The Impact of Interactivity on Simulation-Based Science Inquiry with Variable-Setting Controls

Jung Aa Moon (jmoon001@ets.org)
Educational Testing Service
Princeton, NJ 08541 USA

Michelle LaMar (mlamar@ets.org)
Educational Testing Service
San Francisco, CA 94105 USA

Carol M. Forsyth (cforsyth@ets.org)
Educational Testing Service
San Francisco, CA 94105 USA

Madeleine Keehner (mkeehner@ets.org)
Educational Testing Service
Princeton, NJ 08541 USA

Abstract

The current study investigated how interactivity of simulation controls affects data collection in science inquiry. A chemistry simulation was designed to allow either low or high interactivity in setting experimental variables. Adult participants were randomly assigned to one of the interactivity conditions and solved a series of assessment items. The results from the first item indicated that the highly interactive controls posed challenges in conducting a thorough investigation. Performance in the last item which is a repetition of the first item suggested that the participants were able to overcome the initial challenges over the course of their investigations. The results provide implications for designing educational simulations for learning and assessment.

Keywords: interactivity; simulation; science inquiry; education; assessment

Introduction

Traditional science assessments conducted on a large scale have until recently been limited mostly to the paper-and-pencil format (e.g., Shavelson, Carey, & Webb, 1990). These assessments often measured test-takers' prior knowledge of science facts and principles by collecting their final responses to multiple choice items. In the last few decades, however, criticisms of the traditional assessments and advances in educational technology have spurred a growing interest in using simulation-based tasks for science assessments (Gobert, Sao Pedro, Raziuddin, & Baker, 2013; Wieman, Adams, & Perkins, 2008).

The simulation-based assessments typically provide test-takers with interactive tools that support multiple science inquiry steps such as designing experiments and collecting data to test a scientific hypothesis. The interactive nature of these simulations raises a question of how simulation interactivity impacts cognition and educational outcomes. Answering this question is important for designing simulation environments where test-takers or learners have a sufficient level of freedom to conduct science inquiry. It is also important for ensuring that simulation interactivity does not hinder valid measurements of knowledge and skills.

Interactivity

The topic of interactivity has been actively researched in the field of multimedia learning. While there are various approaches to how interactivity can be interpreted (Domagk, Schwartz, & Plass, 2010), a number of studies have associated interactivity with the amount of control learners have over various aspects of the learning system (Kalyuga, 2007; Moreno & Mayer, 2007). For instance, studies have manipulated whether learners can control the sequence of the learning materials (Swaak & de Jong, 2007) and whether they can select their own answers rather than to receive the correct answers selected by the system (Moreno & Mayer, 2005). While some studies have found that learner-controlled learning results in better outcomes than system-controlled learning (Hasler, Kersten, & Sweller, 2007; Moreno & Valdez, 2005), others suggest negative outcomes associated with the greater learner control (e.g., Moreno & Valdez, 2005; Tuovinen & Sweller, 1999).

Cognitive load theory (Sweller, 1994) provides accounts for those seemingly mixed findings on interactivity. The main claim of the theory is that cognitive load irrelevant to learning results in negative learning outcomes as it takes away limited cognitive resources that could otherwise be used for learning. On the other hand, cognitive load put into creating schemas (i.e., cognitive structures that allow organization of information) is beneficial for learning. One implication of the cognitive load theory for interactivity research is that interactive systems can have positive outcomes if they engage learners in deeper cognitive processing germane to learning (Hasler, Kersten, & Sweller, 2007; Kalyuga, 2007).

Interactivity in Science Inquiry

The prior research on interactivity raises a question of how those findings apply to science simulations which typically provide individuals with a fair amount of freedom to conduct science inquiry. In particular, one of the distinctive features of simulation-based science assessments is that they allow test-takers to collect their own data for scientific hypothesis testing. Test-takers can make their own decisions on various aspects of data collection such as how much data to collect and how exhaustively to search the experimental space. This contrasts to traditional assessments in which test-takers are typically asked to draw a conclusion based on the data prepared by assessment developers.

Collection of adequate data is one of the critical inquiry skills involved in scientific investigations (Kuhn, Schauble, & Garcia-Mila, 1992). Prior research found that people often exhibit suboptimal data collection behaviors. For instance, people often jump to a conclusion based on limited observations without sufficiently sampling variables (Harrison & Schunn, 2004; Kuhn et al., 1992). To understand how simulation interactivity impacts data collection behaviors and inquiry outcomes, we manipulated the design of slider/button controls in a chemistry simulation. While there is a fair amount of human-computer interaction research on how design of slider controls impacts user behaviors (e.g., Roster, Lucianetti, & Albaum, 2015), few studies (e.g., Renken & Nunez, 2013) assessed its impacts on science inquiry behaviors.

Concentration Simulation

Concentration simulation developed by PhET (Wieman, Adams, & Perkins, 2008) was modified for the purpose of our study. The simulation is an HTML5 application written in JavaScript and delivered through a standard web browser. The simulation (Figure 1) allows one to design and run simulation trials to investigate the relationship among solute, water, and concentration level. The top left panel displays the slider and button controls used to set the amounts of solute and water. Once a trial is run, the solute and water are mixed to form a solution. The resulting concentration level of the solution and the variable settings for the trial appear in the data table located in the bottom left panel. In a typical screen, the right panel displays an assessment item and response options and/or text entry.

Method

Participants

Adult participants (N = 308) recruited through Amazon Mechanical Turk completed the study for monetary compensation ($5). There was no particular entry condition for participation. Background information about participants was collected in a separate survey to which 248 of the participants responded (111 females and 137 males, mean age 35, age ranged from 20 to 61). Participants were randomly assigned to the interactivity conditions.

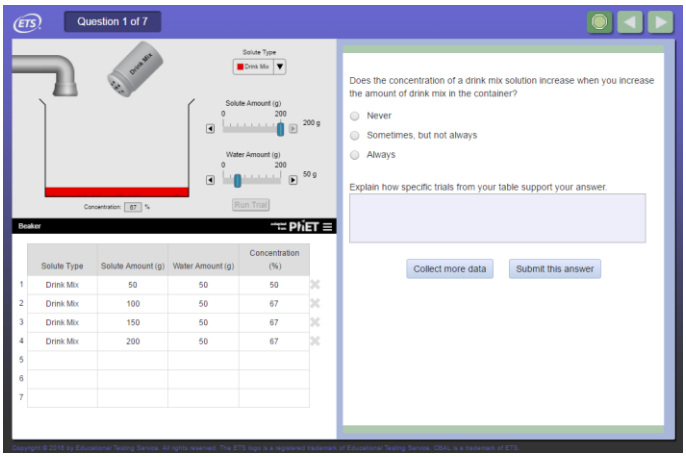

Figure 1. Concentration simulation.

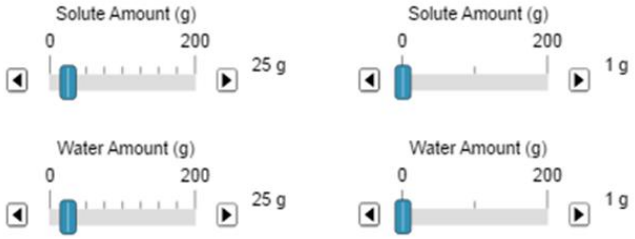

Figure 2. Variable-setting controls in the low (left) and high (right) interactivity conditions.

Design

The study involved a between-subject design in which three features of the simulation were manipulated to allow different levels of interactivity. Those features were: 1) the slider and button controls, 2) the amount of the data one could keep in the data table, and 3) the capability to re-order the rows in the data table. Due to the space limit, the current paper focuses on discussing the results of the slider/button control manipulation. These controls (Figure 2) were manipulated to allow different amounts of control over the values one could choose for the solute and water variables. While the range for each variable was the same in both conditions (0-200 g), the smallest amount of adjustment one could make using the controls was 25 g in the low interactivity condition and 1 g in the high interactivity condition. This manipulation offered a far greater number of choices for the high condition than for the low condition. While the high condition participants could set any of the 201 integer values for each variable, the low condition participants' choices were limited to 9 values (0, 25… 200).

In order to adjust a variable, participants could drag the slider (blue rectangle in Figure 2) to a desired location on the scale and release it. Once released, the slider automatically snapped to the closest value available in the relevant condition. Participants could also click the tweak buttons (buttons with an arrow sign) located next to each scale. A button click incurred 25 g of change (left button for decrement, right button for increment) in the low condition

and 1 g of change in the high condition. In the low condition, a tick mark was placed at every available value on the scale. In the high condition, only three tick marks were placed (0, 100, and 200 g) due to programming constraints and space limit. We intentionally chose not to make the visuals of the sliders the same for the two conditions because of a concern that doing so would bias the participants in the high condition prefer those particular values with tick marks.

Experiment

The experiment consisted of a brief tutorial session followed by a main session. During the tutorial session, participants were introduced to the terminology used in the simulation and were familiarized with various simulation features including the slider and button controls. During the main session, participants solved seven assessment items presented one at a time in a self-paced manner. Participants were asked to use the simulation to conduct investigations necessary for answering each item. The order of the seven items was the same for the two conditions.

These items were designed to assess understanding of the impacts of solute and water amounts on the concentration level. All the items were designed around the concept of saturation (i.e., concentration[1] does not further increase once maximum is reached). The first and last items were identical except for their relative locations in the item series. These two items were different from the five items in the middle in two major aspects. First, some of the middle items directed participants to investigate the impacts of the variables on the concentration level under a specified condition such as with a larger range of solute values or with a smaller amount of water. These directed investigations were expected to facilitate observation of saturation. Second, the first and last items involved a solute type ("drink mix") different from the one used in the middle items ("chemical A"). While the two solutes are different, we expected that some knowledge gained from investigations with one solute could be transferred to investigations with the other solute. In a sense, the first and last items can be viewed as pre- and post-tests that allow assessment of learning gained through investigations. Due to space limit, the current paper focuses on the results in these two items. Participants' performance in the two conditions did not significantly differ in any of the middle items.

Task

In the first and last items, participants were asked to test the following statement: "Does the concentration of a drink mix solution increase when you increase the amount of drink mix in the container?" Participants were asked to choose one of the three response options ("never", "sometimes, but not always", and "always") and explain how specific trials from the data table support their answer by typing their responses to the text entry. The correct answer is "sometimes, but not always" because adding more drink mix does not further increase the concentration level once the saturation point is reached.

Success in these items largely depends on the ability to sample sufficient ranges of the variables. Insufficient sampling is likely to lead to either "always" (concentration keeps increasing before the saturation point is reached) or "never" response (no further increase in concentration past the saturation point). In order for the saturation point to be observed, the amount of water has to be less than or equal to 75 g. The amount of drink mix that results in saturation depends on the amount of water.

Results

We describe the results on response choice (Table 1) and multiple measures of inquiry behaviors (Table 2). The numbers in the parentheses are adjusted standardized residuals in Table 1 and standard deviations in Table 2. The F-stats in Table 2 are from the repeated measures ANOVA with condition (between) and item (within) as variables.

First Item

Choice of Response Table 1(A) shows how many participants in each condition chose each of the response options. Only 37 of the low and 26 of the high participants chose the correct "sometimes" response. The Chi-square test was significant (χ^2 (2, N = 308) = 8.4, p = .015), indicating the significant effect of being in the interactivity condition on the choice of response. The high participants had a greater tendency to choose the "always" response compared with the low participants. When the incorrect responses ("never" & "always") were combined together, the Chi-square test was not significant (p = .107).

Time on Task The amount of time spent on solving the item did not significantly differ between the low (143.6 s) and high (156.5 s) conditions, t(305) = 1.4, p = .158.

Number of Trials The average number of trials run was significantly greater in the low condition (low: 5.0, high: 4.3, t(303) = 2.5, p = .012).

Table 1: Choice of response

(A) First item

	Never	Sometimes	Always	Total
Low	16 (2.2)	37 (1.6)	100 (-2.7)	153
High	6 (-2.2)	26 (-1.6)	123 (2.7)	155
Total	22	63	223	308

(B) Last item

	Never	Sometimes	Always	Total
Low	6 (1)	90 (1)	57 (-1.4)	153
High	3 (-1)	82 (-1)	70 (1.4)	155
Total	9	172	127	308

[1] Before saturation is reached, concentration is mass of solute divided by mass of solution (i.e., solute plus water).

Table 2: Results on inquiry measures.

	First item		Last item		F-stat		
	Low	High	Low	High	Condition (C)	Item (I)	C * I
Time on task (second)	143.6 (71.7)	156.5 (87.6)	143.6 (82.6)	151.6 (96.5)	1.1	.1	.4
Number of trials	5.0 (2.8)	4.3 (2.1)	7.5 (4.2)	6.9 (4.7)	5.1^{*}	96.7^{*}	.0
Solute sampling range	104.2 (54.3)	91.0 (52.4)	139.7 (50.7)	138.8 (50.5)	2.8^{+}	108.8^{*}	2.9^{+}
Water sampling range	32.3 (50.0)	24.3 (46.2)	59.8 (61.6)	58.4 (63.6)	.9	56.0^{*}	.9

*: $p < .05$, $^{+}$: $p < .10$

Sampling Range For each participant who ran at least one trial, the sampling range of each variable was obtained by getting the minimum and maximum values sampled across all trials and subtracting the former from the latter. The average sampling range of the solute variable (black bars in Figure 3) was significantly greater in the low participants (104.2) than in the high participants (91.0) for the first item, $t(303) = 2.2$, $p = .032$. The average sampling range of the water variable (white bars) did not significantly differ between the low (32.3) and high (24.3) conditions, $t(303) = 1.5$, $p = .145$.

The results above suggest that while overall performance did not significantly differ between the conditions, the high participants collected less data and searched a narrower experimental space. A further analysis was performed to understand what might have contributed to their suboptimal data collection behaviors.

Preference for Round Numbers Nowhere in the item were participants asked to try any particular amount of solute or water. The solute and water amounts set by the high participants revealed that even though they could select any of the 201 integer values, they preferred "round" numbers such as multiples of 50 or 100. Across all the trials run by the high participants, 50, 100, and 200g were the three most frequently selected values for both solute and water. These three values accounted for 28% of the solute values and 70% of the water values, which is much higher than chance (.5%). The low participants also frequently selected the three values (43% in solute, 74% in water).

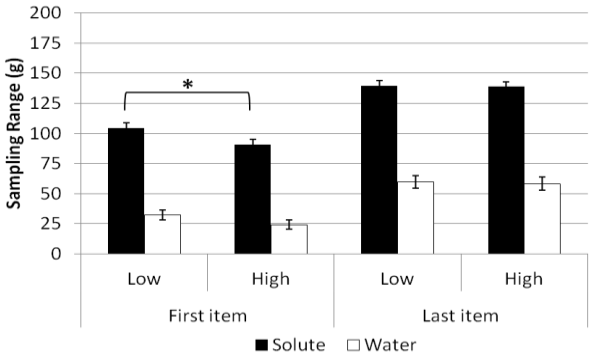

Figure 3. Average sampling ranges of solute and water. The error bars are the standard error of mean.

Interactions with Simulation The preference for the round numbers cost the high participants frequent interactions with the simulation. To adjust solute or water, participants could drag the slider, click the tweak buttons, or combine the two actions. While the two conditions did not differ much in the average number of drag actions (low: 5.4, high: 5.7), the high participants clicked the tweak buttons much more frequently (low: 2.5, high: 22.8). The condition variable was a significant predictor of tweak button click counts in a negative binomial regression (Wald Chi-square ($df = 1$) = 66.4, $p < .001$). Due to the limited interface space, it was almost impossible for the high participants to make very fine adjustments with the slider alone. A likely strategy for setting a round number is to drag the slider close to the desired value and click the tweak buttons to make small adjustments.

Did the high participants' tweak button use influence their sampling range? We categorized participants into two groups based on the type of variable-setting actions: "drag-only" group who used the sliders only and "tweak" group who used the tweak buttons at least once. The numbers of participants in the drag-only and tweak groups were 107 and 43 respectively in the low condition, and 57 and 98 respectively in the high condition. The results of ANOVA on the solute sampling range indicated the significant interaction between condition and action type ($F(1, 301) = 5.9$, $p = .016$) and the main effect of condition ($F(1, 301) = 3.9$, $p = .048$). The sampling range did not differ much between the two groups in the low condition (Figure 4). In the high condition, the tweak group (83.5) sampled a significantly narrower range than the drag-only group (103.8), $t(153) = 2.4$, $p = .020$.

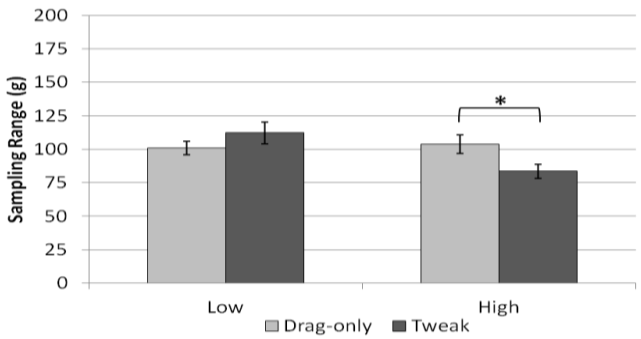

Figure 4. Sampling ranges of solute for the first item. The error bars are the standard error of mean.

The above results suggest that the high participants' preference for the round numbers led to greater workload involved in making frequent interactions with the simulation. One potential interpretation of these results is that experiencing the greater amount of workload deterred the high participants from searching a sufficient experimental space. One may question why the high participants preferred the round numbers despite the workload. A prior study from our laboratory (Koster van Groos & LaMar, 2016) suggests that people choose them because they are easy to remember and keep track of.

Last Item

Choice of Response Compared with the first item, a greater number of participants in both conditions (low: 90, high: 82) chose the correct "sometimes" response in the last item (Table 1B). An exact McNemar's test indicated that the performance improvement was statistically significant ($p < .001$). There was no significant association between the interactivity condition and response choice (χ^2 (2, N = 308) = .98, p = .32).

Time on Task The participants spent about the same amount of time as before (first: 150 s, last: 148 s). There was no significant difference between the low (143.6s) and high (151.6s) conditions, $t(300) = .76$, $p = .443$.

Number of Trials Compared with the first item, the participants ran on average 2.5 more trials in the last item (first: 4.7, last: 7.2). The difference between the conditions was not significant (low: 7.5, high: 6.9, $t(301) = 1.2$, $p = .233$).

Sampling Range The sampling range increased from the first to the last item for both solute (first: 97.5, last: 139.3) and water (first: 28.2, last: 59.1). The difference between conditions was not significant in neither solute (Figure 3, low: 139.7, high: 138.8, $t(289) = .1$, $p = .884$) nor water (low: 59.8, high: 58.4, $t(289) = .2$, $p = .844$) in the last item.

Preference for Round Numbers The participants showed a continued preference for the round numbers. The three most preferred values of 50, 100, and 200 g accounted for about 31% of the solute values and 78% of the water amounts in the high condition. These values accounted for 44% of solute and 73% of water amounts in the low condition.

Interactions with Simulation While the two groups did not differ much in the number of drag actions (low: 7.0, high: 7.5), the high participants used the tweak buttons more frequently (low: 4.4, high: 26.5). The condition variable was a significant predictor of tweak button click counts (Wald Chi-square ($df = 1$) = 39.85, $p < .001$).

In the last item, participants in both conditions showed improved performance in various aspects. A greater number of participants in both conditions selected the correct response. The improved correctness of response choice is consistent with positive changes in their data collection behaviors. While spending about the same amount of time as before, participants collected more data and sampled greater ranges of solute and water.

The results on the sampling range (Figure 3) suggest that the difference between the two conditions became smaller in the last item[2]. The high participants initially sampled a smaller range of solute, yet they later sampled approximately the same range as the low participants. Some other patterns of behaviors, however, did not change much between the two items. The high participants continued to prefer the round numbers despite the greater workload associated with setting those numbers. While the low participants also selected those numbers frequently, their workload is likely to be low because they could set those numbers with a relatively smaller number of button clicks and/or drag actions.

Discussion

The current study investigated how interactivity in variable-setting controls impacts simulation-based science inquiry. Our results suggest that the greater simulation interactivity had initially negative impacts on inquiry performance. Our high participants preferred values that were easier to work with despite the additional workload involved in setting those values. They also ran fewer trials and sampled the experimental variables less exhaustively, likely because experiencing the greater workload hindered thorough scientific investigations.

However, the results in the last item suggest that the initial challenges imposed by the simulation interface became less important over time. On various measures, the low and high participants achieved an equivalent level of performance. It appears that the participants who were initially penalized by the highly interactive simulation interface were able to overcome their challenges over the course of their investigations. One plausible explanation is that observing saturation in the middle items made them realize the importance of sufficiently sampling variables. It is also possible that observing saturation led them to intentionally look for a saturation point in the last item by sampling more exhaustively. The current results alone do not tell us whether the better performance in the last item is due to the improvement of skills, content knowledge, or both. Identifying the contributions of skills and knowledge would be a potentially interesting topic for future research.

Compared with the high condition, the low condition offered much more limited options for variable amounts. One can view that the limited options served as scaffolding for decisions on data collection such as what amount to sample and how exhaustively to sample. The high participants who did not have such scaffolding presumably

[2] Not all participants sampled more than one solute amount in both items. The condition*item interaction in Table 2 is significant when we exclude participants whose sampling range is zero in any of the items, $F(1, 274) = 4.9$, $p < .05$.

had to make those decisions on their own. Their decisions on data collection were less than optimal at least in the first round of their investigations.

While our results appear to suggest negative outcomes of high simulation interactivity, the question of which interactivity level is a better choice cannot be answered without fully understanding the nature of the workload associated with high interactivity. Based on the cognitive load theory (Sweller, 1994), a more appropriate question is whether the extra workload had any relevance to inquiry skills and knowledge. It is possible that our highly interactive controls engaged individuals in deeper cognitive efforts. They had to figure out what is the optimal grain size of data and what is the appropriate sampling range through trials and errors. Their performance improvement suggests that they achieved some learning on those aspects. If the low participants passively followed guidance of the system instead of thinking on their own, their seemingly thorough data collection behaviors may be an overestimation of their genuine inquiry skills and knowledge.

The current discussion focused on the impacts of the variable-setting controls on data collection behaviors. An ongoing analysis is in progress to investigate the impacts of the other two interactivity manipulations (data sorting capability & the amount of data one could keep in the data table) on data collection and organization behaviors. Despite the limited scope, this study provides several implications for designing educational simulations for the purpose of learning and assessment. First, interactive features of simulations can serve as scaffolding that aids cognitive systems to achieve better performance. From the assessment point of view, however, the availability of scaffolding may make it harder to measure genuine knowledge and skills of test-takers. Second, learning and assessment design needs to consider the challenges imposed by simulation interactivity. Especially with highly interactive simulations, providing multiple learning and assessment opportunities seems necessary due to the initial challenges individuals may experience. Overall, the current research suggests the importance of considering how simulation interactivity impacts cognition in learning and assessments.

Acknowledgments

We thank the PhET Interactive Simulations project at the University of Colorado Boulder. Their contributions were funded by the Gordon and Betty Moore Foundation.

References

Domagk, S., Schwartz, R. N., & Plass, J. L. (2010). Interactivity in multimedia learning: An integrated model. *Computers in Human Behavior*, *26*, 1024–1033.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. J. D. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*, 521–563.

Harrison, A. M., & Schunn, C. D. (2004). The transfer of logically general scientific reasoning skills. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 541–546). Erlbaum, Mahwah, NJ.

Hasler, B. S., Kersten, B., & Sweller, J. (2007). Learner control, cognitive load and instructional animation. *Applied Cognitive Psychology*, *21*, 713–729.

Kalyuga, S. (2007). Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective. *Educational Psychology Review*, *19*, 387–399.

Koster van Groos, J. M. & LaMar, M. M. (2016). *Separating inquiry goal and skill during analysis of student performance on a simulation-based science assessment.* Poster presented at the annual conference of the American Educational Research Association, Washington, DC.

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

Moreno, R., & Mayer, R. E. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology*, *97*, 117–128.

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, *19*, 309–326.

Moreno, R., & Valdez, A. (2005). Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Educational Technology Research and Development*, *53*, 35–45.

Renken, M. D., & Nunez, N. (2013). Computer simulations and clear observations do not guarantee conceptual understanding. *Learning and Instruction*, 23(1), 10–23.

Roster, C. A., & Lucianetti, L. (2015). Exploring Slider vs . Categorical Response Formats in Web-Based Surveys. *Journal of Research Practice*, 11(1), 1–15.

Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan*, *71*, 692–697.

Swaak, J., & de Jong, T. (2007). Order or no order: System versus learner control in sequencing simulation-based scientific discovery learning. In F. E. Ritter, J. Nerb, T. M. O'Shea, & E. Lehtinen (Eds.), *In order to learn: How the sequence of topics influences learning*. New York, NY: Oxford.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*, 295–312.

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology*, *91*, 334–341.

Wieman, C. E., Adams, W. K., & Perkins, K. K. (2008). PhET: Simulations that enhance learning. *Science*, *322*, 682–683.