Computer simulations and clear observations do not guarantee conceptual understanding

Maggie D. Renken a, *, Narina Nunez b

a Educational Psychology and Special Education Department, Georgia State University, PO Box 3979, Atlanta, GA 30302-3979, USA
b Psychology Department, University of Wyoming, Dept. 3415, 1000 E. University Ave, Laramie, WY 82070, USA

Abstract

Evidence for cognitive benefits of simulated versus physical experiments is unclear. Seventh grade participants (n = 147) reported their understanding of two simple pendulum problems (1) before conducting an experiment, (2) immediately following experimentation, and (3) after a 12-week delay. Problem type was manipulated within subjects — participants’ understanding of one problem was typically accurate and for the other was typically inaccurate. Experiments were computer-simulated or hands-on and were observed in a slow motion or real time replay. There was no difference between simulated vs. hands-on or slow motion vs. real time replay for conceptual understanding outcomes. Instead, the problem type, the time of posttest, and the participant’s experimentation strategy were significant predictors. Specifically, poor experiments were especially bad at informing understanding of the previously misconceived problem. Furthermore, even for the problem that was previously conceived correctly, accurate understanding declined over twelve weeks for those participants who conducted inadequate experiments. Post hoc tests revealed participants were less likely to control variables during simulations.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

The empirical consideration of computer simulation versus physical experimentation in science education is increasingly important as technology in the classroom advances and more simulation software becomes available for teachers and students (Moreno & Mayer, 2007; Rutten, van Joolingen, & van der Veen, 2011). While computer simulations in the classroom have been the topic of much discussion and research attention as early as the 1980s, emphasis within educational research remains on simulations as enhancements to or replacements for classroom instruction (Akpan, 2001; Hofstein & Lunetta, 2004). Recent reviews largely identify quasi-experimental findings focused on computer simulations in complex settings involving cognitive tutors, student collaboration, or multimedia learning (e.g., Bell & Trundle, 2008; Hofstein & Lunetta, 2004; Rutten et al., 2011). This research approach is informative for instructional strategies and simulation design, but does not address the individual learners’ basic cognitive processes. Therefore, the benefits of computer simulations for student learning void of extraneous variables (e.g., collaboration, multimedia context) remain unclear (Chang, Chin, Lin, & Sung, 2008; Kim & Pedersen, 2011; van der Meij & de Jong, 2006; Moreno, 2009). Furthermore, in the science classroom computer simulations have been lauded for their ability to address difficult- or impossible-to-observe phenomena (e.g., photosynthesis) (Cepni, Tas, & Kose, 2006; Roth, Wosczyna, & Smith, 1996; Urban-Woldron, 2009). But for simple phenomena that can be illustrated via hands-on experimentation, recent research has failed to consider the exact cognitive benefits of computer simulations over physical hands-on experimentation. Before valuable resources are allocated toward implementing computer simulations in the classroom, their unique contribution to successful cognitive outcomes should be identified (Moreno & Mayer, 2007).

1.1. Science conceptual understanding as a primary cognitive outcome

An important cognitive outcome in science domains is accurate conceptual understanding. Classical approaches to knowledge acquisition and conceptual change assert that science learning occurs over short periods of time (c.f., McCloskey, 1983; Posner,
Strike, Hewson, & Gertzog, 1982). More recent approaches suggest, however, that accurate conceptual understanding involves a long process – requiring the coordination of an individual’s naïve theories and incoming evidence over the course of many years (Vosniadou, 1994, 2007a). Over these longer periods of time, individuals interact with new information in some socio-cultural context (e.g., home, school, play). Any knowledge acquired during these “interactions,” then, is applied toward either enriching conceptual understanding (i.e., knowledge acquisition) or restructur- ing existing conceptual understanding (i.e., conceptual change) (Vosniadou, 2007b). Thus, the process of understanding some science concept requires both immediate and long-term cognitive mechanisms that influence an individual’s knowledge structure. It follows that although brief interactions, such as computer-simulated or hands-on experiments, are insufficient for complete conceptual understanding; they contribute to the additive processes of knowledge acquisition and conceptual change.

1.2. The benefit of computer simulations for conceptual understanding

To this end, computer-simulated experiences were assumed, early on, to be valuable tools for students’ conceptual under- standing in science domains. This is in part because they provide a rich environment for eliminating distractors and constraining learning to relevant evidence (ChanLin, 2001; White, 1984). Computer “microworlds” in which real world complexity is elimi- nated were initially introduced in the classroom as a logical solu- tion for informing students’ knowledge. The assumption was that by eliminating complexity, computer simulations are best situated to confront or challenge inaccurate understanding. In other words, early predictions and the research that followed suggest that computer simulations may be better than hands-on experiments at enriching and/or restructuring students’ conceptual understanding through two specific and inter-related mechanisms: 1) reducing and/or eliminating complexity and subsequently, 2) confronting inaccurate conceptual understanding.

A specific form of real world complexity that often arises in the case of physical science concepts occurs when experimental observation is unclear. For example, because dropped items reach the ground in an instant, the moment of contact is dif- ferent from what students observe. As a result, what is often assumed as clear and accurate. By virtue of their artificial nature, one of the ways computer simulations are well suited for reducing complexity is by clarifying observations. For example, many physical science simulations addressing Newtonian laws of motion, which are not only ill-conceived and difficult to observe, include opportunities to view results in slow motion. This allows enhanced opportunity for clearly observing results.

It is important to note that although simulations have the potential to reduce complexity via tools like slow motion experi- mental observations, computer-simulated learning experiences must be carefully supported through background knowledge and assistance during hypothesis formation, experimentation, and interpretation of data (Chang et al., 2008). More recently, research has continued to support the notion that instruction in the form of feedback or cues is necessary when completing simulation assignments (Corbett, Kauffman, Maclaren, Wagner, & Jones, 2010; Roth et al., 1996). Accordingly, computer simulation may not be an adequate substitute to physical experience for basic conceptual understanding, and specifically conceptual change. As the following section highlights, this especially may be the case during adolescence.

1.3. Adolescents’ understanding of science concepts

Adolescence is a unique developmental period for considering conceptual understanding in science domains. First, adolescents’ inaccurate knowledge in physical science is typically robust – meaning their knowledge structures have been consistently rein- forced and are most likely difficult to change (Brna, 1987; Chinn & Malhotra, 2002; Clement, 1982; Penner & Klahr, 1996). For example, previous research, most early adolescents inaccurately reported that the mass of an object affects the speed at which it travels and were unlikely to change their response, even when provided contrary data (Renken & Nunez, 2010). In general, adolescents have marked difficulty accepting evidence that contradicts their existing knowledge structure (Schauble, 1996).

Conceptual understanding specific to physical science is further complicated for adolescents since they also demonstrate difficulties designing sound experiments (Zimmerman, 2007). While this is the case for both computer simulated and hands-on experiments, the modeling nature of computer simulations may make the interpretation of their evidence even less accurate. In early work regarding computer-simulated models of physical science phenomena, adolescents commonly perceived a computer to be broken when the results of their actions did not match their inac- curate beliefs (diSessa, 1982). In other words, the adolescent may attribute a mismatch between belief and observation to the hypo- thetical or incorrect nature of computers. Furthermore, adolescents are unaware of the effect of not controlling variables on modeling outcomes on a computer (Brna, 1987). Even when performing controlled computer-based experiments, adolescent students were incapable of making plausible inferences about stimuli relationships (Brna, 1987). Previous research has indicated that among college undergraduates there is no difference in the conceptual benefits of physical versus simulated experiments (Zacharia & Olympiou, 2011). These particular findings, however, have not been extended to an adolescent population with potentially less experience than undergraduates learning from simulated science experiments.

1.4. The current research

In sum, a common postulation is that computer simulations are better than hands-on experiments at informing students’ science understanding because they can provide clear evidence and, by doing so, are better able to confront any inaccurate beliefs. However, there is some suggestion that the benefits of computer simulations to adolescents’ conceptual understanding may be fewer than those stemming from physical experiments (see Sections 1.2 and 1.3). Our expectation, then, was that if hands-on experiments were capable of providing clear data in similar ways, they would be even more beneficial to adolescents’ conceptual understanding—given their resistance to accept simulated evidence that contradicts their current understanding. Rather than pit simulations and hands-on experiments against one another as two distinct type of experimen- tation, the present research directly addresses the roles of the following factors in determining students’ conceptual under- standing: 1) the clarity of evidence, 2) the type of science problem being examined, 3) the type of experiment (i.e., simulated vs. hands-on), and 4) the quality of experimentation strategies (considering both the control of variables and number of experimental trials run). Note that the type of science problem, as used here, addresses the notion that scientific evidence either enriches knowledge for
problems that are accurately understood or restructures knowledge for problems that are inaccurately understood. Also note that the quality of experimentation strategies is a single measure combining both the control of variables and the presence of more than one trial and is referred to from here on as experimentation adequacy. We expected that observation clarity (Variable 1), problem type (Variable 2), and experimentation adequacy (Variable 4) would best predict students’ conceptual understanding (Hypothesis 1). Specifically, reported conceptual understanding was expected to be most accurate when data was clearly observed (Hypothesis 1a), when the science problem presented an opportunity for knowledge to be enriched rather than restructured (Hypothesis 1b), and when experimentation adequacy was superior (i.e., experiments controlled variables and included more than one experimental trial) (Hypothesis 1c).

Experiment type refers to hands-on versus simulated experiments (Variable 3) and was also expected to predict students’ conceptual understanding. With observation clarity accounted for separately, physical hands-on experiments were expected to be more likely to lead to accurate student understanding than computer-simulated experiments (Hypothesis 2). Finally, the present research assessed student understanding at three time points: (1) before experimentation (prettest), (2) immediately following experimentation (immediate posttest), and (3) 12 weeks following experimentation (delayed posttest). Previous research shows that students tend to revert to prior conceptions after a 12-week delay, so more accurate understanding was expected for the immediate posttest than for the delayed posttest (Hypothesis 3) (Renken & Nunez, 2010).

2. Method

2.1. Experiment design

Two simple pendulum problems depicting specific Newtonian laws of motion were used in the current research. A pendulum consists of an object, called the “bob,” which is suspended, in this case from a string, and allowed to swing freely when released from a certain height. The setup for the current research allowed two pendulums to swing in unison (Figs. 1–3). The first pendulum problem addressed the effect of altering the length of string from which a bob is suspended on the rate at which the bob travels (length problem). Evidence presented in this problem is expected to be consistent with adolescents’ existing knowledge. For example, most students correctly assert that the bob will travel faster with a shorter string length (e.g., White, 1984). Therefore, this problem represents an opportunity for knowledge to be enriched.

The second problem addressed the effect of altering the bob’s mass on the rate at which the bob travels (mass problem). Unlike for the length problem, evidence presented in the mass problem is expected to be inconsistent with adolescents’ existing knowledge, representing an opportunity for knowledge to be restructured (e.g., White, 1984). For example, most students incorrectly assert that a bob with greater mass will travel faster than a bob with less mass. Pendulum problems were chosen because they provide fast, difficult-to-observe results and involve straightforward experimentation that may be accomplished with both a physical apparatus and a computer simulation. (See Figs. 1–3 for images of the pendulum physical apparatus and computer-simulated pendulum.) Observation clarity was manipulated in both the hands-on experiment and computer simulation by replaying experimental results for each trial in either real time or slow motion.

The 2 (observation clarity: real time replay vs. slow motion replay) × 2 (experiment type: hands-on vs. simulation) design was fixed across two problems (problem type: length vs. mass). Problem type was a within-subjects variable, while observation clarity and experiment type were between-subjects variables. Experimentation adequacy (Variable 4) was not experimentally manipulated. Instead, participants’ experiments were assessed and assigned one of three possible rankings (superior, adequate, and inadequate). Although it was not experimentally manipulated, experimentation adequacy was treated as an independent within-subjects variable since it was expected to predict conceptual understanding.

2.2. Participants

Participants were 147 seventh grade students ranging in age from 11 to 14 (M = 12.36, SD = 0.52) recruited from life science classes at a junior high school in the Rocky Mountain region. Seven participants were removed from final analysis due to missing data. Data collection was in two waves: Spring semester (n = 76) and Fall semester (n = 71). Overall, 71% of the students invited to participate consented. This proportion was greater for the Fall (86%) than for the Spring (59%) semester most likely because parent information
letters and consent forms were distributed with a packet of required school-wide permission forms in the Fall and distributed individually in the Spring. Forty-eight percent of the participants were female. None of the students reported previous or current enrollment in a physical science course.

2.3. Procedure and materials

Students with permission to participate were randomly assigned to one of four conditions: 1) hands-on, real time replay ($n = 35$), 2) hands-on, slow motion replay ($n = 35$), 3) simulation, real time replay ($n = 35$), or 4) simulation, slow motion replay ($n = 35$). Each participant completed two counterbalanced experiments involving the two pendulum problems on an individual basis during their life science class meeting time (see Section 2.1 for description of problems). First, participants completed a demographic information survey (Appendix A). For the pretest, before conducting the first pendulum experiment participants were asked verbally to predict the outcome for both the length and mass problems (Appendix B). Length and mass problem pretests were counterbalanced across participants. The experimenter demonstrated, using either the physical apparatus or the computer simulation software, the variable of interest by holding (or pausing on the computer screen) two bobs at the same starting position. Then the experimenter asked participants, “When these two weights swing from here, do you think one of the weights will come back to the starting place before the other, or they will return at the same time?” If a participant answered that one would return before the other, they then indicated which bob would return first. Predictions were coded as correct (1) or incorrect (0). Then, participants were asked if they were sure of their answer or if it was a guess. This method of assessing prediction confidence was used based on previous research in which a Likert-scale assessment (1–5, with 5 indicating the highest level of confidence) resulted in a mode response of three for both the length and mass problem (see Renken & Nunez, 2010). The aim here was to better assess

![Fig. 2. Two screenshot of the computer simulated, length problem. a) Illustration of a trial setup. b) Illustration of the above trial in motion.](image-url)
confidence with a forced-choice item, in which participants reported either being confident or not. Predictions and confidence ratings were reported before participants conducted the experiments in order to confirm that, as expected, the length problem provided opportunity for knowledge enrichment and the mass problem provided opportunity for knowledge restructuring.

After making predictions and indicating confidence for both problems, participants in the physical experiment condition were provided the pendulum used during the prediction demonstration along with one plastic bob, two (heavier) steel bobs, and three strings, each with equal lengths marked at three locations (short, medium, and long) (Fig. 1). Participants in the computer simulation condition were given a laptop with a pendulum experiment computer program (see http://phet.colorado.edu/en/simulation/pendulum-lab to access software or Figs. 2–3 for screen shots). For this condition, the experimenter pointed out on the screen how to alter the string length and bob mass for two pendulums. Participants were instructed to adjust only the settings related to string length and bob mass. Participants in both experiment type conditions were then given written and verbal instructions to test whether the length of the string and the mass of the bob affect the rate at which the bob travels (Appendix C). Rather than provide participants with explicit instructions regarding how to test the two questions of interest, the experimenter told participants to change the pendulum in any way they needed to address the questions. They were also instructed that they could use the paper provided to record information and could repeat their experiment until they had an answer.

As previously mentioned, to manipulate observation clarity, participants were randomly assigned to either real time or slow motion replay. In the real time replay condition, participants conducted their experiments (either with the physical apparatus or on the computer), and after each trial watched a replay. Regardless of condition, participants did not conduct another experimental trial. They simply viewed a replay of the original experimental set up. For the hands-on, real time replay condition, the experimenter videoed
each experimental trial using a laptop, and then participants watched a replay video on the computer screen before either conducting another experimental trial or drawing a conclusion. For the simulation, real time replay condition, participants were instructed to replay each trial. Participants in the slow motion replay condition similarly viewed a replay of their trial, but it was slowed to 1/4 real time. For the hands-on, slow motion condition, this was achieved via an “Effects” option on the video software, and involved one more (time-negligible) step for slow motion replay than for the real time replay. For the simulation, slow motion replay condition, the software allowed participants to slow down the pendulum’s swing to 1/4 time by selecting a button in the program’s toolbar. Like for the simulation, real time replay condition, participants were instructed to replay each trial before moving on to another trial or drawing a conclusion—the difference being that the replay was to follow selecting the “1/4 time” button. Thus, participants in all four conditions saw each of their experimental trials twice.

Participants then completed a posttest for each problem (i.e., length and mass). They received two forced-choice conclusion questions, for which they were instructed to answer whether the bob traveled faster, slower, or at the same speed (see Appendix C). Participants answered the mass conclusion question immediately following the mass experiment and the length conclusion question immediately following the length experiment. Conclusions were coded as correct (1) or incorrect (0). Next, participants were asked to rate, on a scale of 1–5, the enjoyment other students their age would experience and the effort other students their age would exert, if conducting the same experiment (Appendix D). Then participants were given a set of six questions to assess whether they could transfer their conclusions to similar, real-world problems (Appendix E). Three questions addressed the length problem and three questions addressed the mass problem. Each question was a forced-choice item with three answer choices indicating increased, decreased, or same rate of travel, in those or similar terms. An overall concept transfer score for each problem was calculated by coding the responses as correct (1) or incorrect (0) and adding them. Concept transfer scores ranged from 0, indicating no correct responses, to 3, indicating all correct responses, since there were three questions for each problem.

The next assessment was the delayed posttest, and it occurred after an average delay of twelve weeks from participants’ individual experimentation.1 As a class of 15–25 students, participants completed the 10-min written assessment (i.e., forced-choice conclusion and concept transfer questions) from the immediate posttest to evaluate whether they maintained the content knowledge they gained immediately after the experiment. Conclusion coding and concept transfer scoring were conducted the same at delayed as at immediate posttest.

Finally, participants’ experiments were coded for experimentation adequacy. For the physical condition, the replay video clips were coded. For the simulation condition, a screen movie capture was taken for each participant and coded. Screen movie captures were taken using Copernicus software for Macs. Frame rate was at two frames/second and movie scale was 50%. This allowed assessment of screen actions without compromising computer memory space during data collection. The region selected included the pendulum apparatus, the play/pause control, and the settings toolbar.

Videos were coded for control of variables and number of trials run by two blind raters during the Spring Wave and two different blind raters during the Fall Wave. Variables were considered controlled when an experiment displayed the same length or the same mass, but not both. An experimental trial was defined as a pause in the pendulum swing followed by a repeat of the same experimental set up. A pause followed by a change in length and/or mass for one or both of the pendulums was considered a new experiment. Experimentation adequacy was coded as inadequate (0), adequate (1), or superior (2). To receive an inadequate score, an experiment did not control for one of the two variables. To receive an adequate score, the experiment controlled for the respective variable, but included only one trial. To receive a superior score, the experiment controlled for the respective variable, and included two or more trials. In some cases, participants conducted multiple experiments for the same problem. For example, two experiments may have used different string lengths but both addressed the length problem. In these cases, the trials from the separate experiments were combined to determine the experimentation adequacy score. Inter-rater reliability was found to be Kappa = .91 (p < .0001). The primary investigator coded experiments for which the raters disagreed, and the primary investigator agreed with the same blind rater for each of these cases within the two data collection Waves.

3. Results

3.1. Establishing the predicted problem type distinction

First, analysis confirmed that predictions for the two problems were accurate and inaccurate as expected. A 2 (problem type: length vs. mass) x 2 (prediction accuracy: correct vs. incorrect) contingency table revealed that 73.6% of the participants correctly predicted a bob attached to a shorter string to travel faster than one attached to a longer string (length problem), while only 13.4% of participants correctly expected two bobs varying in mass to travel at the same speed (mass problem). A chi-square test indicated that the proportion of correct predictions for the two problems differed significantly (χ²(1, N = 147) = 104.05, p < .0001). (See Table 1 for observed and estimated expected frequencies.) Confidence in predictions did not differ across problem type, however, with 55.8% of participants expressing confidence in their length problem predictions and 53.7% of participants expressing confidence in their mass problem predictions. The indication being that although participants were far less likely to correctly predict that bob mass has no effect on rate of travel than to correctly predict that string length has an effect on rate of travel, they were equally confident in their predictions for both problems.

3.2. Immediate posttest

Participants’ reported conceptual understanding immediately following an experiment included their conclusion accuracy and concept transfer scores. The conclusion accuracy and concept transfer outcomes were expected to be predicted similarly (Hypothesis 1), but were considered in two separate analysis models. Therefore the following sections (3.2.1 and 3.2.2) address

| Table 1 | Frequency of correct and incorrect predictions for the length and mass problems. |
|--------------------------------|-----------------|--------------------------------|-----------------|
| | Problem | Prediction accuracya | Correct | Incorrect |
|--------------------------------|-----------------|--------------------------------|-----------------|
| Problem | | | | |
| Length | 103 | 37 |
| | 60.6 | 79.4 |
| Mass | 19 | 123 |
| | 61.4 | 80.6 |

Note: Observed values are listed first. Expected values are included in italics.

a A chi-square test was significant at the p < .0001 level (χ² = 104.05, df = 1).

1 The range of the delay was ten to fourteen weeks.
Hypotheses 1 and 2 by identifying which variables (i.e., observation clarity, problem type, experiment type, and experimentation adequacy) are significant predictor covariates. The sections are separated into consideration of the 1) conclusion accuracy outcome and 2) concept transfer outcome. (Conclusion accuracy proportions and concept transfer scores are reported in Table 2).

3.2.1. Conclusion accuracy

To assess the role of problem type (length vs. mass), observation clarity (slow motion replay vs. real time replay), experiment type (hands-on vs. simulation), and experimentation adequacy (inadequate, adequate, or superior) in conclusion accuracy, each factor was included in an initial generalized estimating equations (GEE) regression model. GEE analysis was employed because it allows for consideration of the predicted main effects (see hypotheses 1a–1d, 2). It also allows consideration of potentially complex interactions between the predictor variables, which provides insight into the relationships between the covariate predictors. Similar to logistic regression analysis, GEE allows model building to construct a best-fit model. GEE analysis also is appropriate for the present results because predictor variables and outcome data were categorical, and included both within- and between-subjects variables. While logistic regression is appropriate for models composed of categorical data, it does not allow for the inclusion of within- and between-subject effects; GEE does and was therefore deemed appropriate.

The first GEE analysis included each expected predictor variable and was used with problem type as the repeated measure since the length and mass problems were a within-subjects treatment. GEE with logit function and binomial distribution revealed that problem type and experimentation adequacy were significant predictor covariates of conclusion accuracy, but neither observation clarity nor experiment type were significant predictor covariates (Hypotheses 1 and 2) ($\beta_{\text{problem}} = -1.25, \chi^2 = 4.45, df = 1, p = 0.03$; $\beta_{\text{experimentation}} = 0.76, \chi^2 = 8.47, df = 1, p = 0.004$; $\beta_{\text{observation clarity}} = 0.62, \chi^2 = 13.62, df = 1, p = 0.0002$; Goodness of fit: $\chi^2 = 379.61, df = 277$). Because they were not significant predictors, observation clarity and experiment type were removed from the model. To determine how experimentation adequacy and problem type affect conclusion accuracy together, the problem type by experimentation adequacy interaction term was added. Problem type, experimentation adequacy, and the interaction term were significant covariate predictors in the final, best-fit model ($\beta_{\text{problem}} = -1.42, \chi^2 = 19.18, df = 1, p < 0.0001$; $\beta_{\text{experimentation}} = 1.59, \chi^2 = 12.11, df = 1, p = 0.0005$; $\beta_{\text{observation clarity}} = 0.97, \chi^2 = 17.81, df = 1, p < 0.0001$; $\beta_{\text{interaction}} = -0.75, \chi^2 = 4.45, df = 1, p = 0.04$; Goodness of fit: $\chi^2 = 373.71, df = 278$). In other words (and as demonstrated by the conclusion accuracy proportions in Table 2), participants were correct more often (1) when answering the length conclusion question than the mass conclusion question and (2) when conducting adequate experiments. In addition, the significant interaction term suggests that an inadequate experiment leads to fewer accurate conclusions for the mass than for the length problem ($\beta_{\text{mass problem, inadequate experiment}} = 0.07$; $\beta_{\text{length, inadequate experiment}} = 0.45$) (see Table 2).

3.2.2. Concept transfer score

As expected, for concept transfer questions related to the length problem, more than 80% of participants answered each question correctly. Conversely, for the mass problem, less than 33% of participants answered each question correctly. Significant predictive covariates of the concept transfer score were the same as those for conclusion accuracy (see Section 3.2.1). First, GEE with logit function revealed that problem type and experimentation adequacy, but neither observation clarity or experiment type, were significant predictor covariates of concept transfer score (hypotheses 1 and 2) ($\beta_{\text{problem}} = 0.46, \chi^2 = 2.56, df = 1, p = 0.11$; $\beta_{\text{experimentation}} = 1.65, \chi^2 = 203.06, df = 1, p < 0.0001$; $\beta_{\text{observation clarity}} = 0.21, \chi^2 = 6.25, df = 1, p = 0.01$; Goodness of fit: $\chi^2 = 286.08, df = 277$). The problem type by experimentation adequacy interaction term was significant in the final best-fit model, which again excluded observation clarity and experiment type ($\beta_{\text{problem}} = 0.60, \chi^2 = 5.37, df = 1, p = 0.02$; $\beta_{\text{experimentation}} = 0.23, df = 1, p < 0.0001$; $\beta_{\text{problem}} = 0.35, \chi^2 = 7.51, df = 1, p = 0.006$; $\beta_{\text{experimentation}} = -0.32, \chi^2 = 4.08, df = 1, p = 0.04$; Goodness of fit: $\chi^2 = 284.59, df = 278$). Again, participants were correct more often (1) when answering the length questions than the mass questions and (2) when conducting adequate experiments. An inadequate experiment led to lower concept transfer scores for the mass than for the length problem ($M_{\text{mass problem, inadequate experiment}} = 0.32$; $M_{\text{length, inadequate experiment}} = 2.39$) (see Table 2). Overall, conclusion accuracy and the ability to transfer knowledge were best predicted by the problem type, the adequacy of the participants’ experiments, and the interaction of the two. It appears that for a problem that allows an opportunity for knowledge to be restructured or changed, it is unlikely without performing an adequate experiment.

3.3. Delayed posttest

GEE analysis for the delayed posttest data was conducted with two new models. Similar to Section 3.2, the following results are divided into two models: one addressing the conclusion accuracy outcome measure (3.3.1) and the second addressing the conceptual transfer outcome measure (3.3.2). Both models included the following covariate predictors: problem type, observation clarity, experiment type, experimentation adequacy (similar to analysis of the immediate posttest (see 3.2). Problem type and time were within-subjects factors. To consider the impact of the 12-week delay that took place between the immediate and delayed posttest, time (immediate vs. delayed) was added to the model as a predictor covariate. Recall that understanding was expected to be more accurate during immediate than during delayed posttest (see hypothesis 3).

3.3.1. Conclusion accuracy

The observation clarity and experiment type treatment conditions were not significant predictors of conclusion accuracy

---

Table 2

<table>
<thead>
<tr>
<th>Problem</th>
<th>Conclusion accuracy</th>
<th>SE</th>
<th>Transfer score</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>0.59</td>
<td>0.04</td>
<td>2.51</td>
<td>0.06</td>
</tr>
<tr>
<td>Mass</td>
<td>0.44</td>
<td>0.04</td>
<td>0.90</td>
<td>0.10</td>
</tr>
<tr>
<td>Experimentation adequacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inadequate</td>
<td>0.28</td>
<td>0.06</td>
<td>1.44</td>
<td>0.17</td>
</tr>
<tr>
<td>Adequate</td>
<td>0.60</td>
<td>0.04</td>
<td>1.92</td>
<td>0.11</td>
</tr>
<tr>
<td>Superior</td>
<td>0.56</td>
<td>0.05</td>
<td>1.58</td>
<td>0.14</td>
</tr>
<tr>
<td>Problem × experiment type adequacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length problem</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inadequate</td>
<td>0.45</td>
<td>0.09</td>
<td>2.39</td>
<td>0.14</td>
</tr>
<tr>
<td>Adequate</td>
<td>0.67</td>
<td>0.05</td>
<td>2.58</td>
<td>0.08</td>
</tr>
<tr>
<td>Superior</td>
<td>0.56</td>
<td>0.09</td>
<td>2.48</td>
<td>0.15</td>
</tr>
<tr>
<td>Mass problem</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inadequate</td>
<td>0.07</td>
<td>0.05</td>
<td>0.32</td>
<td>0.18</td>
</tr>
<tr>
<td>Adequate</td>
<td>0.51</td>
<td>0.07</td>
<td>0.98</td>
<td>0.17</td>
</tr>
<tr>
<td>Superior</td>
<td>0.57</td>
<td>0.07</td>
<td>1.10</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Note: Conclusions were coded as 0 for inaccurate responses and 1 for accurate responses.
Instead, significant predictor covariates included: problem type, experimentation adequacy, and time (hypotheses 1b, 1c, and 3) ($\beta_0 = -1.31, \chi^2 = 6.55, df = 1, p = 0.01$; $\beta_{problem} = 1.60, \chi^2 = 56.10, df = 1, p < 0.0001$; $\beta_{experimentation} = 0.49, \chi^2 = 12.89, df = 1, p = 0.0003$; $\beta_{time} = 0.37, \chi^2 = 5.66, df = 1, p = 0.02$; Goodness of fit: $\chi^2 = 701.03, df = 555$). The final best-fit model was developed by including the significant predictor covariates and their interaction combinations in the three models that follow. First, all possible interaction terms (i.e., problem type*experimentation adequacy, problem type*time, experimentation adequacy*time, and problem type*experimentation adequacy*time) were added to the model. Only problem type was a significant predictor ($\beta_{problem} = 2.73, \chi^2 = 25.20, df = 1, p < 0.0001$). Then, with the 3-way interaction term removed, the pairwise interaction model: conclusion accuracy = problem type + experimentation adequacy + time + experimentation adequacy*problem type + problem type*time + experimentation adequacy*time was considered. In this model, problem type, experimentation adequacy, and the problem type by time interaction term were significant predictor covariates, and the experimentation adequacy by problem type interaction term approached significance ($\beta_0 = -2.49, \chi^2 = 29.48, df = 1, p < 0.0001$; $\beta_{problem} = 3.20, \chi^2 = 45.97, df = 1, p < 0.0001$; $\beta_{experimentation} = 0.68, \chi^2 = 7.18, df = 1, p = 0.007$; $\beta_{time} = 1.19, \chi^2 = 8.12, df = 1, p = 0.004$; $\beta_{problem*time} = -1.86, \chi^2 = 27.67, df = 1, p < 0.0001$; $\beta_{experimentation*problem} = -0.51, \chi^2 = 3.17, df = 1, p = 0.08$, ns; Goodness of fit: $\chi^2 = 673.27, df = 554$).

The final, best-fit model was conclusion accuracy = problem type + experimentation adequacy + time + problem type*time ($\beta_0 = -2.27, \chi^2 = 56.10, df = 1, p < 0.0001$; $\beta_{problem} = 2.67, \chi^2 = 73.79, df = 1, p < 0.0001$; $\beta_{experimentation} = 0.52, \chi^2 = 13.03, df = 1, p = 0.0003$; $\beta_{time} = 1.39, \chi^2 = 36.72, df = 1, p < 0.0001$; $\beta_{problem*time} = -1.91, \chi^2 = 31.02, df = 1, p < 0.0001$; Goodness of fit: $\chi^2 = 673.77, df = 556$). In sum, conclusion accuracy was still significantly predicted by problem type and experimentation adequacy when considering the role of a 12-week delay. Additionally though, the delay itself significantly predicted conclusion accuracy. The effect of the 12-week delay was a function of the problem, as demonstrated by the significant problem × time interaction term (see Fig. 4).

3.3.2. Concept transfer score

For the initial concept transfer outcome model, problem type, experimentation adequacy, and time, but neither observation clarity nor experiment type were significant predictor covariates of concept transfer scores (hypotheses 1, 2, and 3) ($\beta_0 = 0.41, \chi^2 = 2.99, df = 1, p = 0.08$, ns; $\beta_{problem} = 1.72, \chi^2 = 296.53, df = 1, p < 0.0001$; $\beta_{experimentation} = 0.19, \chi^2 = 8.70, df = 1, p = 0.003$; $\beta_{time} = 0.43, \chi^2 = 33.87, df = 1, p < 0.0001$; Goodness of fit: $\chi^2 = 568.12, df = 557$). Removing observation clarity and experiment type and adding the pairwise and 3-way interaction terms, the final full model revealed that problem type and the 3-way problem type by experimentation adequacy by time interaction term were significant predictors ($\beta_{problem} = 0.28, \chi^2 = 6.71, df = 1, p = 0.01$; $\beta_{problem} = 1.59, \chi^2 = 64.80, df = 1, p < 0.0001$; $\beta_{problem*experimentation*time} = -0.56, \chi^2 = 7.73, df = 1, p = 0.005$; Goodness of fit: $\chi^2 = 567.15, df = 555$). (See Table 3 for concept transfer score means as a function of the significant 3-way interaction term.) Most importantly, as indicated in Table 3, mean concept transfer score was significantly reduced over time for the mass problem regardless of experimentation adequacy. For the length problem, mean concept transfer score was significantly reduced over time as well, but only for inadequate experiments ($M_{time 1} = 2.39, M_{time 2} = 1.70$).

3.4. Differences in experimentation adequacy

Since experiment type was not a significant predictor of conclusion accuracy or conceptual transfer, its role was further examined. First, we expected a significant relationship between the experiment type and experimentation adequacy. A chi-square test of the relationship between experiment type and experimentation adequacy indicated that experimentation adequacy significantly differed across experiment type ($\chi^2 (1, N = 147) = 40.85, p < 0.001$). Specifically, 33.3% of the participants in the computer condition versus 9.5% of the participants in the physical condition performed inadequate experiments. This was despite no significant difference in enjoyment or effort across experiment type (enjoyment ratings: $M_{physical} = 3.54, SD = 0.71, M_{computer} = 3.68, SD = 0.68$; effort ratings: $M_{physical} = 3.31, SD = 0.82, M_{computer} = 3.37, SD = 0.88$).

Next, post hoc analysis turned to exploring the problem type by experimentation adequacy interaction (see Section 3.2). A chi-square test of a 2 (problem type: length vs. mass) × 3 (experimentation adequacy: inadequate, adequate, superior) contingency table analysis confirmed that experimentation adequacy differed across problem type ($\chi^2 (1, N = 147) = 11.96, p = 0.003$). For the length problem, 22.1% of participants ($E = 44.3, O = 31$) performed

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Mean concept transfer score as a function of the significant three-way interaction term (Problem × time × experimentation adequacy).</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Length problem</td>
<td></td>
</tr>
<tr>
<td>Time 1</td>
<td></td>
</tr>
<tr>
<td>Inadequate</td>
<td>2.39</td>
</tr>
<tr>
<td>Adequate</td>
<td>2.58</td>
</tr>
<tr>
<td>Superior</td>
<td>2.48</td>
</tr>
<tr>
<td>Time 2</td>
<td></td>
</tr>
<tr>
<td>Inadequate</td>
<td>1.70</td>
</tr>
<tr>
<td>Adequate</td>
<td>2.34</td>
</tr>
<tr>
<td>Superior</td>
<td>2.35</td>
</tr>
<tr>
<td>Mass problem</td>
<td></td>
</tr>
<tr>
<td>Time 1</td>
<td></td>
</tr>
<tr>
<td>Inadequate</td>
<td>0.32</td>
</tr>
<tr>
<td>Adequate</td>
<td>0.98</td>
</tr>
<tr>
<td>Superior</td>
<td>1.10</td>
</tr>
<tr>
<td>Time 2</td>
<td></td>
</tr>
<tr>
<td>Inadequate</td>
<td>0.04</td>
</tr>
<tr>
<td>Adequate</td>
<td>0.65</td>
</tr>
<tr>
<td>Superior</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Note: Concept transfer scores were coded as 0 for no correct responses and 3 for all correct responses.
superior experiments, whereas for the mass problem, 41.1% of participants ($E = 44.7, O = 58$) performed superior experiments.

It follows that participants were performing experiments differently based on experiment type and problem type. To better understand these differences in experimentation, experimentation adequacy scores were parsed into 1) the presence of control of variables (CV) and 2) the number of trials employed (NT). CV was coded as present or not (i.e., 0 = no CV, 1 = CV), and NT was coded as count data for each of the two problem types. For an experiment to be considered “adequate” in the previous analysis, CV had to be present for the experiment to be assigned to the given problem. This new coding scheme allowed considering the NT for experiments with no CV. For example, if a participant had two experiments—the first controlling for length and the second controlling for nothing—it was assumed that the second experiment should have controlled for mass and was designed to address that problem. In that case, the participant’s experiment received an NT score based on the number of trials s/he ran despite receiving a 0 for CV score. Seven out of 280 cases were excluded because they contained multiple experiments with no CV—making it impossible to assign NT based on this rule.

3.4.1. Specific experimentation differences across experiment type

Because CV scores were categorical, a chi-square test was conducted and indicated that CV varied significantly for the hands-on and simulated experiments ($\chi^2 = 29.6; O_{\text{computer}} = 49; O_{\text{hands-on}} = 12; E_{\text{Hands-on}} = 12; E_{\text{Computer}} = 12$). CV and NT, with greater prevalence of CV for the hands-on experiments (91% of experiments) than for simulated experiments (66% of experiments). In terms of NT, a one-way ANOVA was conducted since data was continuous. With experiment type as the independent factor and NT as the dependent factor, the ANOVA revealed that NT also differed significantly across experiment type ($F(1, 272) = 38.96, p < 0.001; d = 0.76$). For hands-on experiments, mean NT was 1.26 (SD = 0.64), and for simulations, mean NT was 2.01 (SD = 1.25). In sum, the difference in experimentation adequacy across experiment type was a function of both CV and NT, with greater prevalence of CV for the hands-on experiment condition and greater mean NT for the computer simulation condition.

3.4.2. Specific experimentation differences across problem type

A chi-square test revealed CV ($M = 0.78, SD = 0.41$) did not vary significantly across problem type ($\chi^2(1, N = 147) = 0.52, p = 0.28$, ns), with 76% of the length experiments controlling for mass and 80% of the mass experiments controlling for length. A one-way ANOVA revealed, however, that NT ($O = 8; M = 1.63, SD = 1.06$) differed significantly across problem type ($F(1, 272) = 5.25, p = 0.02; d = 0.29$), with more trials run for the mass problem ($M = 1.79, SD = 1.16$) than for the length problem ($M = 1.48, SD = 0.98$). Therefore, the previously reported difference in experimentation adequacy across problem type was a function of participants performing more experimental trials for the mass problem than for the length problem.

4. Discussion

4.1. The role of observation clarity

Based on the suggestion that students were capable of altering inaccurate beliefs in the presence of “successful” observation, the expectation here was that clarifying observation would benefit conceptual understanding (Hypothesis 1a) (see Chinn & Malhotra, 2002). This was not the case for the mass problem. Instead, attempts to make observations more clear may not be as simple as expected. One possibility is that the slow motion result provided here is perfectly clear, but only contributes to conceptual understanding if participants conduct the experiment correctly. Another possibility is that observation clarity is not sufficient for benefiting conceptual understanding, especially in cases requiring conceptual change. In the context of conceptual change, Chinn and Brewer (1993) outline six alternatives to accepting anomalous data: (1) ignoring the data, (2) rejecting the data, (3) excluding the data, (4) holding the data in abeyance, (5) reinterpretating the data without editing previous theory, and (6) reinterpretating the data while making peripheral changes to previous theory. Through the current observation clarity manipulation, we attempted to artificially provide a scenario in which participants could not easily ignore or reject the evidence that a bob of greater mass swings at the same rate as a bob of less mass. However, participants were still able to exclude, put off, or reinterpret data during their interpretation of experiments’ results. Thus, the current findings suggest that observation clarity may be less influential in accepting data when scientific reasoning (e.g., experimentation strategies) and interpretation (e.g., conclusion drawing) are involved. In that case, computer simulations’ built-in slow motion operations may not be as valuable as expected for all problem types.

4.2. The role of problem type

As expected, problem type predicted participants’ conclusions and conceptual transfer (Hypothesis 1b). Participants drew more accurate conclusions and answered more applied questions correctly for the length problem. On the one hand, this finding may seem commonplace. After all, participants were initially more likely to know that string length would affect the bob’s speed. Of course they would be more likely to draw a correct conclusion regarding string length. This finding further supports the notion that evidence may enrich current knowledge structures. On the other hand, despite experimentation—with a physical apparatus, on the computer, with a slow motion replay, with adequate (or even superior) experimentation—participants are unable to alter their inaccurate understanding regarding the effect of mass on rate of travel. This further confirms the notion that while evidence may also be applied to the restructuring of incorrect knowledge structures, the process is slow and gradual, requiring more than one instance of cognitive conflict. After a 12-week delay, the effect of problem type is even more pronounced as reported conceptual understanding related to the length problem does not decline in accuracy over time to the same extent that it does for the mass problem. In other words, of the few who are able to report accurate conceptual understanding at Time 1, even fewer are able to maintain this understanding over a twelve-week period. Again, this supports a conceptual change approach that acknowledges the restructuring of conceptual knowledge as a process that may require more than a brief or isolated interaction with evidence.

4.3. The role of experimentation adequacy

What is most informative and intriguing is the powerful influence of experimentation adequacy in conceptual change and its relationship with the kind of experiment conducted and the kind of problem addressed. Surprisingly, adequate, and not superior experimentalization led to the best participant outcomes at immediate posttest, which introduces the suggestion that multiple trials are not necessary for accurate conclusions. Considering the effect more closely, adequate experimentation (i.e., an experiment with one trial) resulted in the best outcomes for the length problem (Hypothesis 1c). Outcomes did not differ between adequate and
superior experiments for the mass problem, but inadequate experiments were especially bad at informing conceptual change. In other words, for a problem that requires only knowledge structure enrichment (like the length problem here), one trial is sufficient. Contrary to the common expectation that seeing an effect multiple times will strengthen one’s belief in that effect though, for a problem that requires knowledge restructuring, running more trials was not associated with uncovering and, importantly, accepting an accurate concept. Instead, repeated experiences do not function to encourage conceptual change; they most likely represent strong intentions to avoid conceptual change.

This finding works in conjunction with reasoning considerations of argument evaluation bias and social cognitive considerations of motivated skepticism, to suggest that students are being more critical of experiments that demonstrate something with which they disagree. Examined in adult populations, argument evaluation bias considers the cognitive processing of preference-inconsistent information in relationship to self-serving bias (Ditto & Lopez, 1992; Kunda, 1990; Stanovich & West, 1997). In these contexts, people engage in more detailed or more careful evaluation of information/arguments that oppose their preferences/beliefs. This research seems to predict that individuals engaged in heightened critical thinking would be more likely to control variables when their beliefs are challenged. However, this was not the case for the current study’s mass problem. Within the context of inaccurate physical science knowledge, the comparable approach for being “more critical” appears to be repeating experimental trials. Perhaps, in their unwillingness to believe what they see, adolescents become increasingly motivated instead to see what they believe. Therefore, belief-inconsistent experimentation requires more trials in hopes of reaching this highly motivated goal of attaining an observed result that matches belief. It is unclear from the findings here why an individual’s skill in designing a controlled experiment and his or her implied implicit understanding of the importance of such actions for testing and interpreting outcomes does not translate into conceptual change. Thus, this finding again lends support to the role of interpretation skills – above observation and reasoning skills – in the case of conceptual restructuring.

Finally, considering the profound effect of experimentation adequacy over time, concept transfer scores after 12 weeks suggest that the role of experimentation adequacy in conceptual understanding is even more significant over time. As Table 3 shows, this was the case for even the length problem, with the concept transfer score mean significantly decreasing over twelve weeks for those participants that conducted inadequate experiments. This is a powerful suggestion that a poor experiment may not only disable students from developing accurate concepts, but that over time it may cause students to inappropriately restructure, or change, the accurate concepts they once held.

4.4. The role of experiment type

Although hands-on experimentation did not result in more accurate conceptual understanding than simulation as was expected, the way experiments were performed across these two contexts differed distinctly (Hypothesis 2). For computer simulations, participants were less likely to control variables and were more likely to run additional trials than for the hands-on experiments. Recall that participants were instructed to repeat each trial to mimic the video replay that physical condition participants were exposed to, possibly making this measure an observation clarity manipulation check. As such this distinction will not be further discussed. The presence of less control of variables during computer simulated experiments however is not a manipulation check, but instead supports previous findings regarding the need for carefully constructed computer environments and the poor ability of adolescents to design unconfounded experiments—especially in simulated microworlds (Brna, 1987; Chang et al., 2008; Lewis, Stern, & Linn, 1993; White, 1984). The current study design expands previous findings by providing a comparison of adolescents’ natural variable control in isolated computer simulations and hands-on experiments.

The difference in variable control between computer and hands-on experiments may be expected as a result of adolescents’ familiarity and comfort with simulated environments. They may be more likely to “play” in the computer-simulated environment—suggesting less regard for experimental control in this condition. The measures of enjoyment and effort indicated no preference for computer simulations over the physical pendulum apparatus. Instead, students may be more entertained by computer simulations, causing them to alter variables based on entertainment value rather than experimental value (White, 1993). Unfortunately, this form of entertainment is not constructive in terms of conceptual understanding. Since enjoyment and effort were not increased and variables were controlled less often, the suggestion is that unsupported computer simulation is not the best mode of experimentation. This suggestion is consistent with previous work regarding the effectiveness of simulation-based learning in scientific domains (e.g., Bangert-Drowns, Kulik, & Kulik, 1985; Carlsen & Andre, 1992; Regan & Sheppard, 1996). Furthermore, it is important to note that some research has even suggested that while supported computer simulations (i.e., those with experimental prompts) improve the experimentation strategies themselves, conceptual change outcomes are not positively affected (Rivers & Vockell, 1987). We found here that without any form of guidance, computer simulations do not positively affect the isolated improvement of experimentation strategies or conceptual understanding.

4.5. Limitations and future directions

While the physical apparatus used here hypothetically allowed for as much freedom as the computer simulation, it may not have encouraged it at the same level. For example, although students in both environments could test a broad and continuous range of string lengths, the physical apparatus had three distinct lengths marked, while the computer simulation tool indicated no preferential length. The physical apparatus also only provided two options for varying mass (a plastic ball and a steel ball). While participants in the computer simulation condition only needed to use two distinct masses to test their question, they had the opportunity to use a wide range of masses.

Experimentation adequacy scoring also was limited in the current study. For instance, while the control of length and mass were considered, there was no way to more carefully assess participants as experimenters, such as monitoring height of drop or force exerted during drop (i.e., a forced swing, instead of a passive drop). During the additional trials conducted for the mass problem, participants may have actually altered variables, such as bob release time, to better observe what they believed. As a case example, an experimenter had a participant—who believed the lighter ball would travel faster—slightly, but visibly, shorten the length of the plastic, lighter, ball “to make them even.” This adjustment was made after clearly, and admittedly, seeing the two balls travel the same speed in a slow motion replay. As a result of shortening the plastic ball’s string length, the plastic ball now traveled visibly faster than the steel ball. The participant then was satisfied with the result and selected the conclusion answer choice that matched his
inaccurate knowledge. The current study's experimentation adequacy scoring did not provide for any assessment of the possibility that participants made miniscule manipulations related to string length, ball release time, etc., in order to reinforce their knowledge with consistent evidence.

Following the current findings, there remains much to be considered empirically in the realm of computer-simulated learning. First, future research should more closely examine the role of simulation in adolescents’ cognitive processing. Specifically, future studies should assess how observation, general reasoning, and interpretation processes function during simulated experiments. Importantly, how each process contributes to the restructuring of knowledge through simulated scientific experiences should be more fully explored. Second, the current methodological procedure and findings should contribute to considerations of how best to control across real-life and simulated environments. Future studies should aim to better control and match the environments in terms of flexibility and freedom of choice in experimentation. While some may consider such flexibility and freedom a fundamental and integral benefit of simulated environments, it is vital in a discussion of the underlying cognitive and contextual mechanisms responsible for optimal learning to limit such flexibility and freedom in basic, lab studies of simulations. Finally, additional work should be aimed at considering adolescent student perception of and motivation towards the simulated learning environment. Such work has the potential to extend beyond conceptual change underpinnings and to begin to address the influence of simulated environments on attitudes toward science as well as differences in achievement orientation.

5. Conclusions

Overall, a good experiment, whether conducted on a computer or in real life, even with stimuli slowed down for better observations, may not contribute to the restructuring of an adolescent’s inaccurate conceptual understanding. More importantly, a bad experiment may be enough to alter students’ accurate conceptual understanding. The controlling of variables in experimental design—which the majority of adolescents are capable of for the current scenarios—is of utmost importance in both enriching and restructuring adolescents’ knowledge structures. Addressing previous suggestions that unclear data are often responsible for the persistence of inaccurate physical science concepts, creating more clear observations had no effect on adolescents’ conceptual understanding. So we recommend that researchers and educators alike endorse the use of simulated learning environments with greater caution. While more research is necessary to better understand the role of computer simulation in learning, isolated physical science computer simulations are no better in informing conceptual knowledge than are real-life experiments, and in fact, they may be worse in encouraging students’ careful experimental design.

Appendix

Appendix A

Demographic Information

Age: __________

Gender (circle one):  Male   Female

Grade in school (circle one):  7th  8th

Have you taken a physical science class? (circle one):  yes  no

If so, what was the name of the class: _______________________

Appendix B

Predictions

The following questions were asked verbally:

**Same mass, different string length**

1. “When these two weights swing from here, do you think:
   ____ one of the weights will come back to the starting place before the other, or
   ____ they will return at the same time?”

   a) If you think one will come back before the other, which will come back **first**:
      ____ the bob on the longer string
      ____ the bob on the shorter string

2. “Are you guessing?”
   ____ yes  ____ no
Appendix C

Conclusions

Conclusion (circle one):

With a longer string, a pendulum will swing over and back
(a) faster than with a shorter string,
(b) slower than with a shorter string,
-or-
(c) at the same speed as with a shorter string

Conclusion (circle one):

A greater (or bigger) mass will swing over and back
(a) faster than a smaller mass,
(b) slower than a smaller mass,
-or-
(c) at the same speed as a smaller mass

Appendix D

Motivation Measure

The researchers for this project are interested in whether or not students your age would enjoy the experiment you just conducted. To help us decide if we should give this activity to other students your age, please answer the following questions.

How much do you think other students your age would enjoy this experiment?
(Circle one number on the scale below.)

Not enjoy it at all                   Love it!
                                        1  2  3  4  5

How much effort do you think other students your age would put into this activity?
In other words, how hard do you think they would work at conducting the experiment to find the right answer?
(Circle one number on the scale below.)

No effort                     Lots of effort
                                           1  2  3  4  5
Appendix E

Generalization Questions

1) A baby swing is higher off the ground than the rest of the swings on a swing set. A child releases the baby swing and a regular swing at the same time, from the same height. Which swing will come back to her first?
   (a) the baby swing
   (b) the regular swing
   (c) The swings will come back to her at the same time.

2) A bowling ball and a tennis ball are dropped from the same height on a pendulum. Assuming they are the same distance from the top of the pendulum, which ball will swing over and back first?
   (a) the bowling ball
   (b) the tennis ball
   (c) They will swing over and back at the same time.

3) What effect does an increase in the length of string attached to a ball in a pendulum have on the speed the ball swings?
   (a) As the length increases, the speed increases.
   (b) As the length increases, the speed decreases.
   (c) As the length increases, the speed does not change.

4) What effect does an increase in the weight of a ball have on the speed the ball swings on a pendulum?
   (a) As the weight increases, the speed increases.
   (b) As the weight increases, the speed decreases.
   (c) As the weight increases, the speed does not change.

5) Two objects the same shape and mass are released to swing on two pendulums simultaneously. One is attached to a 10 cm length string and the other is attached 20 cm length string. Which object will swing faster?
   (a) the object attached to the 10cm length string
   (b) the object attached to the 20cm length string
   (c) They will swing the same speed.

6) A child weighing 50 pounds and a child weighing 70 pounds are swinging at the playground. Both of the children started swinging at the same height at the same time. Which child is swinging faster?
   (a) the 50 pound child
   (b) the 70 pound child
   (c) They are swinging at the same rate.

References


