the pretest. All students received the same posttests and were given 40 minutes to take this test. The posttest served as the dependent measure of student transfer of CVS knowledge.

Scoring Tests and Hands-On Activity

Pre- and posttest scoring. Student selections on each of the 10 multiple choice questions were assigned a score of “0” if incorrect and a “1” if correct. Student explanations for each of their 10 answers were graded on a 4-point scale ranging from 0 to 4. A “0” indicated that the explanation failed to include any language relevant to CVS. A “1” indicated the student discussed at least one variable (thus, related to CVS) but not correctly (e.g., the explanation “It’s a good experiment because they changed the type of sweetener” would be incorrect if, in fact, it was the oven temperature that needed to be changed). A “2” indicated the student discussed at least one variable correctly (though they may have talked about the other variables incorrectly or they may have only talked about one variable) (e.g., The explanation, “It’s a good experiment because they changed the oven temperature” would be correct if it was the oven temperature that needed to be changed). A “3” indicated the student discussed two of the three variables correctly (though they may have talked about the third incorrectly or they may have only talked about two of the variables) (e.g., The explanation, “It’s a good experiment because they changed the oven temperature and kept the sweeteners the same” would be correct if only the oven temperature should have been changed). A “4” indicated the student discussed (or implied) all three variables correctly (e.g., The explanation, “It’s a good experiment because only the temperature of the oven was changed” or “It’s a good experiment because they changed the oven temperature and kept the sweeteners and number of eggs the same" would be correct if only the oven temperature should have been changed). The total number of points possible on the pre- and posttest was 50.

Hands-on activity. Ramp setups received scores of "0" or "1." A "0" indicated there were one or more confounds in the ramp setup. A score of "1" indicated all variables except the variable of interest (i.e., surface texture) remained constant. Similar to the scoring for pre- and posttest explanations, the ramp explanations received scores from 0 to 4. A "0" indicated the student's explanation did not consist of language relevant to CVS. A "1" indicated the student discussed at least one variable (thus, related to CVS) but not correctly (e.g., I changed the height pieces because…"). A "2" indicated the student discussed at least one variable correctly (though they may have talked about the other variables incorrectly or they may have only talked about one variable) (e.g., "I kept the height pieces the same because…"). A "3" indicated the student discussed two of the three variables correctly (though they may have talked about the third incorrectly or they may have only talked about two of the variables (e.g., "I kept the height pieces the same and changed the surface pieces because…"). A "4" indicated the student discussed all three variables correctly (e.g., I kept the height pieces the same and the starting positions the same but changed the surface pieces because…").

Treatment Fidelity

An independent observer (research assistant B) determined treatment fidelity by observing more than half (56%) of the lessons and scoring the instructor on lesson implementation. The authors developed a checklist for each condition for the independent observer to score the instructor on teaching. This checklist consisted of four checkpoints (a total score of 4 points). The first checkpoint was whether or not the instructor followed the script (with slight deviations allowed if they did not change the tenets of the condition). The other three checkpoints had to do with the important characteristics of the condition (e.g., providing specific feedback or asking probing questions). For example, if the instructor failed to provide feedback during direct guidance in the Modeling/Direct Guidance/Activity condition, the instructor would lose a point. The instructor adhered to the lesson script and condition on each observed lesson and received an overall score of 100% for the fidelity checks.

Inter-Rater Reliability

Student scores on the pre- and posttest and the hands-on activity were ordinal in nature; thus, we calculated inter-rater reliability using several weighted reliability measures. We report weighted Cohen's Kappa (Cohen, 1988) only as the coefficients for Krippendorff's Alpha (Krippendorff, 2011), Gwet's AC_2 (Gwet, 2008), Scott's Pi (Scott, 1955), and Brennan-Prediger (Brennan & Prediger, 1981) had similar coefficients (i.e., they did not deviate from one another by more than .05) and had weighted agreement coefficients that fell within the "almost perfect" benchmark range (.80-1.00) when using cumulative membership probabilities.

Pretest. All pretests were scored independently by two scorers (the first author and research assistant A). A scoring rubric was developed by the authors and utilized by both scorers. The rubric provided information on how an explanation would qualify as receiving a score of 0 to 4 (see previous sub-heading "Scoring Tests and Hands-On Activity" for the explanations of what each score meant) with examples and nonexamples of an answer that would receive each score listed in the rubric. Both scorers were blind to each student's condition and to each other's scores. Once both scorers had scored the pretests, they met to discuss any disagreements in their scoring. The highest number of disagreements for a test question was 16 (out of 145 scores: 11.03%) with a total number of 133 disagreements (out of 1,450 scores: 9.17%) across all 10 test questions and students. There were 9 score disagreements (out of the 133 disagreements across all questions and students: 6.77%) that deviated by more than 1 point (7 deviated by 2 points and 2 deviated by 3 points). Typical disagreements in a score were due to the ambiguity of a student's response (e.g., incomplete sentences/spellings) or the difficulty in reading hand-writing. All scoring disagreements were resolved by the scorers discussing the rubric, the student's answer, and justification for the score they had given the response. Cohen's Kappa coefficients ranged from .90 (question 6) to .96 (question 10) with these values falling in the Landis-Koch (Landis & Koch, 1977) "almost perfect" category for agreement level.

Hands-on activity. All of the ramp setups and the accompanying student explanations were scored by the first author; research assistant B independently scored 56% of the ramp setups and explanations. A scoring rubric—developed by the authors and utilized by both scorers—provided criteria on how to (a) determine if the setup was unconfounded or confounded and (b) assign a score of 0 to 4 to a student explanation (see "Scoring Tests and Hands-On Activity" for explanations of what each score meant). The rubric included examples and nonexamples of explanations that would receive each score. Scorers listened to the student's explanation and wrote down their scores on individual scoring sheets. After the lesson, both scorers discussed any disagreements. The number of disagreements was 6 (out of 145 scores: 4.14%). There was 1 score disagreement (out of the 6 disagreements across all students: 16.67%) that deviated by 2 points. Typical disagreements in a score were due to the ambiguity of a student's response (e.g., poor articulation). All but two scoring disagreements were resolved by the scorers discussing the rubric, the student's answer, and justification for the score they had given the response. For the two unresolved scores, the final score was determined by the first author who consulted the scoring rubric. The Cohen's Kappa coefficient was .95 with this value falling in the Landis-Koch "almost perfect" category for agreement level.

Posttest. All posttests were scored independently by two scorers (the first author and research assistant A). The rubric provided information on how an explanation would qualify as receiving a score of 0 to 4 (see "Scoring Tests and Hands-On Activity" for score explanations). Both scorers were blind to each student's condition and to each other's scores. Once both scorers had scored the posttests, they met to discuss any disagreements. The highest number of disagreements for a test question was 14 (out of 145 scores: 9.66%) with a total number of 85 disagreements (out of 1,450 scores: 5.86%) across all 10 test questions and students. Fourteen of the score disagreements (out of the 85 disagreements across all questions and students: 16.47%) deviated by 2 points (there were no disagreements that differed by more than 2 points). All scoring disagreements were resolved by the scorers discussing the rubric, the student's answer, and justification for the score they had given the response. Typical disagreements in a score were due to the ambiguity of a student's response (e.g., incomplete sentences/spellings) or the difficulty in reading hand-writing. Cohen's Kappa coefficients ranged from .95 (question 3) to .99 (question 2) with these values falling in the Landis-Koch (1977) "almost perfect agreement" category.

Statistical Considerations

As previously stated, our conditions represent a continuum ranging from highly teacher-directed to highly student-directed. We chose the condition "minimal guidance/minimal guidance/activity" as our reference group due to it having the least amount of teacher support and requiring students to engage in more discovery-oriented processes. Thus, we wanted to compare our active learning conditions with higher levels of teacher support to the condition with the lowest level of teacher support. Dummy variables were created for each condition.

Including dummy variables in the statistical model to represent conditions is likely to improve the precision of the impact estimate.

Data were analyzed in SPSS 25 and STATA 15. A three-level Hierarchical Linear Model (HLM) was used to estimate condition impacts on student outcomes. HLM takes into account the nested structure of the data—students nested within groups that were clustered within classes—to estimate condition effects (Goldstein, 1987; Murray, 1998; Raudenbush & Bryk, 2002). A three-level hierarchical linear model (HLM) was constructed to estimate condition impacts. Level-1, the students level was specified by:

$$Y_{i:g:c} = \alpha_{0g:c} + \beta_{1g:c}Pre_{i:g:c} + \Sigma\beta_T T_i + \varepsilon_{i:g:c} \qquad \textbf{[1]}$$

Where subscripts i, g & c denote student, group and class, respectively; the nesting is reflected by the colons (:); Y represents student achievement in designing simple experiments; Pre represents the baseline measure of the outcome variable; T_i is a variable indicating student enrollment assigned to which condition. Lastly, μ represents a random variable for groups, and $\varepsilon_{i:g:c}$ is an error term for individual sample members. In this model, each condition effect compared to reference group is represented by β_T, which captures covariate-adjusted differences in the outcome variable. Level-2, the group level was specified by:

$$\alpha_{0g:c} = \gamma_{00c} + \gamma_{01c}(Group\_COV_{g:c}) + \mu_{0c} \qquad \textbf{[2]}$$

Where γ_{00c} is the adjusted grand mean of α_0 across groups in class c, $Group\_COV_{gc}$ is a matrix of group-level covariates and γ_{01c} is a vector of fixed effects in class c, and μ_{0c} is the group-level random effect for α_0. Level-3, the class level was specified by:

$$\gamma_{00c} = \pi_{000} + \pi_{01}(Class\_COV_c) + \varepsilon_{00c} \qquad \textbf{[3]}$$

Where π_{000} is the adjusted grand mean of γ_{00} across classes, Class_COV$_{gc}$ is a matrix of class-level covariates and π_{01} is a vector of corresponding fixed effects, and ε_{00c} is the class-level random effect for γ_{00c}.

In our study, student level fixed effects included assigned conditions and pretest; there were no group level or class level covariates included. The purpose of including statistical controls was to minimize random error and to increase the precision of the estimates. Even though our data was from three different schools, to avoid introducing more confounding factors, we considered the class differences instead of including a school level. Our analytic sample had 145 students nested in 36 groups, that were clustered in 8 classes. There were small sample sizes in each group, which is a limitation of our study. It might lead to biased estimates of the upper-level (level two and level three) standard errors. However, the upper-levels were not of interest to our study, we only have explicit interest in student level. Simulated conditions showed that the estimates of the regression coefficients, the variance components, and the standard errors at level one should be unbiased and accurate (Browne & Draper, 2006; Maas & Hox, 2005; McNeish & Wentzel, 2017).

Results

Analyses

The pre- and posttests were treated as parallel forms measuring CVS learning in novel domains (i.e., outside of the context of the ramps). These questions were chosen from a question bank used in previous CVS studies (e.g., Chen & Klahr, 1999) and were matched by wording and complexity. The Pearson Correlation for individual items between the pre- and posttest measures was 0.76, indicating a fair amount of rank-order stability across the parallel forms. Our analyses aimed to address the three research questions: (a) are the active learning implementations used in this study equivalent in their effects on student learning of CVS? (b) if there are differences in learning, which implementations are most effective in promoting students' acquisition and transfer of CVS? and (c) if there are differences, what specifically led one implementation to be more effective than another?

Acquisition of CVS (Dependent Variable 1) in a familiar domain (i.e., with the ramp materials) was assessed during the hands-on activity (Part 4) in each condition. There was no baseline knowledge measure for hands-on activity. HLM results (Table 3) showed Modeling/Direct Guidance/Activity effects was statistically significantly better (0.79) than Minimal Guidance/Minimal Guidance/Activity effects. Modeling/Modeling/Activity and Minimal Guidance/Direct Guidance/Activity were better than Minimal Guidance/Minimal Guidance/Activity as well, although not statistically significant. Table 3 revealed the following rank order of performance (from least to greatest) in hands-on activity: Minimal Guidance/Minimal Guidance/Activity < Minimal Guidance/Direct Guidance/Activity < Modeling/Modeling/Activity < Modeling/Direct Guidance/Activity.

To estimate how the amount of learning differs across the four conditions, Table 4 displays paired comparisons between all four conditions. There was only one pair of condition differences that were statistically significant: Modeling/Direct Guidance/Activity vs. Minimal Guidance/Minimal Guidance/Activity. The other paired differences between conditions were from 0.22 to 0.57, but neither reached statistical significance.

The learning of CVS (Dependent Variable 2) in novel domains was also assessed by looking at pre-post test scores. As the threat of attrition bias of this randomized controlled study is a tolerable threat, there is no need to check the baseline equivalence (What Works Clearinghouse Standards Handbook 4.0, 2017). Table 5 revealed significant learning gains from pretest to posttest in all conditions. Additionally, we were interested in examining which condition(s) produced the greatest increase in pre- to posttest performance. Using pretest as the baseline measure, impact analysis from the HLM model showed that both "Modeling/Direct Guidance/Activity" and "Modeling/Modeling/Activity" had large effect sizes on students' pre-posttests (1.25 and 1.16, respectively); "Minimal Guidance/Direct Guidance/Activity" had medium effects (0.44); Minimal Guidance/Minimal Guidance/Activity had a small effect size of 0.09 (see Figure 3). In examining which condition(s) produced the greatest standardized increase in pre- to posttest performance, the rank order of performance (from least to greatest) was: Minimal Guidance/Minimal Guidance/Activity < Minimal Guidance/Direct Guidance/Activity < Modeling/Modeling/Activity < Modeling/Direct Guidance/Activity, which is consistent with the order findings from the hands-on activity.

To estimate how the amount of learning differs across the four conditions, Table 6 displays paired comparisons between all four conditions. The difference between Modeling/Direct Guidance/Activity and Modeling/Modeling/Activity were quite close, and not statistically significant. Modeling/Direct Guidance/Activity was statistically significantly better than both Minimal Guidance/Direct Guidance/Activity and Minimal Guidance/Minimal Guidance/Activity. In the same way, Modeling/Modeling/Activity was statistically significantly better than both Minimal Guidance/Direct Guidance/Activity and Minimal Guidance/Minimal Guidance/Activity. Minimal Guidance/Direct Guidance/Activity was not statistically significantly better than Minimal Guidance/Minimal Guidance/Activity, though.

Discussion

The fundamental question that motivated this study was whether or not different active learning implementations would lead to differences in learning and performance. We focused our efforts on the first source of variation in active learning classrooms—the instructional techniques employed by teachers. Thus, active learning was implemented in a lesson that was structured according to four different instructional techniques. The lesson occurred in the context of teaching third and fourth graders a simple but important science topic: CVS. A secondary goal of our study was to determine which pedagogical elements of the different active learning implementations led one to be more effective than another. Our analyses revealed that when compared to Minimal Guidance/Minimal Guidance/Activity, Modeling/Direct Guidance/Activity resulted in significantly higher levels of CVS knowledge on the hands-on activity.

When examining student learning of CVS from pre- to posttest, students in all conditions had significant learning gains. However, the largest effect sizes were for Modeling/Direct Guidance/Activity followed by Modeling/Modeling/Activity and the weakest effect size was for Minimal Guidance/Minimal Guidance/Activity. Further, when looking at the paired comparisons between conditions, Modeling/Direct Guidance/Activity and Modeling/Modeling/Activity were statistically significantly better than both Minimal Guidance/Direct Guidance/Activity and Minimal Guidance/Minimal Guidance/Activity. Students thus had larger learning gains when assessed in a novel domain (i.e., the posttest) when instruction moved from more inquiry-based (i.e., Minimal Guidance/Minimal Guidance/Activity and Minimal Guidance/Direct Guidance/Activity) to more teacher-directed (i.e., Modeling/Direct Guidance/Activity and Modeling/Modeling/Activity). The two conditions that included at least one *modeling* instructional treatment (i.e., Modeling/Direct Guidance/Activity and Modeling/Modeling/Activity) provided the greatest learning of CVS. Direct guidance also seemed to be important as the strongest condition contained a *direct guidance* instructional treatment (i.e., Modeling/Direct Guidance/Activity) and the partially inquiry-based condition (i.e., Minimal Guidance/Direct Guidance/Activity) had a larger effect size than the full inquiry-based condition (i.e., Minimal Guidance/Minimal Guidance/Activity).

Many forms of inquiry-based learning have been criticized for being a “straw man” (Davis, Goulding, & Suggate, 2017) as they do not include any guidance and are not considered representative of many of the richer forms of inquiry-based learning (Herman & Gomez, 2009). As the argument goes, if students are simply told to explore without any support, they are unlikely to learn much. However, our more inquiry-based conditions (i.e., Minimal Guidance/Minimal Guidance/Activity and Minimal Guidance/Direct Guidance/Activity) included guidance and did lead to student learning. Thus, the minimal guidance instructional treatment in Minimal Guidance/Minimal Guidance/Activity and in Minimal Guidance/Direct Guidance/Activity was successful in helping students to acquire CVS knowledge.

Differences in CVS knowledge emerged both when students were tested on familiar (i.e., the hands-on activity involving the ramps) and novel domains (i.e., the posttest questions). Our results are consistent with past research examining inquiry-based approaches versus more explicit approaches (e.g., Alfieri, Brooks, Aldrich, & Tenenbaum, 2011), especially in that (a) the more teacher-directed approaches (which often include modeling and direct guidance) help students in novel contexts and more inquiry-based methods (which often include minimal guidance) are less effective in these contexts (see Klahr & Nigam, 2004). When active learning is combined with more teacher support and direction, we see greater learning gains than when active learning is combined with more inquiry-based methods. In fact, consistent with our findings, Lorch et al. (2010) found that interactive lecture combined with hands-on experimentation led to greater learning than hands-on experimentation without instruction.

The primary theoretical argument against inquiry-based learning approaches is that their processing demands exceed the limited capacity of working memory (Kirschner et al., 2006). These demands are particularly costly when learners are unable to retrieve previously learned information from long-term memory. In contrast, scaffolded approaches are designed to reduce students' cognitive load by (a) breaking information into small units (chunks), (b) sequencing these units in a logical progression, and (c) and having students master these units before moving to more complex units (Hughes, Morris, Therrien, & Benson, 2017).

Another interesting finding from our study is that providing minimal guidance *prior* to providing direct guidance did not lead to greater learning than when instruction was provided to students before they had the opportunity to practice and receive feedback. Although we did not design our study or include all of the necessary conditions to test "Productive Failure" (e.g., Kapur, 2015) or "Preparation for Future Learning" (e.g., Schwartz & Martin, 2004), our results are consistent with Matlen and Klahr (2012) in that our teacher-directed conditions (i.e., Modeling/Direct Guidance/Activity and Modeling/Modeling/Activity) were more effective than our inquiry-based conditions (i.e., Minimal Guidance/Minimal Guidance/Activity and Minimal Guidance/Direct Guidance/Activity) where exploration came first in the instructional sequence.

Our results are not consistent with the claim that students do better when they are engaged in interactive learning. According to the ICAP framework (Chi & Wylie, 2014), interaction among students or between teacher and student should lead to the greatest learning gains as compared to students working independently. One could argue that the Minimal Guidance/Minimal Guidance/Activity and Minimal Guidance/Direct Guidance/Activity conditions used in our study met the criteria for the "I" in ICAP (i.e., interactive) because "minimal guidance" is usually interpreted as situations in which students interact actively and constructively with the teacher through turn taking. However, our most effective conditions—Modeling/Direct Guidance/Activity and Modeling/Modeling/Activity—did not involve the teacher and students interacting in a constructive manner where "each speaker's utterances generate some knowledge beyond what was presented in the original learning materials and beyond what the partner has said" (Chi & Wylie, 2014, p. 223). One research question to ponder is whether or not *how* active learning is implemented (i.e., from a pedagogical standpoint) is more important to consider than *if* students are engaged on an active, constructive, or interactive level. In other words, should our focus when researching active learning be shifted from student behaviors to the pedagogical strategies employed by the instructor?

Through this study, as well as from the mixed results in the broader literature on active learning, it appears not all active learning approaches are created equal. The way in which the imprecise notion of "active learning" is operationally defined and implemented in specific instructional procedures can lead to significant differences in student learning, specifically when measuring students' transfer of knowledge to novel questions or settings. Most of the published studies on active learning compare it to passive learning (also known as the "traditional lecture method") (see Freeman et al., 2014). Our study suggests that a simple global contrast between the relative effectiveness of active learning versus lecture is too narrow; rather, we should be deconstructing active learning into its components, and then attempting to determine which combinations and implementations of those pedagogical elements are (a) more effective than lecture and (b) more effective than other active learning implementations. Additionally, active learning studies tend to be conducted at the college level but should also be utilized in K-12 settings.

Although there were several interesting findings in this investigation, there are several limitations to consider. First, CVS instruction for all students in all conditions was provided by a “domain expert” (the first author of this paper). The extent to which the results reported here would generalize to other teachers or topics is unknown, although we believe that by using our detailed instructional script and materials, high treatment fidelity could be achieved. Second, we assessed only near transfer. Future studies could be conducted on the various dimensions of transfer that have been described in the literature (Chen & Klahr, 2008; Klahr & Chen, 2011). Third, students were taught in small groups, whereas in many schools, teachers’ resources often constrain them to use only whole class instruction, and the relative effectiveness (and feasibility) of implementing the instructional conditions used in our study to an entire class at once remains to be determined. Fourth, we taught only one isolated (albeit very important) science process skill, CVS, and it is not clear what the effects of different active learning implementations would be if several skills were taught simultaneously. Fifth, students were engaged in one 40-minute lesson, leading to a question of whether or not greater learning would occur in certain conditions if students received additional time with CVS. However, our study does address a question of efficiency given our single lesson duration. Finally, we did not have a full factorial design of all possible combinations of the different instructional approaches. Future research could be conducted to determine if there are other ways to combine varying levels of instructor modeling, guidance, and feedback to improve student learning and if there are ordering effects where certain forms of instruction are better implemented first or second, for example, in a lesson. Active learning is a promising pedagogical method that would benefit from a more precise description and analysis of its essential features as well as a better understanding of the learning mechanisms evoked by specific features of active learning.

Acknowledgements

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grants R305B150008 and R305A170176 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. We thank the students, teachers, and administrators at Winchester-Thurston School, Propel Pitcairn Elementary School, and Propel Northside Elementary School.

References

Adams, A. E. M., Garcia, J., & Traustadottir, T. (2016). A quasi experiment to determine the effectiveness of a “partially flipped” versus “fully flipped” undergraduate class in genetics and evolution. *CBE-Life Sciences Education, 15*(2), 1–9. doi:10.1187/cbe.15-07-0157

Akinoglu, O., & Tandogan, R. O. (2007). The effects of problem-based active learning science education on students’ academic achievement, attitude and concept learning. *Eurasia Journal of Mathematics, Science and Technology Education, 3*(1), 71-81. doi:10.12973/ejmste/75375

Alfieri, L., Brooks, P. J., Aldrich, N. J., Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1-18. doi: 10.1037/a0021017

Archer, A. L., & Hughes, C. A. (2011). *Explicit instruction: Effective and efficient teaching.* New York, NY: Guilford Press.

Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom* [Monograph]. Retrieved from http://files.eric.ed.gov/fulltext/ED336049.pdf

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*(3), 687-699. doi 10.1177/001316448104100307

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1, 473–514.* doi:10.1214/06-ba117

Case, R. (1974). Structures and structures: Some functional limitations on the course of cognitive growth. *Cognitive Psychology*, *6*(4), 544-573. doi:10.1016/0010-0285(74)90025-5

Cattaneo, K. H. (2017). Telling active learning pedagogies apart: From theory to practice. *Journal of New Approaches in Educational Research*, *6*(2), 144-152. doi:10.7821/naer.2017.7.237

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120. doi:10.1111/1467-8624.00081

Chen, Z., & Klahr, D. (2008). Remote transfer of scientific reasoning and problem-solving strategies in children. In R. V. Kail (Ed.) *Advances in child development and behavior,* Vol. 36, (pp. 419-470). Amsterdam: Elsevier.

Chi, M. T. H., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M.,…Yaghmourian, D. L. (2018). Translating the ICAP theory of cognitive engagement into practice. *Cognitive Science, 42*(2018), 1777-1832. doi: 10.1111/cogs.12626

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219-243. doi: 10.1080/00461520.2014.965823

Clark, R. E., Kirschner, P. A., & Sweller, J. (2012). Putting students on the path to learning: The case for fully guided instruction. *American Educator, 36,* 6-11.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge.

Davis, A., Goulding, M., & Suggate, J. (2017). *Mathematical knowledge for primary teachers* (5th ed.). New York, NY: Routledge.

Dean, D., & Kuhn, D. (2006). Direct instruction vs. discovery: The long view. *Science Education, 91*(3), 384-397. doi: 10.1002/sce.20194

Eddy, S. L., & Hogan, K. A. (2014). Getting under the hood: How and for whom does increasing course structure work? *CBE-Life Sciences Education, 13*(3), 453–468. doi:10.1187/cbe.14-03-0050

Eichler, J. F., & Peeples, J. (2016). Flipped classroom modules for large enrollment general chemistry courses: A low barrier approach to increase active learning and improve student grades. *Chemistry Education Research and Practice*, *17*(1), 197-208. doi:10.1039/C5RP00159E

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Science*, *111*(23), 8410-8415. doi:10.1073/pnas.1319030111

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Oxford University Press.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*, 29-48. doi 10.1348/000711006X126600

Herman, P., & Gomez, L. M. (2009). Taking guided learning theory to school. In S. Tobias & T. M. Duffy (Eds.), *Constructivist instruction* (pp. 62-81). New York, NY: Routledge.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*, 99– 107

Hughes, C. A., Morris, J. R., Therrien, W. J., & Benson, S. K. (2017). Explicit instruction: Historical and contemporary contexts. *Learning Disabilities Research & Practice, 32*(3), 140-148. doi:10.1111/ldrp.12142

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science, 38*(5), 1008-1022. doi: 10.1111/cogs.12107

Kapur, M. (2015). Learning from productive failure. *Learning: Research and Practice*, *1*(1), 51-65. doi:10.1080/23735082.2015.1002195

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist, 51*(2), 289-299. doi: 10.1080/00461520.2016.1155457

Kinoshita, T. J., Knight, D. B., & Gibbes, B. (2017). The positive influence of active learning in a lecture hall: An analysis of normalized gain scores in introductory environmental engineering. *Innovations in Education and Teaching International*, *54*(3), 275-284. doi:10.1080/14703297.2015.1114957

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75-86. doi:10.1207/s15326985ep4102_1

Klahr, D. (2010). Coming up for air: but is it Oxygen or Phlogiston? A response to Taber's review of constructivist instruction: Success or failure? *Education Review, 13*(3), 1-6.

Klahr, D. (2013). What do we mean? On the importance of not abandoning scientific rigor when talking about science education. Proceedings of the National Academy of Science, 110, 14075-14080. doi:10.1073/pnas.1212738110

Klahr, D., & Chen, Z. (2011). Finding one's place in transfer space. Child Development Perspectives, 5(3), 196-204. doi:10.1111/j.1750-8606.2011.00171.x

Klahr, D., & Nigam, M. (2004). The equivalence of learning oaths in early science instruction: Effects of direct instruction and discovery learning. Psychological Science, 15, 661-667. doi:10.1111/j.0956-7976.2004.00737.x

Kolb, A. Y., & Kolb, D. A. (2005). Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of Management Learning & Education, 4*(2), 193-212. doi: 10.5465/amle.2005.17268566

Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability*. Retrieved from https://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers

Kuhn, D., Garcia-Mila, M., Zohar, A., Anderson, C., White, S. H., Klahr, D., & Carver, S. M. (1995). Strategies for knowledge acquisition. *Monographs of the Society for Research in Child Development, 60*(4), 1-160.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research, 86*(3), 681-718. doi: 10.3102/0034654315627366

Lin, X., & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching, 36*(7), 837-858. doi:10.1002/(SICI)1098-2736(199909)36:73.0.CO;2-U

Lorch, R. F., Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Dunham Freer, B. (2010). Learning the control of variable strategy in higher and lower achieving classrooms: Contributions of explicit instruction and experimentation. *Journal of Educational Psychology, 102*, 90-101. doi:10.1037/a0017972

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1, 86–92.* doi:10.1027/1614-2241.1.3.85

Marx, R. W., Blumenfeld, P. C., Krajcik, J. C., Fishman, B., Soloway, E., Geier, R., & Tal. R. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of Research in Science Teaching*, *41*(10), 1063-1080. doi:10.1002/tea.20039

Matlen, B. J., & Klahr, D. (2012). Sequential effects of high and low instructional guidance on children’s acquisition of experimentation skills: Is it all in the timing? Instructional Science, 41(3), 621-634. doi:10.1007/s11251-012-9248-z

McNeish D. & Wentzel, K.R. (2017). Accommodating Small Sample Sizes in Three-Level Models When the Third Level is Incidental. *Multivariate Behavioral Research, 52(2),* 200-215. doi: 10.1080/00273171.2016.1262236

Murray, D.M. (1998). *Design and analysis of group randomized trials*. New York: Oxford University Press.

Rau, M. A., Kennedy, K., Oxtoby, L., Bollom, M., & Moore, J. W. (2017). Unpacking “active learning”: A combination of flipped classroom and collaboration support is more effective but collaboration support alone is not. *Journal of Chemical Education, 94*(10), 1406–1414. doi:10.1021/acs.jchemed.7b00240

Raudenbush, S.W., and Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage: Newbury Park, CA.

Rosenshine, B. (2008). *Five meanings of direct instruction*. Lincoln, IL: Center on Innovation & Improvement. Retrieved from http://www.centerii.org/search/Resources/FiveDirectInstruct.pdf

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129-184. doi:10.1207/s1532690xci2202_1

Schwichow, M., Croker, S., Zimmerman, C., Hoffler, T., & Hartig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37-63. doi:10.1016/j.dr.2015.12.001

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321-325. doi 10.1086/266577

Stockard, J. & Wood, T. W. & Coughlin, C., & Khoury, C. R. (2018). The effectiveness of Direct Instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479-507. doi 10.3102/0034654317751919

Taraban, R. T., Box, C., Myers, R., Pollard, R., & Bowen, C. W. (2007). Effects of active-learning experiences on achievement, attitudes, and behaviors in high school biology. *Journal of Research in Science Teaching*, *44*(7), 960-979. doi 10.1002/tea.20183

Webb, D. J. (2017). Concepts first: A course with improved educational outcomes and parity for underrepresented minority groups. *American Journal of Physics*, *85*(8), 628–632. doi:10.1119/1.4991371

What Works Clearinghouse. (2017). *What Works Clearinghouse (WWC) standards handbook Version 4.0*. Retrieved from What Works Clearinghouse: https://ies.ed.gov/ncee/wwc/Handbooks

Wieman, C. E. (2014). Large-scale comparison of science teaching methods sends clear message. *Proceedings of the National Academy of Sciences*, *111*(23), 8319-8320. doi:10.1073/pnas.140730411

Table 1

Participant Demographics

Schools (N = 3)	
Free and reduced price lunch (measure of SES)	37.9%
Participants (N = 145)	
Mean age	9.1
Sex:	
Male	54.5%
Female	45.5%
Race/Ethnicity	
Caucasian	33.1%
African American	54.5%
Asian	4.8%
Other	7.6%

Table 2

Ramp Problem Domain

Primary materials	Two ramps, each with adjustable heights and "starting gate" locations Two golf balls Two two-sided surface inserts (for ramps) with different coefficients of friction	
To be determined	How to design an experiment that will let you know if a specific factor affects how far a ball will roll down a ramp	
Variables: 2 independent values for each of 3 variables	height	short, tall
	starting position gate	short, long
	surface	smooth, rough
Dependent measure	Distance ball rolls at end of ramp	
Subject activity Experimental design	For each of two ramps Select one of two heights One of two surfaces One of two starting positions	
Experiment execution	Release gates (not necessarily simultaneously), allowing balls to roll Observe distance balls roll after leaving ramp	
Notable aspects of domain and procedure	Variables are independent, object is constructed from choice of values for each variable Comparison objects are constructed; variable values are not clustered Outcome is evanescent (if based on speed), or stable (if based on final distance)	

Table 3

Impact Analysis of Hands-on Activity

Condition	N students	Mean	SD	Difference compared to reference condition	P Value
Modeling/Direct Guidance/Activity	36	3.23	2.39	0.79	.04*
Modeling/Modeling/Activity	35	2.92	2.40	0.48	.21
Minimal Guidance/Direct Guidance/Activity	37	2.66	2.44	0.22	.55
Minimal Guidance/Minimal Guidance/Activity (**reference condition**)	37	2.44	2.43	-	-

Note: * = significant at p < .05; ** = significant at p < .01; *** = significant at p <.001.

Table 4

Paired Differences Between Conditions for Hands-on Activity

Condition	N students	Mean	SD	Difference	P Value
Modeling/Direct Guidance/Activity vs. Modeling/Modeling/Activity					
Modeling/Direct Guidance/Activity	36	3.23	2.39	0.31	.41
Modeling/Modeling/ Activity	35	2.92	2.40	-	-
Modeling/Direct Guidance/Activity vs. Minimal Guidance/Direct Guidance/Activity					
Modeling/Direct Guidance/Activity	36	3.23	2.39	0.57	.14
Minimal Guidance/Direct Guidance/Activity	37	2.66	2.44	-	-
Modeling/Direct Guidance/Activity vs. Minimal Guidance/Minimal Guidance/Activity					
Modeling/Direct Guidance/Activity	36	3.23	2.39	0.79	.04*
Minimal Guidance/Minimal Guidance /Activity	37	2.44	2.43	-	-
Modeling/Modeling/Activity vs. Minimal Guidance/Direct Guidance/Activity					
Modeling/Modeling/ Activity	35	2.92	2.40	0.26	.51
Minimal Guidance/Direct Guidance/Activity	37	2.66	2.44	-	-
Modeling/Modeling/Activity vs. Minimal Guidance/Minimal Guidance/Activity					
Modeling/Modeling/ Activity	35	2.92	2.40	0.48	.21
Minimal Guidance/Minimal Guidance/Activity	37	2.44	2.43	-	-
Minimal Guidance/Direct Guidance/Activity vs. Minimal Guidance/Minimal Guidance/Activity					
Minimal Guidance/Direct Guidance/Activity	37	2.66	2.44	0.22	.55
Minimal Guidance/Minimal Guidance/Activity	37	2.44	2.43	-	-

Note: * = significant at p < .05; ** = significant at p < .01; *** = significant at p <.001.

Table 5

Impact Analysis of Pre-Post Tests

Condition	N students	Pretest Mean	Pretest SD	Model Adjusted Posttest Mean	Posttest SD	Post-Pre Difference	Standardized Difference Between Post-Pre	P Value
Modeling/Direct Guidance/Activity	36	10.94	9.55	22.83	13.28	11.89	1.25	.00***
Modeling/Modeling/ Activity	35	9.60	11.49	22.95	13.33	13.35	1.16	.00***
Minimal Guidance/Direct Guidance/Activity	37	11.62	12.70	17.16	13.58	5.54	0.44	.00**
Minimal Guidance/Minimal Guidance/Activity	37	12.97	13.57	14.15	13.53	1.18	0.09	.01*

Note: * = significant at p < .05; ** = significant at p < .01; *** = significant at p <.001.

Table 6

Paired Differences Between Conditions for Post-Pre Tests

Condition	Pretest mean	SD	Model adj. Posttest mean	SD	P Value
Modeling/Direct Guidance/Activity vs. Modeling/Modeling/Activity					
Modeling/Direct Guidance/Activity	10.94	9.55	22.83	13.28	.96
Modeling/Modeling/ Activity	9.60	11.49	22.95	13.33	-
Modeling/Direct Guidance/Activity vs. Minimal Guidance/Direct Guidance/Activity					
Modeling/Direct Guidance/Activity	10.94	9.55	22.83	13.28	.01*
Minimal Guidance/Direct Guidance/Activity	11.62	12.70	17.16	13.58	-
Modeling/Direct Guidance/Activity vs. Minimal Guidance/Minimal Guidance/Activity					
Modeling/Direct Guidance/Activity	10.94	9.55	22.83	13.28	.00***
Minimal Guidance/Minimal Guidance /Activity	12.97	13.57	14.15	13.53	-
Model/Model/Activity vs. Direct Guidance/Minimal/Activity					
Modeling/Modeling/ Activity	9.60	11.49	22.95	13.33	.01**
Minimal Guidance/Direct Guidance/Activity	11.62	12.70	17.16	13.58	-
Modeling/Modeling/Activity vs. Minimal Guidance/Minimal Guidance/Activity					
Modeling/Modeling/ Activity	9.60	11.49	22.95	13.33	.00***
Minimal Guidance/Minimal Guidance/Activity	12.97	13.57	14.15	13.53	-
Minimal Guidance/Direct Guidance/Activity vs. Minimal Guidance/Minimal Guidance/Activity					
Minimal Guidance/Direct Guidance/Activity	11.62	12.70	17.16	13.58	.17
Minimal Guidance/Minimal Guidance/Activity	12.97	13.57	14.15	13.53	-

Note: * = significant at p < .05; ** = significant at p < .01; *** = significant at p <.001.

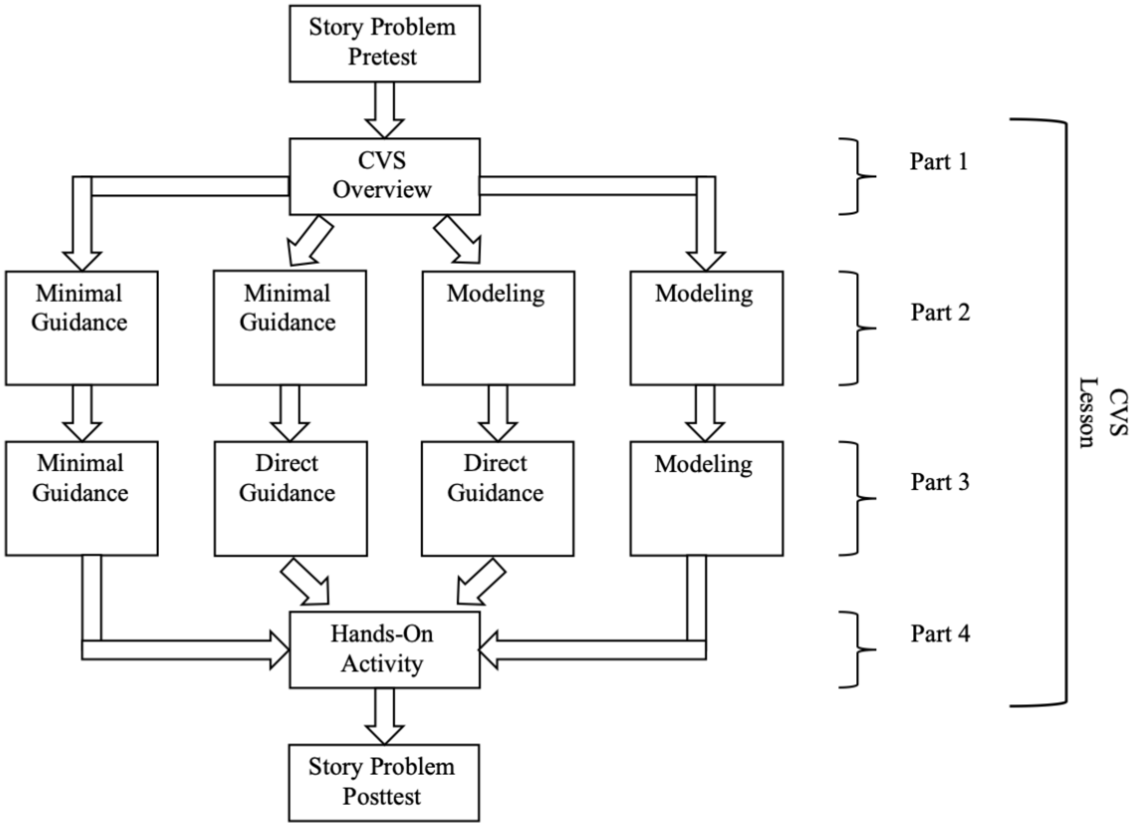
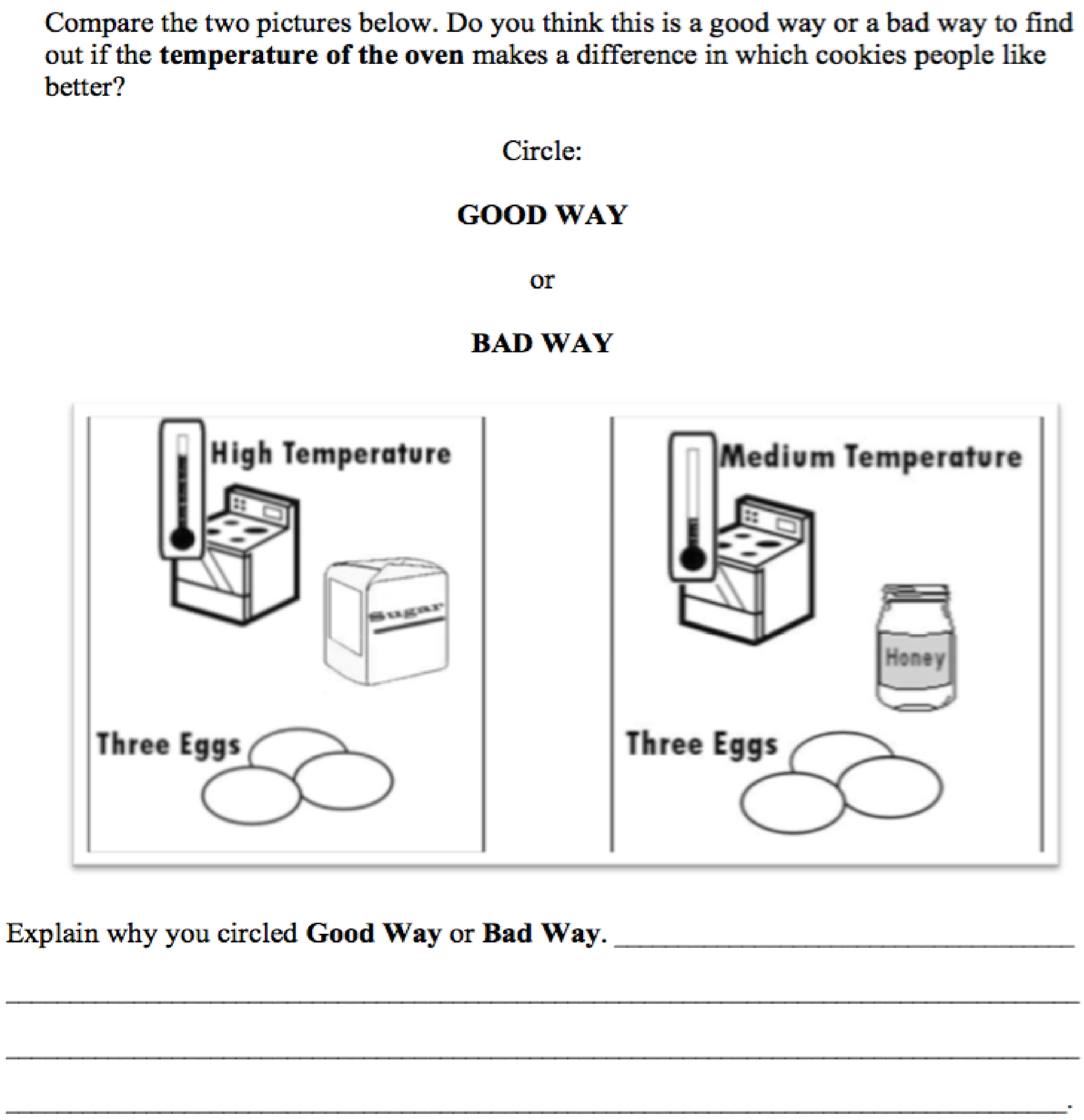


Figure 1. The main elements of the study, with students receiving a pretest, a CVS lesson with 4 parts (Parts 2 and 3 distinguished one condition from another), and a posttest.



Compare the two pictures below. Do you think this is a good way or a bad way to find out if the **temperature of the oven** makes a difference in which cookies people like better?

Circle:

GOOD WAY

or

BAD WAY

High Temperature
Sugar
Three Eggs
Medium Temperature
Honey
Three Eggs

Explain why you circled **Good Way** or **Bad Way**. ______________________________

____

____

____.

Figure 2. One of the 10 questions on the pretest.

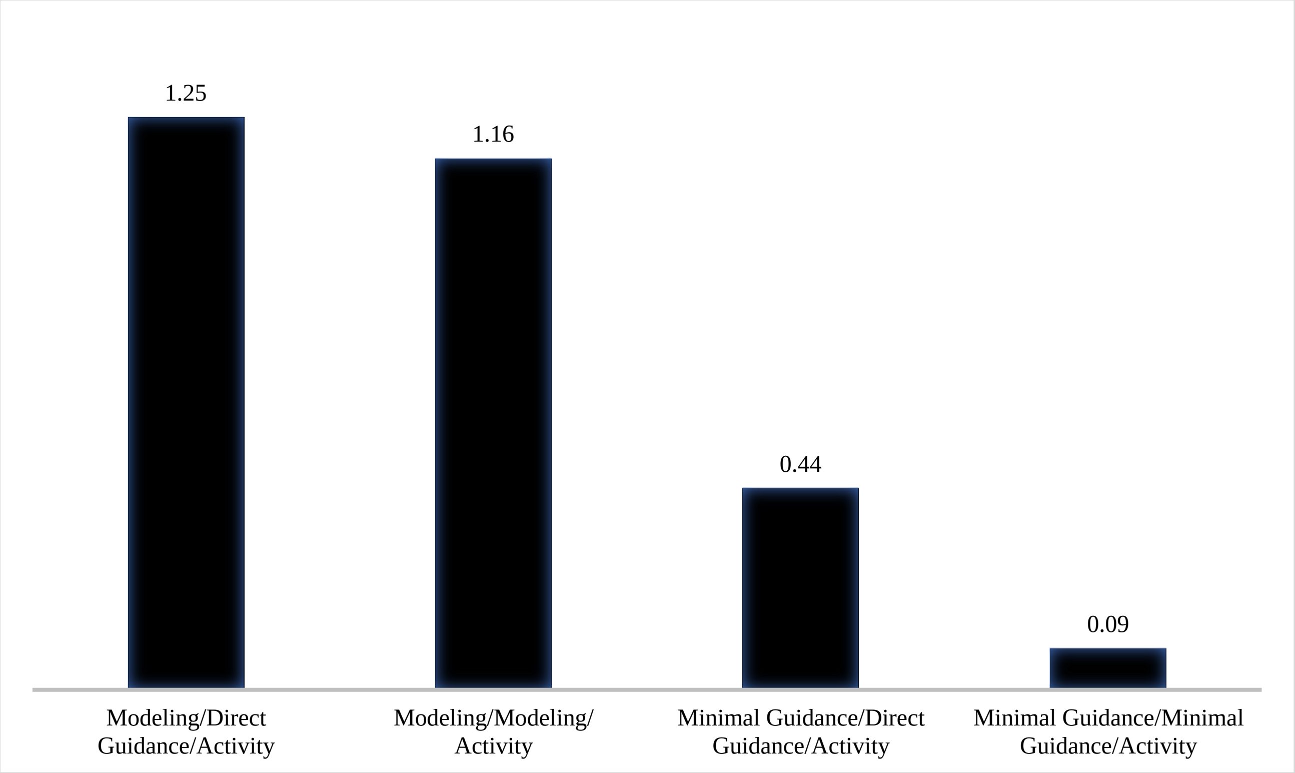


Figure 3. Effect sizes of Post-Pre tests for the four conditions.