

## Hands on What? The Relative Effectiveness of Physical Versus Virtual Materials in an Engineering Design Project by Middle School Children

David Klahr,<sup>1</sup> Lara M. Triona,<sup>2</sup> Cameron Williams<sup>1</sup>

<sup>1</sup>*Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213*

<sup>2</sup>*Department of Psychology, University of California, Santa Cruz, California 95065*

*Received 7 December 2005; Accepted 31 March 2006*

**Abstract:** “Hands-on” activities play an important, but controversial, role in early science education. In this study we attempt to clarify some of the issues surrounding the controversy by calling attention to distinctions between: (a) type of instruction (direct or discovery); (b) type of knowledge to be acquired (domain-general or domain-specific); and (c) type of materials that are used (physical or virtual). We then describe an empirical study that investigates the relative effectiveness of the physical–virtual dimension. In the present study, seventh and eighth grade students assembled and tested mousetrap cars with the goal of designing a car that would go the farthest. Children were assigned to four different conditions, depending on whether they manipulated physical or virtual materials, and whether they had a fixed number of cars they could construct or a fixed amount of time in which to construct them. All four conditions were equally effective in producing significant gains in learners’ knowledge about causal factors, in their ability to design optimal cars, and in their confidence in their knowledge. Girls’ performance, knowledge, and effort were equal to boys’ in all conditions, but girls’ confidence remained below boys’ throughout. Given the fact that, on several different measures, children were able to learn as well with virtual as with physical materials, the inherent pragmatic advantages of virtual materials in science may make them the preferred instructional medium in many hands-on contexts. © 2006 Wiley Periodicals, Inc. *J Res Sci Teach*

The merits and definitions of “hands-on” activities in early science education have been debated for over a century (Deboer, 1991; Huxley, 1899). Advocates for hands-on science argue that it promotes learning because it is consistent with the concrete-to-abstract nature of cognitive development, because it provides additional sources of brain activation via kinesthetic involvement, and because its intrinsic interest increases motivation and engagement (Flick, 1993; Haury & Rillero, 1994). Such arguments have led the National Science Teachers Association to recommend that the minimal amount of science instruction time devoted to hands-on

---

Contract grant sponsor: National Science Foundation; Contract grant number: BCS-0132315.

Correspondence to: D. Klahr; E-mail: klahr@cmu.edu

DOI 10.1002/tea.20152

Published online in Wiley InterScience (www.interscience.wiley.com).

activities should be at least 60% in elementary school, 80% in middle school, and 40% in high school (NSTA, 1990). Critics of hands-on activities argue that they make learning less efficient and effective by producing confusing and inconsistent feedback, by allowing learners to engage in off-task activities that produce irrelevant information, and by providing inadequate mappings between the behavior of physical materials and their abstract representation in formal diagrams and equations. Moreover, hands-on instruction tends to have higher logistical, financial, and temporal costs when compared with other approaches to science instruction (Hodson, 1996).

Both empirical and definitional factors contribute to the longevity and intensity of this debate, which, at times, can become contentious, as in the recent deliberations by the California State Board of Education about whether to put specific constraints on the proportion of instructional time allocated to hands-on activities (California State Teachers Association, 2004; Galley, 2004). The empirical problem is that the few carefully controlled experimental studies examining this issue have measured outcomes at a high level of aggregation, contrasting hands-on and hands-off curricula at the classroom or school level (Ruby, 2001), with unavoidable covariation and somewhat inconclusive results. The definitional problem is that “hands-on science” is sometimes used to describe a broad educational philosophy and, at other times, to refer to a specific instructional practice. As Flick (1993) noted:

There are two ways that we find the term “hands-on science” in common use today. The first uses “hands-on science” to refer to a general approach to instruction. . . . as a philosophy guiding when and how to use the broad range of teaching strategies needed to address diversity in contemporary classrooms. . . . The second way “hands-on science” is commonly used is in terms of a specific instructional strategy where students are actively engaged in manipulating materials, usually called a “hands-on science activity.” In this respect, “hands-on-science” can be found as part of other general approaches to instructional design that are not necessarily based on constructivist philosophy. (p. 9)

### Key Dimensions of Hands-On Activities

In this study, we use the term “hands-on” in the second sense, as a particular type of activity that could be consistent with a variety of educational philosophies. Even within this more focused sense, it is useful to identify some key dimensions of hands-on instruction and, insofar as possible, to study them in isolation. We suggest three important aspects worth considering when assessing hands-on instruction:

1. Whether the learner’s hands are on *physical* or *virtual* materials. This is an important factor because computers may provide a unique opportunity for “hands-on” activities with virtual materials that avoid many of the disadvantages of physical hands-on materials.
2. Whether the type of scientific knowledge that is being learned is *domain-general* or *domain-specific*.
3. Where, on the broad continuum from *discovery learning* to *direct instruction*, the pedagogy associated with the hands-on activity lies.

This three-dimensional space of different types of hands-on science is depicted in Table 1. In addition, Table 1 includes a row for “hands-off” instruction (lectures, reading, discussions) in which the physical–virtual dimension is not relevant, and the discovery learning feature is unlikely (indicated by parentheses). In the following paragraphs we describe each of the dimensions in Table 1.

Table 1  
*Space of some potential contrasts in studies of science activities*

	Instructional Goal			
	Domain-General Knowledge		Domain-Specific Knowledge	
	Direct Instruction	Discovery Learning	Direct Instruction	Discovery Learning
Hands-on materials				
Physical	A	B	C	D
Virtual	E	F	G	H
Hands-off	I	(J)	K	(L)

See text for explanation of cell lettering scheme.

### *Instructional Materials: Physical or Virtual?*

The learner's hands could be on physical materials or virtual materials. We use "physical" in the conventional sense, to mean real materials such as ramps, test tubes, plants, mechanical devices, chemicals, instruments, and electrical components typically found in science kits. By virtual materials we mean computer programs under control of mouse and keyboard that display and enact—on the computer screen—animations or videos that depict the same range of events that occur when the physical materials are used in the real world. In both physical and virtual situations, children's hands remain active and in control of the materials under investigation.

Surprisingly little is known about the instructional consequences of this physical–virtual distinction because almost all of the empirical investigations of the relative effectiveness of computer-based versus non-computer-based science instruction include differences in addition to the instructional medium being compared. Given that such contrasts are intentionally confounded—because the virtual versions facilitate changes not only to medium but also to instructional content and process—it is impossible to attribute differences in outcomes solely to the effects of instructional medium.

There are a few exceptions to this tendency to confound changes in medium with curricular changes, such as the study by Johnson-Gentile, Clements, and Battista (1994) comparing the effectiveness of physical manipulatives and LOGO programs for teaching geometric motion and Marshall's (2005) comparison of physical and virtual versions of the balance beam task. Further discussion of this issue can be found in a study by Triona and Klahr (2003) and in Clements' (1999) review of physical versus computer manipulatives in early mathematics instruction. The study to be described in this paper examines the instructional effects of physical versus virtual materials while holding constant all other likely causal factors.

### *Instructional Goal: Domain-General or Domain-Specific Knowledge?*

The second important distinction is between domain-general versus domain-specific knowledge. Domain-general knowledge, often referred to as "process skill," includes knowledge that transcends any specific branch of science, such as knowledge about the relation between theory and evidence (Kuhn, 2002), or about the procedural and conceptual basis for the design and interpretation of experiments (Chen & Klahr, 1999; Masnick & Klahr, 2003). Domain-specific knowledge includes facts and understanding about particular domains, such as physics, chemistry, or ecology. Our study focuses on domain-specific knowledge. This distinction has been discussed extensively in the literature on cognitive development (e.g., Zimmerman, 2000) and science education (e.g., Lehrer & Schauble, 2006).

*Type of Instruction: Direct or Discovery?*

The third contrast depicted in Table 1 revolves around the instructional context in which the materials are being used: either as part of direct instruction about particular concepts and procedures, or in a discovery mode in which little explicit instruction is provided. That is, instruction using physical or virtual materials, and focusing on domain-general or domain-specific knowledge, can be located anywhere along the direct-to-discovery spectrum. In the present study, all children, whether in the virtual or physical condition, were in “discovery” mode.

The taxonomy in Table 1 reveals how easy it is to conduct a confounded comparison between, say, hands-on and hands-off instruction, or between physical or virtual materials. For example, a comparison between Cell D (physical materials, domain-specific knowledge, and discovery learning), and the hands-off condition in Cell I would make it impossible to attribute any learning differences to just the “hands-on” versus “hands-off” distinction because other potentially causal features of instruction (type of knowledge, and type of instruction) had also been changed.

In some cases, the simultaneous contrast between several potentially causal features is intentional, as when researchers take a “design approach” rather than a simple experimental comparison. For example, Hmelo, Holton, and Kolodner (2000), compared children’s acquisition of knowledge about the human respiratory system in either a hands-on discovery context with physical materials (Cell D in Table 1) or a conventional lecture, discussion, textbook context (Cell K in Table 1). They found that the students in the design condition learned more. Because their focus was on the nuances and fine structure of the design condition, they did not attempt to isolate other potentially causal variables in the differences in learning gains between the two groups.

In other cases, potential confounds of this sort appear to have been unnoticed in the conclusions reached. For example, NAEP (2000) reported knowledge gains for children who responded affirmatively to the question: “Have you ever done hands-on activities or projects in school with chemicals (e.g., mixing or dissolving sugar or salt in water)?” Responses to this question are typically interpreted as if the hands-on activity took place in a highly nondirective discovery context (Cell D). However, there is no way to estimate the proportion of these recalled hands-on activities that occurred under a high degree of teacher control.

A few studies have attempted to isolate the factors included in Table 1. For example, Cells A and B were compared in a series of investigations by Klahr and colleagues (Chen & Klahr, 1999; Klahr & Nigam, 2004; Toth, Klahr, & Chen, 2000) using physical hands-on materials (physical ramps, springs, and sinking objects). The results of these studies demonstrated the relative effectiveness of direct instruction over discovery learning when teaching the “control of variables strategy” (CVS), a domain-general procedure for designing unconfounded experiments. Triona and Klahr (2003) compared physical materials (springs and weights) and virtual materials (computer depictions of the same materials) in the context of direct instruction about CVS (contrasting Cells A and E). Their results indicated that physical and virtual instructional materials were equally effective in achieving several instructional objectives, including children’s ability to design, justify, and derive correct predictions from unconfounded experiments. In the present study, we also focus entirely on the physical–virtual distinction—comparing children’s learning in Cells D and H—so as to provide another contrast that may contribute to our understanding of its different aspects. We describe an experiment that compares the effects of physical versus virtual materials on how well children acquire domain-specific knowledge while in discovery learning mode. We presented children with an engineering design challenge in which they had to produce an optimal design for a toy car (details to follow). Children constructed a series of cars with different features, and then observed their performance. No instruction was provided about how to

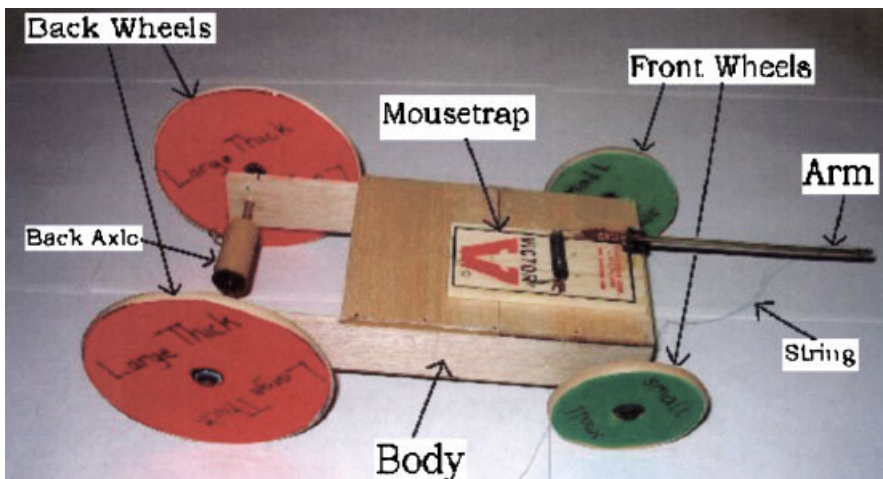
approach the comparisons between one design and another beyond explaining how to construct specific instances and the intended goal of the investigation.

### Designing and Testing Physical and Virtual Mousetrap Cars

Seventh and eighth grade children engaged in an engineering design task in which they created and tested a series of “mousetrap cars”: small mobile cars powered by an ordinary mousetrap that can travel dozens of feet (see Fig. 1). The children’s challenge was to discover the combination of features that yielded an optimal design for the car that could travel the farthest.

Mousetrap cars are used extensively in middle school science labs and science fairs, and a rich supply of kits, manuals, and teacher guides for using these materials is readily available, as a casual search of “mousetrap cars” on the internet will reveal. Their ultimate purpose is to provide a highly motivating context in which students can learn about conservation of energy, torque, friction, and mechanical advantage. Such lessons typically start with a simple challenge to construct cars that go far (as in our study). This focus on causal features is then used to motivate the second, and more theoretically important, questions about underlying physical laws that determine the behavior of the cars. However, it is precisely in this first stage about what features cause which effects that virtual–physical differences might be expected to have their greatest influence. Therefore, in this study, we focus on the initial construction stage in which children attempt to “engineer” high-performing cars.

One group of children worked with physical cars of the type shown in Figure 1. They selected various components (Fig. 2), assembled cars from them, and then ran the cars to see how far they would go. The other group worked with virtual cars of the type shown in Figure 3. They used a computer interface that allowed them to “point and click” to select components, assemble cars, and then “run” them in a virtual window. Before and after building several cars in one mode or the other (physical or virtual) children were assessed on their knowledge about how different values of individual features contributed to the car’s performance and how to combine those features into a car that could go the farthest.



*Figure 1.* A fully assembled physical mousetrap car with a short body, large, thick back wheels on a thick back axle, and small front wheels. The car is shown before the string is wound around the back axle and the mousetrap has been “set.” See text for detailed explanation.

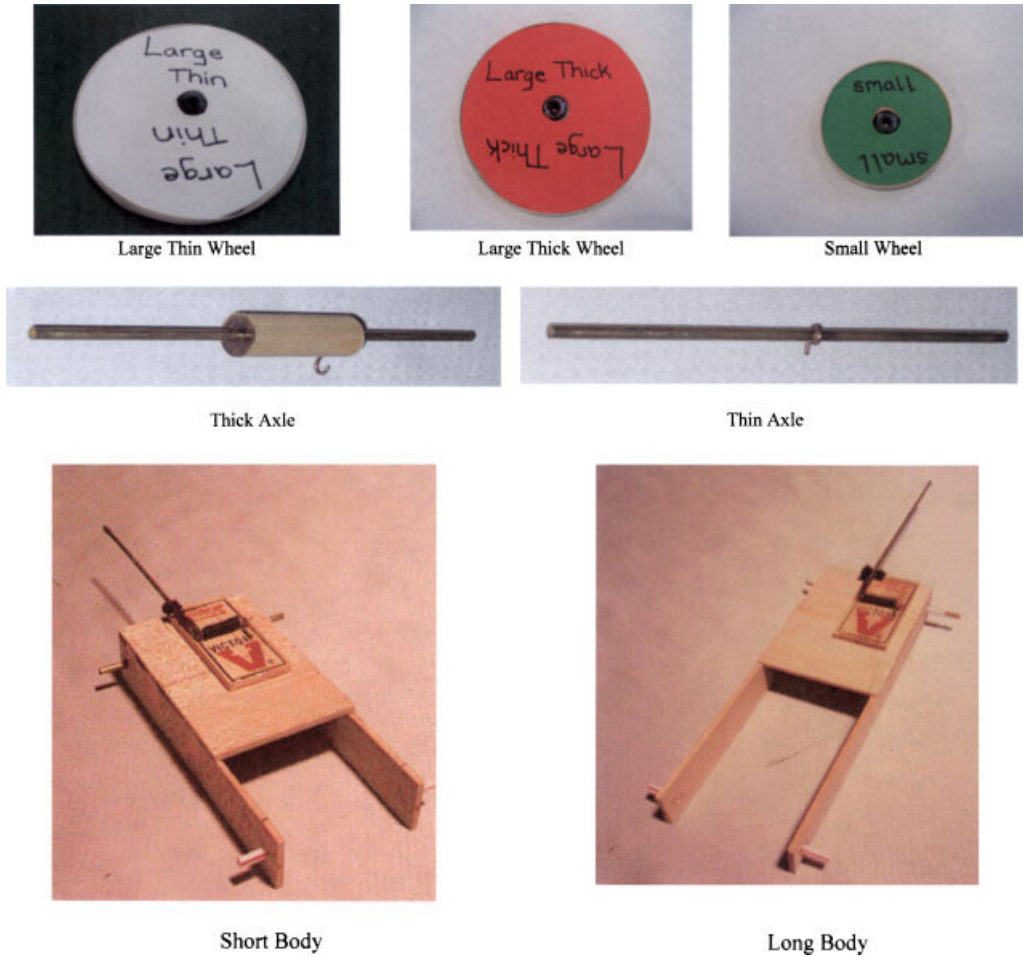
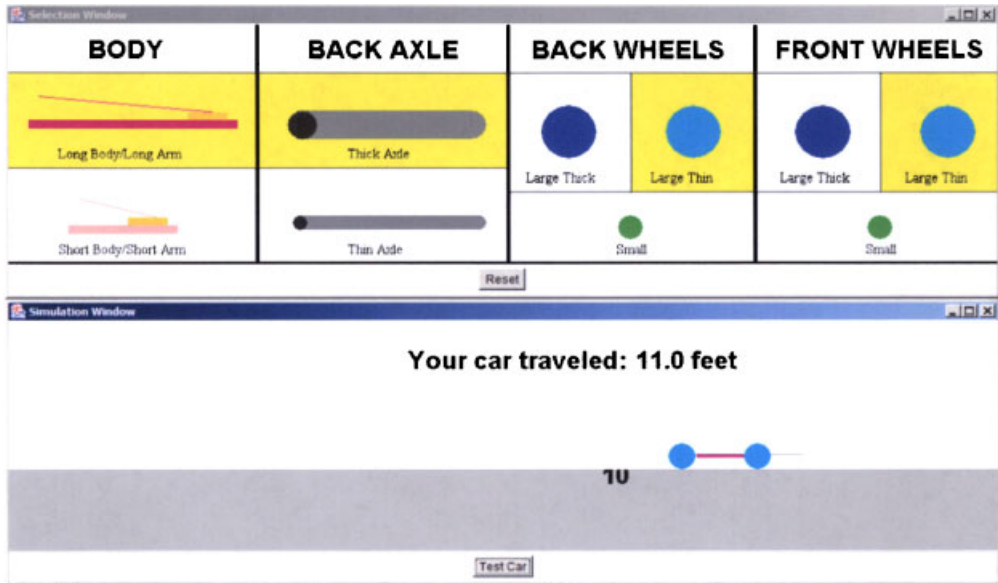


Figure 2. Components of physical mousetrap cars (materials were purchased from a commercial website that sells mousetrap cars, manuals, and related information: [www.docfizzix.com](http://www.docfizzix.com)).

The study extends the work of Triona and Klahr (2003) in several important ways. Triona and Klahr found that virtual and physical materials were equally effective in producing significant gains in children’s ability to design unconfounded experiments and to learn about the properties of springs from those experiments. Several factors may have contributed to this unanticipated finding of equivalence between physical and virtual materials.

First, the primary instructional objective for Triona and Klahr (2003) was for children to learn a domain-general procedure—the control of variables strategy (CVS), rather than about specific causal factors in the domains of ramps or springs. Although mastery of CVS would ultimately enable students to acquire such causal information, the main goal was to acquire the procedural and conceptual knowledge associated with CVS. Thus, the children’s ability to learn may have been relatively insensitive to the instructional medium because the procedural and conceptual basis of CVS is not tied to the physical attributes of *any* specific domain. In contrast, because the instructional goal in the current study—to learn the features of “good distance” cars—is highly domain-specific, the potential impact of physical versus virtual materials may be more discernible.



*Figure 3.* A screen shot of the virtual mousetrap car display. In this example, the highlighted panels at the top indicate that the student has constructed a car with a long body, thick back axle, large thin back wheels, and large thin front wheels. The bottom panel indicates that the car has completed its animated “run” across the bottom of the screen, and traveled 11 feet.

Second, the use of effective instruction by Triona and Klahr may have overwhelmed the potential impact of differences in instructional materials. In the current study, children were in “discovery mode.” That is, they were not provided with any instruction beyond a general explanation about how to assemble and run a mousetrap car.

Third, the information that children extracted from their physical manipulation of springs and weights in the Triona and Klahr study may not have been particularly salient. Although the physical materials did provide tactile information that was absent in the virtual condition, the length, width, and wire size of the various springs were equally apparent in both physical and virtual conditions, as were the outcomes of experiments (in one case, observing springs stretch to a varying extent and, in the other, observing videos of those very same events on a computer monitor, with similar visual resolution). In contrast, the physical and virtual conditions in the current study differ substantially in both the visual and the tactile information that they provide because the virtual materials depict the mousetrap cars as cartoon-like animations of simple two-dimensional line drawings rather in the form of photographs and video images. This provides a more extreme contrast between physical and virtual materials.

Finally, the children in the Triona and Klahr study were third and fourth graders, and the potential extra information extracted from the physical world may have exceeded their cognitive capacity. The increased capacity of the seventh and eighth graders in the present study might facilitate the recognition and integration of this kind of additional tactile and sensory–motor information.

In summary, the goal of the present study is to look at a physical–virtual contrast in hands-on science instruction, but under several different conditions from the Triona and Klahr (2003) study, all of which were chosen so as to enrich our understanding of the conditions under which physical–virtual differences might emerge. Instead of direct instruction, we used a discovery context. Instead of third and fourth graders, we used seventh and eighth graders. Instead of

teaching children domain-general procedural knowledge about how to design unconfounded experiments, the instructional goal was the discovery of domain-specific knowledge: the components of the best “distance” car. Instead of asking children to engage in simple actions such as choosing a pair of springs and a pair of weights and hanging them on a rack to observe their stretching, we asked children to explore an inherently interesting and enjoyable domain in which they had to assemble a series of (physical or virtual) multifeatured cars and run them (either across a screen or down the corridor in their school), and then note the distance they traveled.

In addition to all of these changes, we also investigated two factors that might be expected to interact with the virtual–physical distinction: time and gender. Physical mousetrap cars take several minutes to set up and test because some dexterity is needed to manipulate and assemble their components. In contrast, virtual experimentation requires only a few seconds worth of points, clicks, and drags to assemble and test a car. Thus, each virtual trial takes much less time than a corresponding physical trial, so that children in the virtual condition can run many more cars in a fixed amount of time than can children in the physical condition. To examine the learning effects of this difference we compared the effects of providing a constraint on either the amount of time children had or the number of cars they could construct and test.

Although none of the work just cited reported any gender effects on learning rates, knowledge, or transfer, it is well known that around the middle school grades, girls become less interested in, and less confident about, science, even when they may know more about it than boys (Helgeson, 2005). Although this effect is typically exacerbated when girls are asked to estimate their performance on “male tasks” (Beyer, 1990) such as our mousetrap car assembly task, it is possible that virtual presentation might ameliorate this gender effect, so we included gender as a factor in several of our secondary analyses.

We sought to determine whether the Triona and Klahr findings of no virtual–physical difference in hands-on science would be replicated with an older age group, a new class of instructional objectives, and a new learning context, and whether any of these effects were gender-dependent.

## Method

### *Participants*

Participants included 56 seventh and eighth graders (20 girls and 36 boys;  $M = 13.1$  years,  $SD = 0.69$  years) from two private middle schools in an urban area of southwestern Pennsylvania. Participants were recruited with notices sent to parents explaining that we would be assessing children’s performance on a task that was not a part of their normal science instruction. Children were randomly assigned to one of the four conditions (see Design section).

### *Materials*

*Physical mousetrap cars.* A fully assembled physical mousetrap car is shown in Figure 1 and its components are shown in isolation in Figure 2. At the center of every car there is a body, attached to which is a conventional mousetrap with a long lever arm rigidly attached to the “business end” of the mousetrap. Attached to the end of the lever is a string with a loop at its free end. Metal rods serve as the front and back axles, and the back axle has a small hook that engages the loop attached to the string on the lever arm. The back and front wheels are fitted to the ends of the axles. Each car was assembled by choosing from two different bodies (short or long), two different back axles (thin or thick), three different back wheels (large thick, large thin, or small), and three different front wheels (large thick, large thin, or small). Thus, the design space consists of 36 distinct cars. Further specification of the physical cars is shown in Appendix A.

After all components had been chosen and assembled, the car was “energized” by attaching the loop at the free end of the string to the hook on the back axle, and slowly pulling the arm toward the back axle while winding the string around it, thereby moving the mousetrap to its “armed” position. Once the arm was fully wound (i.e., rotated 180 degrees from its resting position to its armed position), the car was placed at the start line, and released. As the mousetrap spring returned the lever arm to its initial position, the string rotated the axle, which in turn propelled the car forward down a corridor until the string was completely unwound from the axle (but still connected to the hook), at which point the car stopped. (In the few cases where the string came off the hook at the end of the run, and the car “free-rolled,” the trial was not recorded, and was repeated.) A tape measure extended along the side of the corridor enabled the child to note the distance that the car traveled (within an inch or so of accuracy). The distance traveled by the cars constructed in this study ranged from approximately 20 to approximately 40 feet.

*Virtual mousetrap cars.* A computer program was developed to create a virtual world for the assembly and testing of mousetrap cars. The display consists of two windows, one above the other (see Fig. 3). The upper window has four panels, one for each part (body, back axle, back wheels, and front wheels). Each panel is divided into different sections, which contained the choices for that part. For example, the “body” panel is divided into two sections: one for a long body and one for a short body. Each section contains a simple line drawing of the part and a corresponding label.

Children assembled virtual cars by clicking on one of the sections for each part. Once a part was selected, the background of the section for that part turned yellow and the child could not choose another part in that category unless they clicked on the “reset” button (located at the bottom of that window.) After all of the parts were selected, the assembled car appeared—as a simple line drawing of the car—at the left side of the bottom window. The child could then click the “test car” button, which would cause the car to move from left to right across the screen at a constant rate until it neared the right side of the screen. At this point, the car would stay in place and the screen would appear to scroll to the left, and numbers would appear below the car, indicating the distance it had traveled, to continue the illusion of motion. When the car finished running, the distance traveled appeared at the top of the window. The displayed virtual distances corresponded to how far the physical version of that car would have traveled. (These distances were determined by averaging several runs of the equivalent physical car. Thus, children in the physical and virtual conditions received similar feedback about how far a particular design would travel.) Then the child pressed the reset button to build another car.

*Knowledge assessment questionnaire.* Before children assembled and tested any cars, they completed a knowledge assessment questionnaire. After all cars had been assembled and tested, a nearly identical knowledge assessment questionnaire was also administered. These assessments were used to measure changes in children’s knowledge about the features that contributed to the distance a car would travel. Questions 1 through 7 on the pre- and posttest versions of the questionnaire were identical multiple-choice items (see Appendix B). They asked which body length, back axle width, back wheel size, back wheel thickness, front wheel size, and front wheel thickness would make a car travel farther, or whether that factor had no effect. In addition, after responding to each question, children indicated their confidence in their answer on a 5-point scale. The seventh question contained a table with categories for body length, axle width, back wheels, and front wheels and all the parts the children could choose for each category. Children were asked to circle the part from each category that they thought would—in combination—produce a car that would travel the greatest distance. The pre- and posttest differed by one additional question: the pretest asked children whether they had any previous experience with mousetrap cars and the posttest had a final open-ended question asking if there was anything else, other than the parts that had been selected, that might help a car travel farther.

*Datasheet.* In each condition, children recorded the features of their cars and the distance they traveled on a datasheet, containing a table with column headings for car number, body length, back axle, back wheels, front wheels, and distance (see Appendix C). In each row of the table, children could describe the car that they built by recording the parts they chose in the appropriate columns and recording the distance the car traveled in the “distance” column. All datasheets had one row for the sample car. The datasheet for the fixed-number-of-cars condition (see Design section for description of conditions) had six rows and the datasheet for the fixed-time condition had many more rows.

### *Design*

We used a 2 (material: physical vs. virtual)  $\times$  2 (constraint: fixed amount of time vs. fixed number of cars)  $\times$  2 (test phase: pretest vs. posttest) factorial design with test phase as a within-participant factor. Between the pretest and posttest, children had the opportunity to build and test several different mousetrap cars.

*Material factor.* In the physical materials condition, children used real mousetrap car parts constructed from the materials described earlier and ran the cars in a hallway in their school. In the virtual materials condition, children used the computer program described earlier to assemble the parts into complete cars and then “ran” the cars by watching them travel across the screen.

*Constraint factor.* In the fixed-time condition, all children had 20 minutes to build, run, and record the results from as many cars as possible. In this condition, children using virtual materials were expected to complete many more trials than children in the physical condition. In the fixed-number condition, all children had as much time as they needed to build and test exactly six different cars. We set the constraint at six cars because pilot work indicated that this was approximately how many physical cars could be built in 20 minutes, so we expected children in the physical condition to take about 20 minutes to complete the fixed-number condition, whereas children in the virtual condition were expected to take much less time.

### *Procedure*

*Set-up and pretest.* Children were tested individually in a quiet room in their school, adjacent to a corridor where they could run their cars (if they were in the physical car condition). All children, regardless of condition, were shown a mousetrap car assembled from physical parts. The experimenter then pointed to and named each part of the car and demonstrated how the car worked. Once the children understood the parts of the mousetrap car and how it worked they were told whether they would be building the cars using the physical parts or using the computer and whether they would have 20 minutes to build the cars or if they were to build exactly six cars.

Next, the children were shown all the parts that they would be using to construct the cars (physical parts were shown regardless of whether children were in the physical or virtual condition), and they were reminded that their goal was to find what combination of the parts would make a car travel the farthest. Then the pretest was administered. Children were told that it was acceptable to guess on the pretest because they had never worked with the cars before. In addition, they were informed that the posttest was almost the same as the pretest and that it would be used to see what they learned from their experience in building and running cars.

*Assembling and testing cars.* When children finished the pretest, they were directed to either the physical car parts or the computer. They were instructed on how to use the datasheet to record, for each car that they built, its features and how far it traveled. Then they were instructed on how to assemble and test a car, using either the computer program or the physical parts. The experimenter

demonstrated how to build and test a car by assembling a sample car from a short body, a thick axle, large thick back wheels, and small front wheels. (This car was one of the worst cars: It traveled only about 8 feet.) Children then entered the sample car's features on the datasheet, tested it, and recorded the distance it traveled.

Once this instructional phase was completed, children were told—depending on condition—either that they had 20 minutes to construct as many cars as they wanted, or that they could construct six more cars. While children were assembling their cars, the experimenter monitored their entries on the datasheet from the previous experiment and corrected any data entry errors.

*Posttest and distance car.* After the children finished experimenting, they were given the posttest and reminded that they could use all the results on their datasheets to answer any of the questions. Upon completion of the posttest, all children were asked to build the physical car that they felt would travel the farthest (based on their responses to Question 7 on the posttest).

### Results

Our primary question was whether children's knowledge gains about mousetrap cars would be different in the physical or virtual conditions. A related question was whether constraints on either time or number of cars would affect learning. The criterion measure was the number of correct answers on the pretest and posttest. Each child's choice of the "best" value for the causal variables (body length, back axle width, back wheel diameter, front wheel diameter) was scored as correct (1) if that choice would contribute maximally to distance traveled and incorrect (0) if it would not. For the noncausal variables (back wheel thickness and front wheel thickness), a response of "doesn't matter" was scored as correct (1) and any other response was scored as incorrect (0). (With respect to the transmission of energy from the mousetrap spring to the car, front wheel diameter is not a causal factor. However, larger front wheels tended to keep cars going straight, instead of veering to one side or the other, and some students noticed this effect. Thus, for this factor, responses were scored as correct if they said either that a large front wheel would contribute to distance traveled, or that front wheel size did not matter.) All six responses were summed into a "total knowledge" score.

As expected, children were much faster in constructing and testing cars in the virtual than in the physical condition. Thus, in the fixed-time condition, children constructed many more virtual ( $M = 20.1$ ,  $SD = 3.2$ ) than physical cars ( $M = 6.1$ ,  $SD = 0.95$ ),  $t(26) = 15.8$ ,  $p < 0.0001$ . Similarly, when constrained to build only six cars, children completed the task in less time in the virtual than in the physical condition.

#### *Knowledge about Factors Contributing to Distance Traveled*

For each of the six questions, children could choose one of two values or respond "doesn't matter." Thus, a completely random responder would guess the correct answer on one-third of the questions, producing an expected score of 2. Children's initial knowledge was better than that,  $M = 2.50$ ,  $SD = 0.191$ ,  $t(55) = 2.62$ ,  $p = 0.01$ , although it was far from the ceiling value (6.0), and the mean number of questions answered correctly increased significantly from pretest to posttest ( $M = 4.0$ ,  $SD = 0.12$ ) (see Fig. 4). A 2 (phase: pre- or posttest)  $\times$  2 (material: physical or virtual)  $\times$  2 (constraint: time or number of cars) repeated-measures ANOVA on children's test scores, with phase as the repeated measure, showed a main effect for phase,  $F(1, 52) = 54.0$ ,  $p < 0.0005$ , with no other main effects or interactions. To look at individual children, we classified children as "learners" or "nonlearners," according to whether or not their knowledge scores increased by two or more correct answers between pre- and posttest. In the physical condition there were 14 learners and 14 nonlearners, and in the virtual condition there were 16 learners and

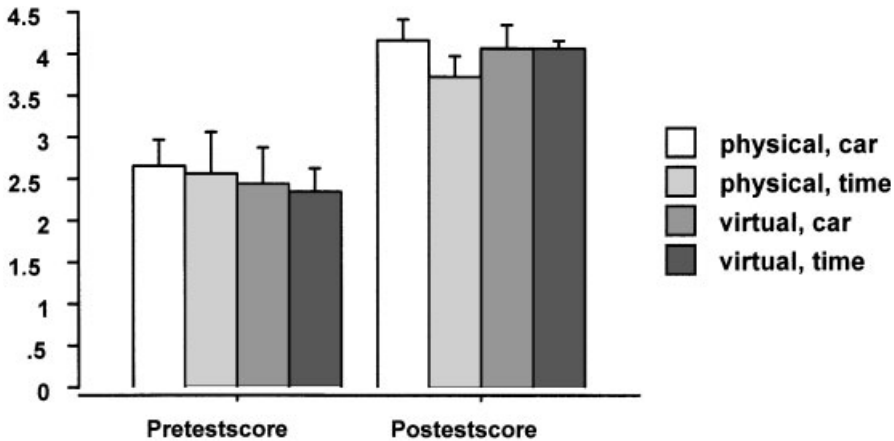


Figure 4. Mean number of correct answers (with standard errors) on the pretest and the posttest, by condition (maximum = 6).

12 nonlearners,  $\chi^2(1, 56) = 0.29$  (not significant). Increasing or decreasing the criterion for being classified as a learner did not change this null effect of material on the number of learners.

Recall that, of the six factors involved in mousetrap car designs, only four are causally related to distance traveled (body length, back axle width, front wheel diameter, and back wheel diameter), whereas the other two are noncausal (back wheel thickness and front wheel thickness). To determine whether children's knowledge gains were specific to causal or noncausal factors we conducted independent analyses of learning gains for the aggregated four causal factors and the aggregated two noncausal factors. For causal variables, knowledge scores increased from 51% correct to 91%,  $F(1, 52) = 121$ ,  $p < 0.0001$ . However, for the two noncausal variables, there was no significant change, with children correctly responding to 23% on the pretest and 17% on the posttest. For these variables, an incorrect answer meant that children were attributing causality to a variable that had no effect. Children's lack of improvement on the identification of noncausal variables is consistent with Kuhn, Schauble, and Garcia-Mila's (1992) finding that it is easier "for subjects to recognize the presence of a causal effect where none was expected than it is for them to abandon the belief that a causal effect exists" (p. 307). In addition, the failure to identify noncausal variables was, in part, a consequence of children's tendency to vary several features from one car design to the next. This is not surprising, given that children were given no instruction about how to design unconfounded experiments in this study.

In addition to the questions about the role of each of the possible factors, the posttest included one open-ended question ("What else do you think would be important for building a distance car?") designed to assess the extent to which children could propose additional features that might contribute to the distance that a car traveled. Children were classified as "poor responders" if they replied "I don't know," or if they provided an incorrect answer. They were classified as "good responders" if they mentioned at least one thing that would, in fact, make a car go farther. Good answers included statements such as: "Make sure the car goes straight"; "Let the string come loose from the axle after it fully unwinds so the car can free roll"; or "Make sure the surface of the floor is smooth." A  $2 \times 2$  cross-classification of individual children by material condition (virtual or physical) and type of responder yielded 11 poor and 17 good responders in the physical condition and 18 poor and 10 good responders in the virtual condition,  $\chi^2(1, 56) = 3.5$ ,  $p = 0.06$  (Fisher's exact  $p = 0.11$ ). Of all the measures, this was the only one for which an advantage for physical materials approached significance. Although children in the virtual condition had no

direct experience with such things as running the cars on smooth floors or cars veering off to the left or right, their responses to this final question were no worse than those of children in the physical condition.

### *Distance Traveled by Assembled Cars*

In an engineering design task, the “proof of the pudding” is how well the design actually performed. For mousetrap cars, the criterion was how far the car traveled. Thus, a pragmatic measure of learning is the difference in the distance traveled by a child’s initial and final “ideal” cars. (Because some children did not actually construct and run their initial ideal distance car during the testing phase, the distance it *would have* traveled, had it been constructed, was estimated by averaging the distance traveled by equivalent cars constructed by other children.) The mean distance traveled by children’s “ideal” cars—had they actually built them—increased significantly (pretest:  $M = 24.4$  feet,  $SD = 11.9$  feet; posttest:  $M = 38.6$  feet,  $SD = 2.6$ ) (see Fig. 5). A 2 (phase: pre or post)  $\times$  2 (material: physical or virtual)  $\times$  2 (constraint: time or number of cars) repeated-measures ANOVA, with test phase as the repeated measure, showed a main effect for phase,  $F(1, 52) = 78.7$ ,  $p < 0.0001$ , but no other main effects or interactions.

### *Confidence*

Recall that on both pre- and posttest, after answering each knowledge question, children indicated their confidence in that answer on a 5-point Likert scale (see Appendix B). Children’s average confidence increased significantly from pretest ( $M = 3.1$ ,  $SD = 0.12$ ) to near ceiling value in each condition at posttest ( $M = 4.5$ ,  $SD = 0.07$ ). Once again, a 2 (test phase)  $\times$  2 (material)  $\times$  2 (constraint) repeated-measures ANOVA, with test phase as the repeated measure, showed a main effect for test phase,  $F(1, 52) = 240.6$ ,  $p < 0.0001$ , with no other main effects or interactions.

### *Gender Effects*

When gender was added to the analyses of learning and performance, no significant differences between boys and girls emerged. Thus, the initial and final measures of domain knowledge and distance traveled were the same for boys and girls, regardless of materials or time constraint. However, there *was* a significant effect of gender with respect to confidence. Although

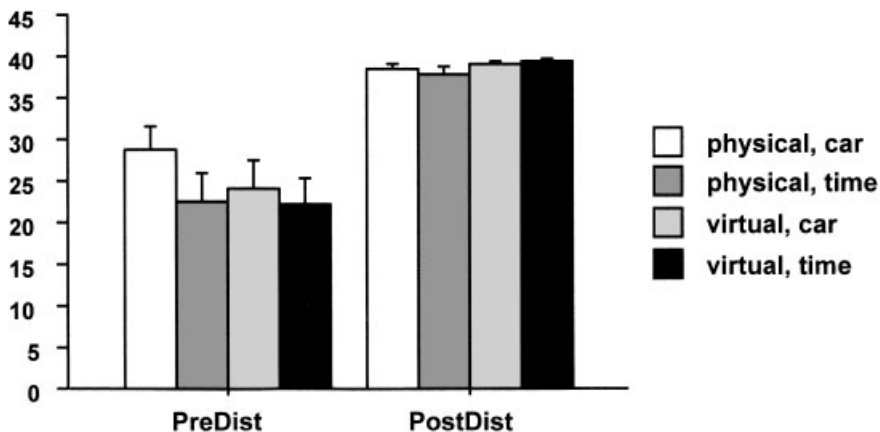


Figure 5. Mean distance (in feet) traveled by students’ “ideal” distance cars on pretest and posttest, by condition (with standard errors).

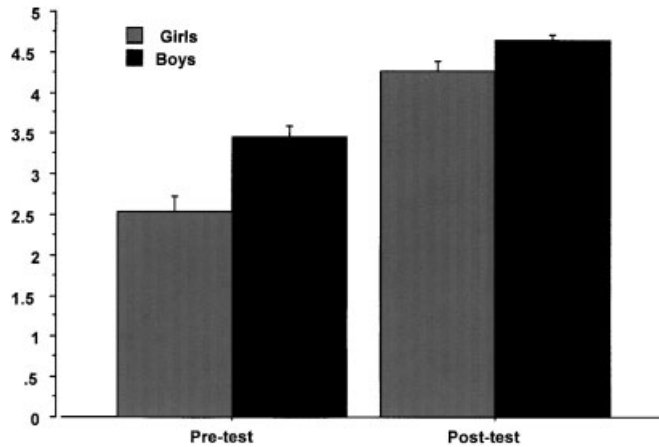


Figure 6. Boys' and girls' average confidence on the pretest and the posttest (with standard errors).

both boys and girls showed increases in confidence from pretest to posttest, the absolute level of girls' confidence was significantly lower than boys' on both pre- and posttest. On the pretest, girls' confidence was  $M = 2.5$ ,  $SD = 0.19$ , whereas boys' confidence was  $M = 3.5$ ,  $SD = 0.12$ , out of a maximum of 5. On the posttest, girls' confidence was  $M = 4.3$ ,  $SD = 0.12$ , whereas boys' confidence was  $M = 4.7$ ,  $SD = 0.08$ . A 2 (test phase)  $\times$  2 (gender)  $\times$  2 (material) repeated-measures ANOVA, with test phase as the repeated measure, showed a main effect for phase,  $F(1, 52) = 240.6$ ,  $p < 0.0001$  (as shown earlier), a main effect for gender,  $F(1, 52) = 16.7$ ,  $p < 0.0002$ , and a phase-by-gender interaction,  $F(1, 52) = 6.8$ ,  $p < 0.012$ . Girls' confidence increased more than boys' from pretest to posttest, although it still remained below boys' at posttest (see Fig. 6).

Of particular interest is the finding that this difference in confidence was not accompanied by a difference in effort. Recall that, in the fixed-time condition, children were free to construct and test as many cars as they could during a 20-minute period. Of course, many more cars were constructed in the virtual than in the physical condition, but in neither condition was there a significant gender difference in the number of cars constructed and tested. In the physical condition, the mean number of cars constructed and tested by girls was 6.3 ( $SD = 0.87$ ) and by boys was 5.9 ( $SD = 0.44$ ); in the virtual condition, girls tested, on average, 13.3 cars ( $SD = 7.9$ ), and boys tested 13.0 cars ( $SD = 7.6$ ). A gender-by-material ANOVA on number of cars built and tested yielded a main effect for material,  $F(1, 52) = 20.8$ ,  $p < 0.0001$ , but no effect for gender and no interactions.

### Discussion

The purpose of this study was to determine the effects of putting learners' hands on virtual rather than physical materials in a scientific discovery context. This is an important contrast because most recommendations about hands-on science implicitly or explicitly exclude computer simulations and virtual labs from their definition of "real" hands-on activities. For example, the National Science Teachers' Association's position statement on the use of computers in science education recommends that "computers should enhance, but not replace, essential 'hands-on' laboratory activities" (NSTA, 1999). The clear implication is that NSTA views the use of virtual materials as distinct from hands-on science activities. However, as Clements and Sarama (2003) pointed out in their extensive review of the literature on children's use of computers, there is neither theoretical nor empirical justification for such an exclusion. More specifically, our results

show that instructional medium—virtual or physical—had no effect on children’s ability to learn from their own hands-on attempts to discover the causal factors in the distance traveled by mousetrap cars, or in their ability to “engineer” an optimal distance car. In addition, children learned equally well with either medium regardless of whether they were limited in the number of cars they could build or the amount of time they could spend on the task. This null effect cannot be attributed to measurement insensitivity or floor or ceiling effects, because in all conditions these measures revealed significant gains in knowledge, in confidence, and in the performance of the cars that children constructed.

These results extend the Triona and Klahr (2003) study in several ways. First, Triona and Klahr investigated the effect of material in the context of direct instruction—contrasting Cells A and E in Table 1. In contrast, the present study compared virtual and physical materials in the context of open-ended discovery, as children constructed and tested mousetrap cars without any instruction, and with feedback only from the behavior of the cars, rather than from an instructor (Cells D and H in Table 1). Second, Triona and Klahr focused on the control of variables strategy (CVS)—a domain-general procedure—whereas the present study looked at children’s acquisition of knowledge about causal factors in the specific domain of mousetrap cars. Third, like Triona and Klahr, we found no difference between physical and virtual conditions on a variety of measures: children demonstrated substantial gains in knowledge and confidence, regardless of the material used. Finally, Triona and Klahr focused on third and fourth graders, whereas this study used seventh and eighth graders, thereby extending the age range over which physical and virtual materials have proven to be equally effective.

The lack of a physical–virtual effect in the present study is particularly interesting, given that the task involves building and testing a mechanical device for which physical examination and manipulation might be expected to contribute to the learning process. Moreover, the use of set time and set car conditions revealed that children were able to increase their knowledge and confidence equally well even when they spent significantly less time on the task (which occurs in the virtual set number of cars condition).

Perhaps the most surprising finding is that the physical–virtual distinction had no effect on the quality of children’s answers to the final open-ended questionnaire item (“What else do you think would be important for building a distance car?”). Children in the physical materials condition often observed mistakes such as the car going crooked, the string coming loose (which means the car did not stop after the string ran out), or watching the car move slowly because the wheels were on too tight, which caused friction. In contrast, children in the virtual condition never experienced any such unanticipated events, because such features were not modeled in the virtual simulation. We conjectured that the information yielded by such unanticipated events would enrich the knowledge base of children in the physical condition, by providing them with a more nuanced understanding of the underlying physics of mousetrap cars, but our analysis of children’s responses to the final question gave no support to this conjecture.

Although not designed primarily as an investigation of gender effects on children’s scientific thinking, this study does add a few findings of interest to the literature on gender differences (for an extensive review see Helgeson, 2005). It is well established that girls are less confident than boys about their performance in domains that are stereotypically perceived as male-type tasks (Kahle & Meece, 1994) and in situations having ambiguous feedback (Lenney, 1977). Assembling mousetrap cars—whether in the physical or virtual condition—is likely to have been perceived as a male task, and although the task provided concrete feedback with respect to the distance each car traveled, there was no clear criterion of success and no way for children to know how far their best car traveled compared with other children’s cars. Thus, either gender specificity or ambiguous feedback could be responsible for these confidence discrepancies. The fact that girls’ confidence

increased more than boys' suggests that, by posttest, girls may have found the task somewhat less ambiguous.

Recall that the confidence discrepancies were not accompanied by male–female differences in any of our performance and knowledge measures. Trankina (1993) suggested that one source of such confidence discrepancies might be biological differences with respect to quantitative skills and spatial visualization. However, in this study, girls were as good as boys in designing and learning about mousetrap cars. Therefore, biological differences cannot account for the disparity in confidence. Our fixed-time condition also allows us to rule out at least one additional explanation for the gender difference. Some studies have found that girls do not try as hard as boys in domains that are supposed to be “boy tasks”; that is, they disengage (Kumar & Helgeson, 2000). If this were the case, then we would expect girls to construct fewer cars than boys in the constrained time conditions. However, as noted earlier, girls constructed just as many cars as did boys.

The boy–girl confidence discrepancy was not affected by instructional medium. There is little basis for predicting which way, if any, the virtual–physical distinction could have been expected to impact this confidence discrepancy. On the one hand, because the physical condition required children to manipulate and assemble mechanical devices, it could present a context in which girls would be less confident than in the virtual condition, which required no manipulative skills beyond conventional computer use. On the other hand, girls might have viewed the computer interface as an even more “boy-appropriate” context than the physical materials. The fact that male–female confidence differences were not affected by type of material suggests that, in this context at least, girls are not disadvantaged by using virtual materials.

Because these results showed no differences between physical and virtual materials in either learning or confidence, other factors relating to the materials should be considered in choosing a hands-on instructional environment. In addition to a relatively easy development, the virtual materials are easier to implement. Unlike children in the physical condition, who needed access to an available school corridor to collect their data, children in the virtual condition were able to remain in a single location throughout the entire experimentation period. More generally, virtual experimentation usually takes less space and effort and affords easier classroom management than physical experimentation. Easy duplication and distribution of virtual materials is another obvious advantage over physical “science kits” (assuming, of course, that computers are available to run the programs). Moreover, once a virtual lab is configured on a computer, it can be extended to include data logging, instructional guidance, and so on, ultimately facilitating the creation of a computer-based intelligent tutor (Koedinger, Anderson, Hadley, & Mark, 1997). In addition, the time savings of using virtual materials can be substantial: It took children in the physical condition about 20 minutes to build six cars, whereas children in the virtual condition built the same number of cars in only 5–6 minutes. Finally, virtual materials are easier than physical materials to replicate and distribute. Overall, on this type of task at least (as well as the springs environment studied by Triona & Klahr, 2003, virtual materials would seem to offer several advantages when compared with physical materials.

Nevertheless, this study and its precursor (Triona & Klahr, 2003) represent only a small part of the space of hands-on science materials, and much remains to be learned about the relative efficacy of physical and virtual materials when they are used in different domains, with different instructional goals, approaches, outcome measures, and types of students. For example, physical materials are likely to have an advantage in domains requiring physical manipulation and tactile senses such as pouring and mixing of chemicals, and there may be domains—such as the life sciences—where having learner’s hands on “the real thing” may have important effects on learning (cf. Eberbach and Crowley’s [2005] study of different types of materials to learn about plant pollination processes or Apkan’s [2002] study of dissections in biology class). In contrast,

there are domains in which virtual materials are the only way to dynamically depict the phenomena being studied, such as when very large or very small temporal or physical dimensions are involved (e.g., astronomy, geology, molecular biology). The relative efficacy of virtual materials in any of these domains would, of course, depend on the fidelity of the simulation, and the extent to which the essential features and interactions of the domain were retained in the virtual world. Moreover, type of material may influence the learner's attitude, long-term recall and transfer, and other aspects of hands-on science instruction not addressed here. Clearly, a large space of experimental designs remains to be explored in order to fully understand the nuances of hands-on science instruction and to further its optimal use.

The authors thank the administrators, teachers, parents, and especially the children from the Carlow College Campus School and the Winchester–Thurston School for their enthusiastic support, cooperation, and participation. We also thank Vicki Helgeson for guidance on the gender differences literature, and also the members of the Discovery Lab research group (Mari Strand Cary, Norma Chang, Amanda Jabbour, Elida Laski, Junlei Li, Amy Masnick, Audrey Russo, and Stephanie Siler) for their suggestions at various stages of the project. This research is based on an undergraduate senior thesis by the third author.

## Appendix A: Components of Mousetrap Cars

### *Body*

The body was made of Balsa wood and had two sides and a top. Holes were drilled at the ends of both sides for the axles to go through. The front axle was put through the front of the body and made so that it could not fall out. The back axle was put through the back of the body and made so that it could be interchanged. The mousetrap was glued to the front of the body and a lever arm made of a brass rod extended mousetrap arm. The end of the lever arm lined up perfectly with the back axle when the mousetrap was pulled back. A string was fastened to the end of the arm and a loop was tied at the end of the string. The length of the string was equal to the distance between the end of the lever arm (when fully extended) to the back axle.

*Short Body*—9 inches long.

*Long Body*—1 foot long.

### *Back Axle*

The back axle was made of a hollow brass rod, which was 6 inches long and 0.25 of an inch in diameter. A hook was attached to the back axle for the string to hook on to.

*Thin Axle.* The brass rod with the hook attached directly to the rod.

*Thick Axle.* The thick axle was created by slipping a wooden “sleeve” over the thin axle. The sleeve had a hook on its outside surface to which the string could be tied. When this axle was used, the string was wound around wooden dowel rather than the brass rod.

### *Wheels*

All wheels were perfectly circular and had rubber bands stretched around the rims so that they were better able to grip the ground. In the center of the wheel there was a rubber washer that fit into the hole in the center of the wheel. This washer had a hole in the center, which perfectly fits the axle and made it possible for children to easily take wheels on and off the axles. Wheels were covered

by different colored pieces of construction paper so that different wheels were color coded and labeled.

*Large Thick Wheels.* —each wheel was made of two regular CDs glued together. The large thick wheels had red wheel covers and were labeled “Large thick.”

*Large Thin Wheels.* —each wheel was made of a single CD. These wheels had white wheel covers and were labeled “Large thin.”

*Small Wheels.* —The small wheels were made of 3-inch-diameter wooden circles, which are approximately the same thickness as a CD. A hole the same size as the hole in the center of a CD was drilled in the center of each of the circles so that the same rubber washers were able to fit into them. These wheels had green wheel covers and were labeled “small.”

#### Appendix B: Pretest and Posttest Questions

##### 0. (Pretest only):

Have you ever worked with mousetrap cars before? Yes \_\_\_ No \_\_\_

If so, describe what you did with them: \_\_\_\_\_

##### 1. Which body length will make the car travel the farthest?

- a. Long body
- b. Short body
- c. Body length does not matter

Confidence:

1	2	3	4	5
Not Confident		Somewhat Confident		Very Confident

##### 2. Which axle width will make the car travel the farthest?

- a. Thick
- b. Thin
- c. Axle width does not matter

Confidence:

1	2	3	4	5
Not Confident		Somewhat Confident		Very Confident

##### 3. What back wheel diameter will make the car travel the farthest?

- a. Back wheel with a large diameter
- b. Back wheel with a small diameter
- c. Back wheel diameter does not matter

Confidence:

1	2	3	4	5
Not Confident		Somewhat Confident		Very Confident

##### 4. What back wheel thickness will make the car travel the farthest?

- a. Thick back wheel

- b. Thin back wheel
- c. Back wheel thickness does not matter

Confidence:

1                      2                      3                      4                      5  
 Not Confident                      Somewhat Confident                      Very Confid

5. What front wheel diameter will make the car travel the farthest?
- a. Front wheel with a large diameter
  - b. Front wheel with a small diameter
  - c. Front wheel diameter does not matter

Confidence:

1                      2                      3                      4                      5  
 Not Confident                      Somewhat Confident                      Very Confid

6. What front wheel thickness will make the car travel the farthest?
- a. Thick front wheel
  - b. Thin front wheel
  - c. Front wheel thickness does not matter

Confidence:

1                      2                      3                      4                      5  
 Not Confident                      Somewhat Confident                      Very Confid

Hands on What?

7. Circle the parts that you believe would create the car that travels the farthest.

<p><b><u>Body Length:</u></b>                  Long Body                  Short Body</p>	<p><b><u>Axle Width:</u></b>                  Thin Axle                  Thick Axle</p>
<p><b><u>Back Wheels:</u></b>                  Large Thick Wheels                  Large Thin Wheels                    Small Wheels</p>	<p><b><u>Front Wheels:</u></b>                  Large Thick Wheel                  Large Thin Wheels                    Small Wheels</p>

8. (Posttest only):

What else do you think would be important for building a distance car?

## Appendix C

*The Following datasheet is for the fixed number of cars condition. For the fixed-time condition, the datasheet had 35 numbered rows.*

<b>Car#</b>	<b>Body: Long or Short</b>	<b>Back Axle: Thick or Thin</b>	<b>Back Wheels: Large Thick or Large Thin or Small</b>	<b>Front Wheels: Large Thick or Large Thin or Small</b>	<b>Distance (Feet, Inches)</b>
Sample					
1					
2					
3					
4					
5					
6					

## References

- Apkan, J.P. (2002). Which comes first: Computer simulation of dissection or a traditional laboratory practical method of dissection. *Electronic Journal of Science Education*, 6. Retrieved July 6, 2005 from: <http://unr.edu/homepage/crowther/ejse/akpan2.pdf>
- Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology*, 59, 960–970.
- California State Teachers Association. (2004). K-8 instructional materials adoption. Retrieved July 6, 2005 from: <http://www.cascience.org/IMCriteria.html>
- Clements, D.H. (1999). ‘Concrete’ manipulatives, concrete ideas. *Contemporary Issues in Early Childhood*, 1, 45–60.
- Chen, Z., & Klahr, D. (1999). All other things being equal: Children’s acquisition of the control of variables strategy. *Child Development*, 70, 1098–1120.
- Clements, D.H., & Sarama, J. (2003). Strip mining for gold: Research and policy in educational technology—A response to “Fool’s Gold.” *Educational Technology Review*, 11. Retrieved July 6, 2005 from: <http://www.aace.org/pubs/etr/issue4/clements.cfm>
- Deboer, G. (1991). *A history of ideas in science education*. New York: Teachers College Press.
- Eberbach, C., & Crowley, K. (2005). From living to virtual: Learning from Museum objects. *Curator: The Museum Journal*, 48, 317–338.
- Flick, L.B. (1993). The meanings of hands-on science. *Journal of Science Teacher Education*, 4, 1–8.
- Galley, M. (2004). California state board backs hands-on science. *Education Week*. Retrieved July 6, 2005 from: <http://www.edweek.org/ew/ewstory.cfm?slug=28Caps.h23>
- Haury, D.L., & Rillero, P. (1994). *Perspectives of hands-on science teaching*. Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.
- Helgeson, V.S. (2005). *The psychology of gender* (2nd ed.) Upper Saddle River, NJ: Prentice-Hall.
- Hmelo, C.E., Holton, D.L., & Kolodner, J.L. (2000). Designing to learn about complex systems. *Journal of the Learning Sciences*, 9, 47–298.
- Hodson, D. (1996). Laboratory work as scientific method: Three decades of confusions and distortion. *Journal of Curriculum Studies*, 28, 115–135.
- Huxley, T. (1899). *Science and education*. New York: Appleton.

International Labor Organization. (2004). More women join world's workplace. Studies and Statistics. Retrieved July 6, 2005 from: [http://www.us.ilo.org/studies/workingwomen\\_04.cfm](http://www.us.ilo.org/studies/workingwomen_04.cfm)

Johnson-Gentile, K., Clements, D.H., & Battista, M.T. (1994). Effects of computer and noncomputer environments on students' conceptualizations of geometric motions. *Journal of Educational Computing Research*, 11, 121–140.

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.

Kahle, J.B., & Meece, J. (1994). Research on gender issues in the classroom. In D.L. Gable (Ed.), *Handbook of research on science teaching and learning*. New York: Macmillan.

Koedinger, K.R., Anderson, J.A., Hadley, W.H., & Mark, M.A. (1997). Intelligent tutoring goes to the big city. *International Journal of Artificial Intelligence*, 8, 30–43.

Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 371–393). Malden, MA: Blackwell.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction*, 9, 285–327.

Kumar, D.D., & Helgeson, S.L. (2000). Effect of gender on computer-based chemistry problem solving: Early findings. *Electronic Journal of Science Education* 4. Retrieved July 6, 2005 from: <http://unr.edu/homepage/crowther/ejse/kumaretal.html>

Lehrer, R., & Schauble, L. (2006). Scientific thinking and science literacy. In W. Damon, R. Lerner, K.A. Renninger, & I.E. Sigel (Eds.), *Handbook of child psychology: Vol. 4. Child psychology in practice* (6th ed.) Hoboken, NJ: John Wiley & Sons.

Lenney, E. (1977). Women's self-confidence in achievement settings. *Psychological Bulletin*, 84, 1–13.

Marshall, P. (2005). Tangibles in the balance: a comparison of physical and screen versions of the balance beam task. *Proceedings of the 8th Human-Centred Technology Postgraduate Workshop*. Cognitive Science Research Paper 576, University of Sussex.

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

NAEP (2000). IES National Center for Education Statistics. The Nations's Report Card: Science 2000. Retrieved 1.10.2005, from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubiod=2003453>

National Science Teachers Association. (1990). Position statement on laboratory science. Retrieved July 6, 2005 from: <http://www.nsta.org/positionstatement&psid=16>

National Science Teachers Association. (1999). Position statement on the use of computers in science education. Retrieved July 6, 2005 from: <http://www.nsta.org/positionstatement&psid=4>

Ruby, A. (2001). *Hands-on science and student achievement*. Santa Monica, CA: RAND.

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

Trankina, M.L. (1993). Gender differences in attitudes toward science. *Psychological Reports*, 73, 123–130.

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