British Journal of Educational Technology *Vol 0 No 0 2018* 1–20
doi:10.1111/bjet.12653

Earth surface modeling for education: How effective is it? Four semesters of classroom tests with WILSIM-GC

Wei Luo , Thomas J. Smith, Kyle Whalley, Andrew Darling, Carol Ormand, Wei-Chen Hung, Jui-Ling Chiang, Jon Pelletier and Kirk Duffin

Wei Luo is a professor with research interests in Web-based technology in enhancing teaching and learning, geomorphology, and GIS applications. Thomas Smith (Department of Educational Technology, Research and Assessment, Northern Illinois University, Dekalb, IL) is a professor with expertise in statistical analysis in education research. Kyle Whalley (Land & Water Resources Department, Dane County, Madison, WI) is a GIS and data management specialist with interests in water resources and environment quality. Andrew Darling (Warner College of Natural Resources, Colorado State University, Fort Collins, CO) is a research scientist with interests in tectonic geomorphology with an emphasis in canyon country of the Colorado River system. Carol Ormand (Science Education Resource Center, Carleton College, Northfield, MN) is a science education and research associate with research interests in education research and spatial thinking. Wei-Chen Hung (Department of Educational Technology, Research and Assessment, Northern Illinois University, Dekalb, IL) is a professor with research interests in problem solving, human computer interface, and performance support system. Jui-Ling Chiang (Department of Educational Technology, Research and Assessment, Northern Illinois University, Dekalb, IL) is a post-doc with research interests in simulation and problem solving in STEM education. Jon Pelletier (Department of Geosciences, University of Arizona, Tucson, AZ) is a professor with expertise in landform evolution and computer modeling. Kirk Duffin (Department of Computer Science, Northern Illinois University, DeKalb, IL) is an associate professor with expertise in computer vision and 3D rendition. Address for correspondence: Wei Luo, Department of Geographical and Atmospheric Sciences, Northern Illinois University, DeKalb, IL 60115, USA. Email: wluo@niu.edu

Abstract

This paper presents results from a randomized experimental design replicated over four semesters that compared students' performance in understanding landform evolution processes as measured by the pretest to posttest score growth between two treatment methods: an online interactive simulation tool and a paper-based exercise. While both methods were shown to be effective at enhancing students' learning of the landform concepts and processes, there was no statistically significant difference in score growth between the two instructional methods. However, the attitudinal survey indicated that students consistently favored the simulation approach over the paper-based exercise. With the simulation method, female students showed greater score growth than males, especially for test items requiring higher level thinking. This indicates that the visually rich interactive simulation tool may be integrated to better support female students' learning in geoscience. Science major students generally outperformed non-science major students in terms of score growth, which suggests that background knowledge played an important role in realizing the potential of computer modeling in enhancing students' learning. Sufficient scaffolding is necessary to maximize the effect of interactive earth surface modeling in geoscience education.

Practitioner Notes

What is already known about this topic
- There have been mixed or inconclusive results regarding the effect of simulations on enhancing students' learning in general;
- Quantitative empirical research on the effect of simulations in geoscience education in comparison with traditional paper-based method has been limited;
- Previous research focused on the gender difference in the attitudes toward the use of technology, but few studies examined the difference in the actual effect of technology use on learning between genders.

What this paper adds
- Paper-based curricular material was as effective as computer-based simulation in enhancing students' learning, but students consistently favored the simulation approach over the paper-based exercise;
- Background knowledge played an important role in realizing the potential of computer modeling in enhancing students' learning;
- With computer-based simulation, female students performed significantly better than male students in answering application-type questions, which require higher level thinking.

Implications for practice and/or policy
- We should take advantage of the fact that students favor computer-based simulation over paper-based approach in learning science;
- Providing students with enough background knowledge and sufficient scaffolding is necessary to maximize the effect of interactive earth surface modeling in geoscience education;
- Both paper-based and simulation approaches of teaching should be integrated to maximize the effect of simulation tools in enhancing students' learning, especially for online or hybrid courses and flipped classrooms.

Introduction

Background on computer simulation

Computer simulations are dynamic, and often interactive, computer models that capture the essential processes behind the phenomena of the real world (Smetana & Bell, 2012). They allow students to manipulate parameters that control the processes and immediately observe the associated results. This dynamic and near instant feedback enables students to better understand these processes and the variables that control them (de Jong, 2006; Luo *et al.*, 2016; Perkins *et al.*, 2012). Thus, computer simulations are well suited for engaging students in inquiry-based exploration of real world processes and have increasingly been used in education to enhance students' learning (Gordin & Pea, 1995). Many previous studies have examined the effectiveness of using computer simulations in science education. Most of these studies have indicated that the interactivity of computer simulation and its ability to engage students are the keys to maximizing its effects in improving students' learning (eg., Day, 2012; Tversky, Morrison, & Betrancourt, 2002). In addition, interactive computer simulations allow students to explore different scenarios and compare/contrast the associated results. This gives students a sense of control and ownership of their learning, which can help build their confidence and promote a positive attitude towards science (Podolefsky, Moore, & Perkins, 2013).

Previous studies on the effectiveness of computer simulation

Geoscience phenomena such as landforms we observe often have taken millions of years to develop, and most of these processes are not repeatable at relevant spatial and temporal scales (Luo *et al.*, 2016). As a result, the computer simulation approach can offer a particularly beneficial means for geoscience students to observe and analyze long term geological processes (Luo *et al.*, 2016). However, previous studies that have quantitatively assessed the effectiveness of computer simulation in geoscience teaching, as opposed to more traditional teaching methods, have been limited and revealed mixed results. Edsall and Wentz (2007) compared students' performance in understanding map projections using a computer-based model versus a physical model and in understanding coastal geomorphology using geographic information system (GIS) produced maps versus paper maps. The authors found that both computer-based methods and traditional methods are effective at improving students' understanding and that computer-based approaches, although generally more appealing to students, are not significantly more beneficial in enabling their understanding of complex concepts. Stumpf, Douglass and Dorn (2008) compared the performance of students who learned desert geomorphology through a virtual field trip with those who went on a traditional (real) field trip. They found no statistically significant difference between the two groups in their basic knowledge about desert geomorphology. However, the qualitative survey data revealed deeper personal ownership of knowledge among real field trip participants. On the other hand, the virtual field trip has tremendous advantages in terms of cost effectiveness and it offers an alternative learning environment, which is especially beneficial for students with physical disabilities (Stumpf *et al.*, 2008).

Similar mixed or inconclusive results have also been reported in the literature regarding the effect of simulations in non-geoscience settings (Bell & Trundle, 2008; Edsall & Wentz, 2007; Scalise *et al.*, 2011). Traditional methods, such as simply lecturing or using a paper-and-pencil exercise, have been found to be as effective as computer simulations; simulations alone are inadequate in helping students understand more complicated processes and concepts because, without necessary support, these more complex simulations can potentially overwhelm or even confuse students (Adams *et al.*, 2008; Podolefsky *et al.*, 2010). Some studies suggest that scaffolding is needed to help students develop enough background knowledge to better understand and take full advantage of computer simulation (eg, Khan, 2011; Schneps *et al.*, 2014). Others argue that too much scaffolding (eg, step-by-step cookbook style guide) can undermine the potential for exploration. Instead, maintaining a proper balance between sufficient guidance and flexibility and freedom for students to explore is a key to success (Adams *et al.*, 2008; Bell & Trundle, 2008). There is a need to conduct more empirical studies with quantitative comparison and experimental control to confirm or resolve these mixed findings, especially in geosciences. Thus, our first set of research questions (RQ) for this study are:

RQ1: What is the effect of the use of computer simulation on students' learning?

RQ2: How much of a role does different background knowledge (of science vs. non-science majors) play in enhancing students' learning when using computer simulations?

Previous studies on gender differences in the effectiveness of computer simulation

Earlier studies on gender and education have focused on the gender differences in attitude toward the use of technology. The stereotypical view is that females have a more negative attitude towards technology than males (eg, Whitley, 1997). While some recent studies found no significant gender difference in students' attitudes toward technology and its use (Rhema & Miliszewska,

2014; eg, Sáinz, Meneses, López, & Fàbregues, 2016), a recent meta-analysis indicated males still hold more favorable attitudes toward technology use than females (Cai, Fan, & Du, 2017).

Previous studies on gender differences in the effectiveness of computer simulation have shown mixed findings. For example, Koh *et al*. (2010) compared engineering students' performance in learning Machining Technology using simulation approach (treatment) with that using traditional lecturing (control). The study revealed that both male and female students in the treatment group were significantly more satisfied and motivated than students in the control group. The male students achieved a significantly higher mean score than the females in the treatment group, but the difference was not statistically significant in the control group, suggesting that simulation based learning experience had a higher positive impact on the male students than that on their female counterparts. Koh *et al*. recommended that instructional strategies should consider gender differences to balance the discrepancy in performance. In contrast, Kickmeier-Rust, Holzinger, Wassertheurer, Hessinger, and Albert (2007) investigated gender differences in medical students' performance in learning blood flow effects using a traditional static image and text-based learning material versus using an interactive computer simulation-based material. The researchers found that both male and female students performed equally well using both types of learning materials and that there were no differences in their learning styles and strategies. Mihindo, Wachanga, and Anditi (2017) conducted a quasi-experimental research comparing the effect of a computer based simulation (CBS) and a traditional teaching method in instructing a chemistry class of secondary school students. While the researchers found that students taught with CBS performed significantly better than those taught with traditional method, there was no significant gender difference among these students.

The present study aims to investigate whether students' science learning through simulation differs by gender, using repeated experiments and larger sample size in a geoscience setting.

In other words, our third research question is:

RQ3: Is there a gender difference in the effectiveness of computer simulation on student learning?

Preliminary findings from using WILSIM-GC

Luo *et al*. (2016) reported preliminary results of using the Web-based Interactive Landform Simulation Model-Grand Canyon (WILSIM-GC, https://serc.carleton.edu/landform/) with a quasi-experimental design to assess the efficacy of WILSIM-GC as a tool for teaching landform development and evolution in comparison with a paper-based exercise. They found that both methods were effective in helping students understand the landform evolution processes as measured by the pre/posttest, but there were several advantages of the simulation approach. The improvement in scores from the pretest to posttest was large for the simulation group, but small-to-moderate for the paper-based group. In addition, for those test items requiring higher-level thinking, the percentage of students answering correctly was higher in the simulation group than in the paper-based group. Furthermore, attitudinal surveys indicated that students generally favored the interactive simulation approach. However, that study was based on data from only one semester, and involved a small sample size. The group assignment was by alphabetical order, not truly random, and some of the model interface and associated curriculum materials were still in their development stage; eg, the river longitudinal profile (hereafter, "river profile") and topographic profile across the canyon (hereafter, "cross-section profile" or simply "cross-section") were not built-in. The pre/posttest items are multiple choice items and students could potentially guess the answers correctly. Since the preliminary study, we have addressed these limitations and continued the experiment for four more semesters. The present study, then, offers the opportunity to answer our last research question:

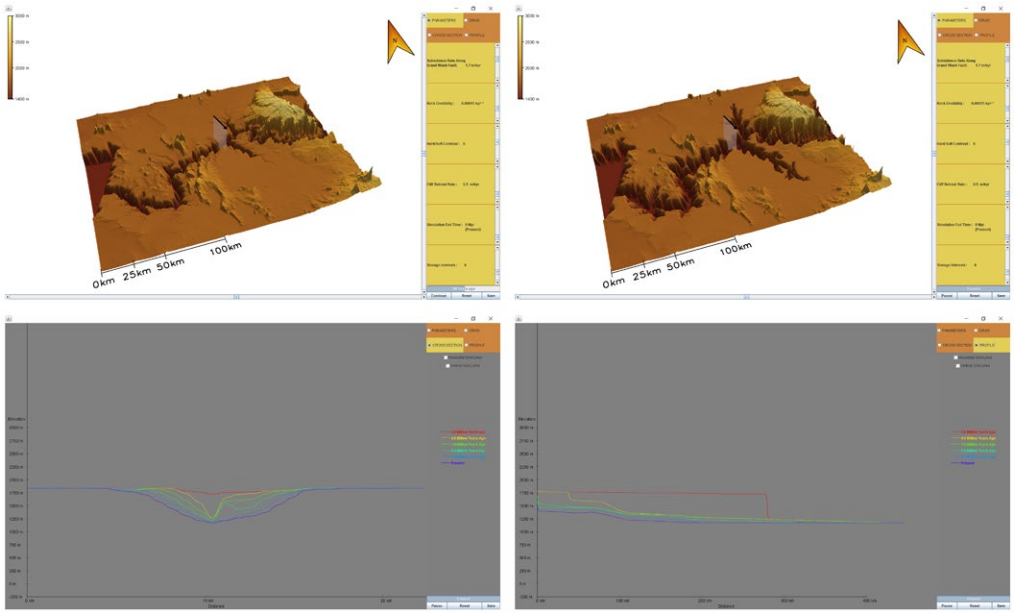

Figure 1: WILSIM-GC: (a) Perspective view (3 million years ago), (b) Perspective view (present), (c) Cross-section view, (d) Long profile view

RQ4: Are the findings from the preliminary study upheld with a larger sample collected over repeated experiments, and across multiple semesters?

Description of WILSIM-GC

WILSIM-GC is an accessible, interactive environment for students to engage in scientific inquiry and to enhance students' understanding of the processes involved in landform evolution through meaningful manipulation of parameters for different scenarios (Luo *et al.*, 2016). A screenshot of the model is shown in Figure 1. More details about the model can be found at the project website: https://serc.carleton.edu/landform/.

WILSIM-GC is a simplified version of a state-of-the-art physically-based model (Pelletier, 2010) that simulates bedrock channel erosion and cliff retreat processes responsible for the development and evolution of the Grand Canyon. We aim to leverage WILSIM-GC in teaching students not only geomorphology principles, but also the process of conducting scientific research. For example, in scientific research on the Grand Canyon, scientists are able to establish the geologically youthful, approximately 6 million year age of the canyon by comparing empirical datasets of erosion rates with theoretical, modeled landscape outputs (Darling & Whipple, 2015; Pelletier, 2010). Further, analysis of erosion rate patterns relative to numerical models with different patterns of rock strength have been a key part of interpreting the incision history of Grand Canyon. For instance, testing whether the canyon has been formed due to an increase in down-cutting rate in the past or perhaps through a constant rate of incision and distinct control from rock strength variation was accomplished by comparing field collected erosion rate and data patterns predicted by different theoretical models (Darling & Whipple, 2015). In WILSIM-GC, rock strength adjustments influence canyon form in a teaching-friendly manner that is similar to approaches employed in real scientific work.

Key Concepts and Curriculum

The key ideas we wish students to develop during the lesson are founded in extensive scientific research on erosion by rivers (eg, Whipple, DiBiase, & Crosby, 2013) and are incorporated in the curriculum material. These include down-cutting erosion, headward erosion, rock strength and erodibility, relief, base level, knickpoint and its upstream migration, and how these processes interact to shape the landform. Rivers collect rainfall and sediment and transport them downstream. This flux of materials and the accompanying expenditure of energy provide tools (the sediment) and power (moving water) to attack exposed bedrock and to remove that bedrock, creating further potential energy to be released by hillslopes and side-streams adjacent to the incising river. The point to which streams flow is called base level and base level can move in time (eg, due to down-cutting). Base level fall rate drives the changes in energy available for erosion, thus controlling vertical erosion rates. Rock strength variation, however, can alter the shape of the channel to accomplish that incision at the rate of base level fall. Channels change shape primarily in width and slope, where steeper slopes and narrower channels are associated with faster base level fall rates. Rock strength inhibits erosion patterns, such as high rock strength necessitating steeper and narrower channels for a given erosion rate.

As a simplified model designed for education, WILSIM-GC captures the key concepts described above, but does not include all the fine details, eg, erosion on the plateau is assumed negligible and the effect of plate flexure in response to sediment unloading as result of erosion is also excluded for simplicity. However, not including these fine details would not inhibit the model's ability to help students understand the key concepts and processes and how these processes interact to shape the landform we observe today, provided that instructors are aware of the limitations and discuss them as needed with students.

Methodology

Participants and experiment setting

The sample for this study consisted of 122 undergraduate students enrolled in GEOG 102: Survey of Physical Geography Laboratory, a 1-credit hour general education lab course at Northern Illinois University (Luo *et al.*, 2016), over four semesters. Students were randomly assigned into two groups that used traditional paper-based or WILSIM-GC-based learning materials (see further details about the procedure below). Tables 1 and 2 show sample sizes for each condition (group) across semesters, by both gender and science major status. Overall, there was approximately the same number of students in each group, but there were more males than females and more non-science majors than science majors.

Table 1: Sample size by group, gender and semester

Semester	Paper-based		WILSIM-GC		*Total by semester*
	Female	*Male*	*Female*	*Male*	
1. Fall 2015	2	7	8	6	23
2. Spring 2016	5	9	12	6	32
3. Fall 2016	6	14	6	9	35
4. Spring 2017	7	10	4	11	32
Total by gender	20	40	30	32	122
Total by group	60		62		

Table 2: Sample size by group, science major status and semester

Semester	Paper-based		WILSIM-GC		*Total by semester*
	Science	*Non-science*	*Science*	*Non-science*	
1. Fall 2015	3	6	6	8	23
2. Spring 2016	2	12	6	12	32
3. Fall 2016	7	13	4	11	35
4. Spring 2017	6	11	5	10	32
Total by Major	18	42	21	41	
Total by group	60		62		122

Instrumentation

Students' understanding of basic landform concepts and the processes involved in the formation of the Grand Canyon was assessed using the same instrument used in Luo *et al.* (2016), which consisted of 10 multiple-choice items (five concept-type items and five application-type items, see supplement material A). These items were revised and narrowed down from an earlier version based on feedback from testing at a workshop with eight local community college geoscience instructors (Luo *et al.*, 2016). The test–retest correlation between the pretest and posstest scores of 14 students in a lab conducted in fall 2014 reached $r = 0.7$, suggesting good reliability. The test in this study was administered at two time points before and after each treatment method (see research design below). Additionally, a 27-item attitudinal survey was administered to all students at the conclusion of the study. This survey included items that assessed students' experience, the degree of satisfaction with the user interface design of WILSIM-GC and learning activities, and their attitude toward computer simulation in comparison to the traditional learning method (Luo *et al.*, 2016).

Design and procedure

The research design followed that used in the preliminary study by Luo *et al.* (2016), and is illustrated in Figure 2. The experiment was conducted in one lab session that lasted 1 hour 50 minutes. Students in each lab section were randomly assigned into one of two groups. Prior to the labs, both groups were required to read some background material about the Grand Canyon posted on the course online management system (BlackBoard) and to complete a pretest assignment. During the lab, the two groups used different curricular materials to learn about the processes involved in forming the Grand Canyon: the treatment group (Group A) used WILSIM-GC

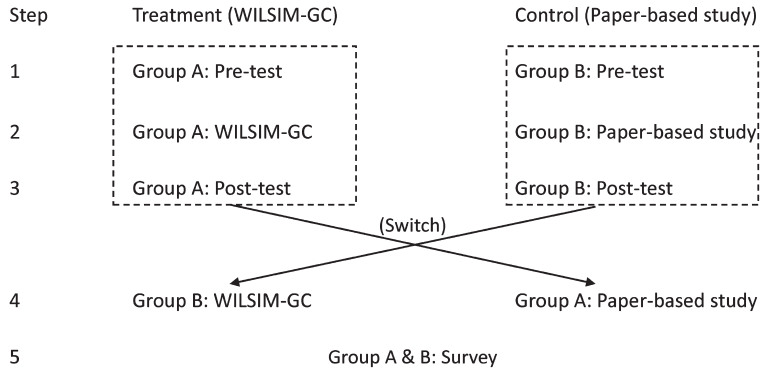

Figure 2: The procedure of the experiment

and the control group (Group B) used traditional paper-based written material (the details of each set of teaching materials will be described next). Both groups then completed a posttest immediately after their respective learning activities to assess the effect of using the two different instructional methods (Figure 2). Then, the two groups switched instructional methods so that, at the study's conclusion, both groups had received the same experience. An attitudinal survey was administered at the conclusion, after both groups had experienced both teaching methods. The labs were run by teaching assistants (TAs), who received the same training on how to use WILSIM-GC. To further minimize the influence of different TAs in different labs, we asked the TAs to provide only technical and procedural help.

Compared to Luo *et al.* (2016), the research design has the following differences/improvements: (1) WILSIM-GC interface and lab material were improved based on previous experience and student feedback; (2) students were randomly assigned to two groups, resulting in a true experimental design; (3) the experiment was repeated for four consecutive semesters, resulting in a larger student sample size; (4) for each pre- and posttest item, we also asked students to indicate whether they had guessed the answer to minimize the uncertainty that they may correctly guess the multiple choice item. All guessed answers were counted as wrong answers.

Data analysis

To address whether the two groups differed in their pretest to posttest score growth (RQ1), as well as to assess the potential effects of background knowledge (implied in science major status) (RQ2), gender (RQ3), and semester (RQ4), 2 × 2 × 2 × 4 repeated-measures ANOVAs were carried out using the total knowledge test score as the repeatedly-measured outcome, and either (1) group membership, gender, and semester as between-subjects factors, or (2) group membership, student major (science vs. non-science), and semester as between-subjects factors. The first repeated-measures ANOVA was then repeated using concept-type knowledge (items #1-#5 on the knowledge test) and application-type knowledge (items #6-#10) as separate repeatedly measured outcomes. Finally, distributional patterns from the attitudinal survey were descriptively examined.

Table 3: Repeated-measures ANOVA results for test scores across time by treatment group and gender

Source	*SS*	*df*	*MS*	*F*	*p*
Time	22516.497	1	22516.497	101.187	<0.001***
Time × group	62.820	1	62.820	0.282	0.596
Time × gender	10.591	1	10.591	0.048	0.828
Time × semester	1808.825	3	602.942	2.710	0.049*
Time × group × gender	1107.632	1	1107.632	4.978	0.028**
Time × group × semester	1709.233	3	569.744	2.560	0.059
Time × gender × semester	1348.415	3	449.472	2.020	0.116
Time × group × gender × semester	1770.884	3	590.295	2.653	0.052
Error	23587.614	106	222.525		

Note. *$p < .05$, **$p < .01$, ***$p < .001$.

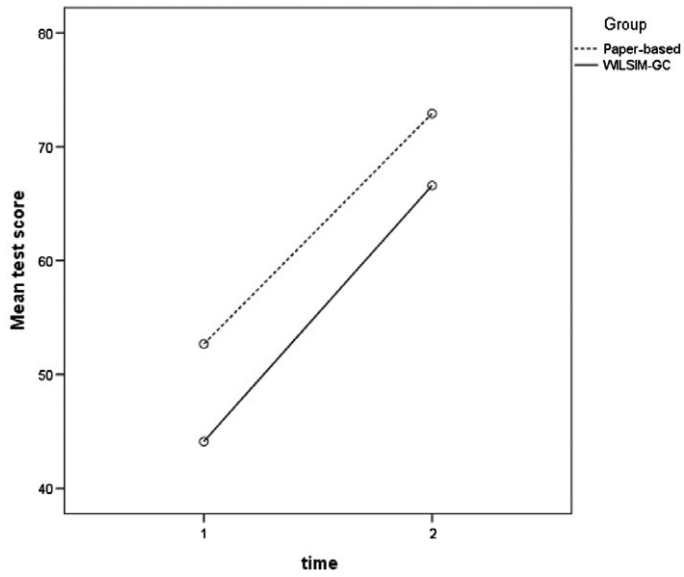

Figure 3: Mean test score values over time by treatment group. Time 1 = pretest, time 2 = posttest

Results

Overall

Results from the repeated-measures ANOVA to assess group differences in achievement test score growth are shown in Table 3. Model residuals showed some departure from normality (see Figure S1); however, analyses were robust to this non-normality due to the sample size. Box's test showed equality of covariance matrices across groups ($p = .836$). The ANOVA results indicated that the combined groups showed statistically significant growth in scores from pretest to posttest [$F(1, 106) = 101.19$, $p < .001$], with a large effect size ($\eta^2 = .42$). No statistically significant time × group interaction effect was observed [$F(1, 106) = 0.282$, $p = .596$] indicating that, overall (across the four semesters), there was no difference between the two groups in test score growth from pretest to posttest. Figure 3 provides a plot of the mean test score values across time of test administration by group.

Differences by gender and semester

Examination of other model effects, however, indicated a significant three-way time × group × gender interaction effect [$F(1, 106) = 4.928$, $p = .028$], with a small effect size ($\eta^2 = .02$). Specifically, the effect of the treatment on test score growth was distinct for females versus males (see Figure 4). Females showed greater growth from pretest to posttest than males with WILSIM-GC than with the paper-based intervention, while males showed greater growth using the paper-based intervention than with WILSIM-GC. A marginally significant [$F(3, 106) = 2.56$, $p = .059$] and small ($\eta^2 = .03$) time × group × semester effect was evident. In semesters 1 and 3, the WILSIM-GC group showed greater growth than the paper-based group, while in semester 2, the paper-based group showed greater growth (no group difference was evident in semester 4). Figure 5 illustrates this effect. Complete descriptive statistics for the pretest and posttest scores of the two treatment groups (paper-based vs. WILSIM-GC) by gender and by semester can be found in the supplemental material Table S1.

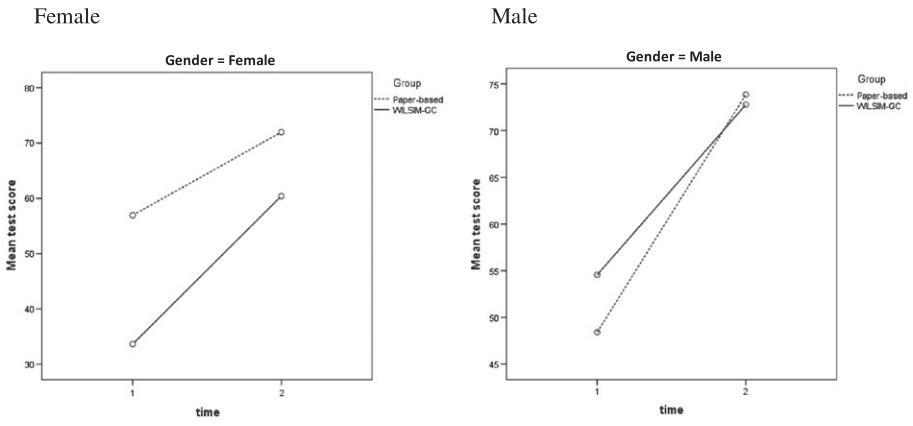

Figure 4: Mean test score values across time by treatment group and gender

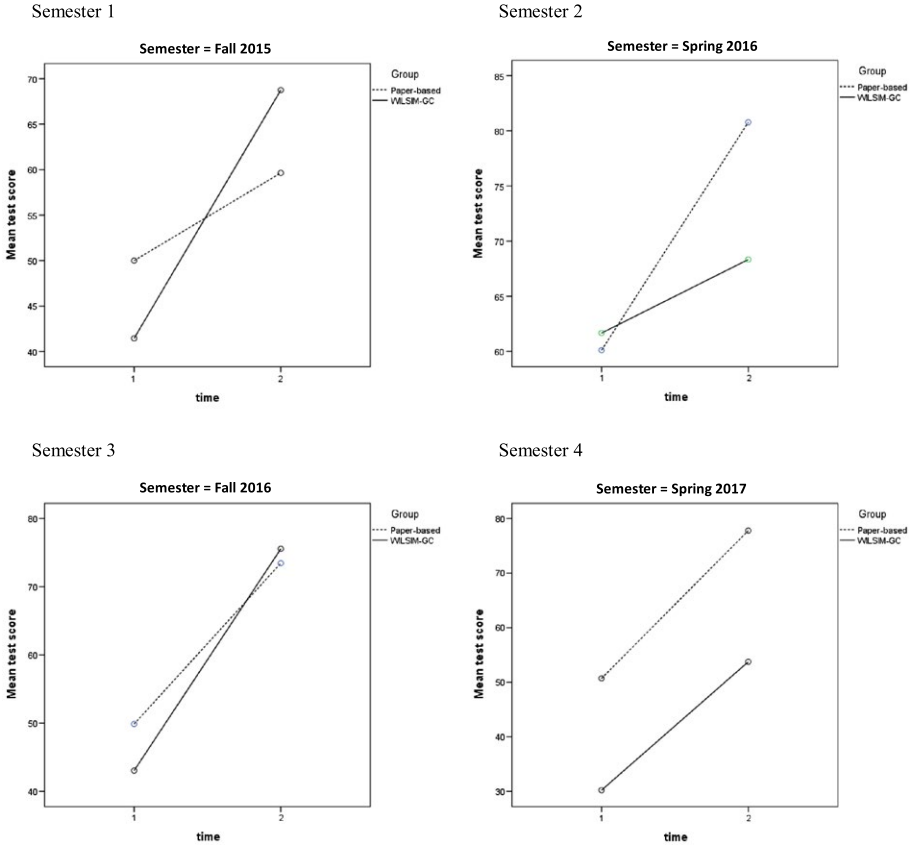

Figure 5: Mean test score values across time by treatment group and semester

Difference by science major status

To address whether students' background knowledge (implied in science major status) played any role in their performance in terms of pretest to posttest score growth, a 2 × 2 × 2 × 4

Table 4: Repeated-measures ANOVA results for test scores across time by treatment group and science major status

Source	*SS*	*df*	*MS*	*F*	*p*
Time	27197.587	1	27197.587	120.653	<0.001***
Time × group	272.757	1	272.757	1.210	0.274
Time × science major	958.530	1	958.530	4.252	0.042*
Time × semester	643.682	3	214.561	0.952	0.418
Time × group × science major	733.973	1	733.973	3.256	0.074
Time × group × semester	836.135	3	278.712	1.236	0.300
Time × science major × semester	2307.908	3	769.303	3.413	0.020*
Time × group × science major × semester	728.706	3	242.902	1.078	0.362
Error	23894.442	106	225.419		

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

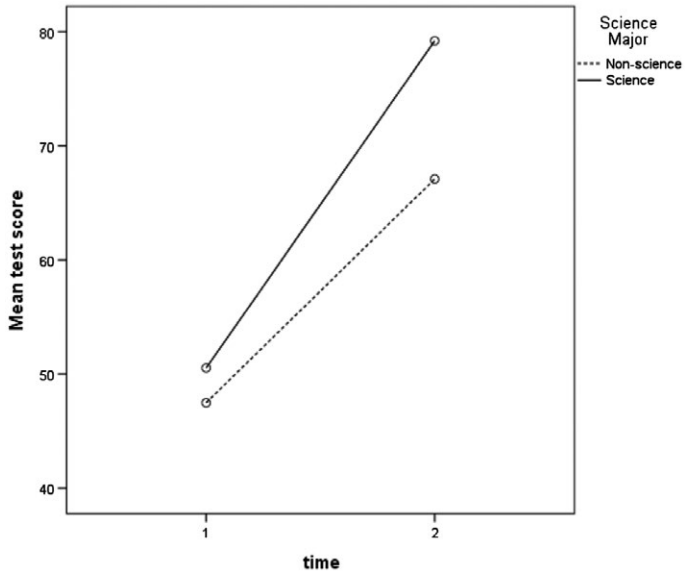

Figure 6: Mean test score values over time by science major status

repeated-measures ANOVA was carried out with group membership, science major status, and semester as the three between-subjects factors, and time of test administration (pretest vs. posttest) as the within-subjects factor. Model residuals showed close-to-normal distributions (see Figure S2). Box's test showed equality of covariance matrices across groups ($p = .383$). Results from the ANOVA (Table 4) indicated that the combined groups showed statistically significant growth in scores from pretest to posttest [$F(1, 106) = 120.65$, $p < .001$], with a large

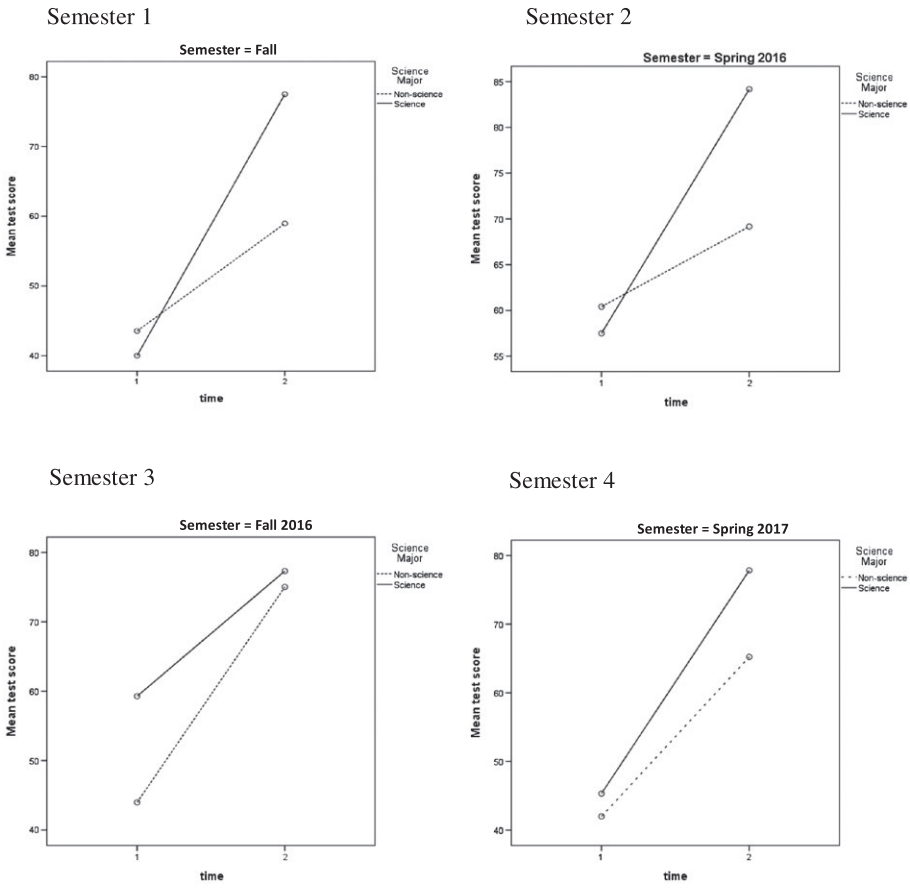

Figure 7: Mean test score values across time by science major status and semester

effect size (η^2 = .47). No statistically significant time × group interaction effect was observed [$F(1, 106) = 1.21$, $p = .274$] indicating that, overall (across the four semesters), there was no difference between the two treatment groups in test score growth from pretest to posttest. Examination of other model effects, however, indicated a significant two-way time × science major status interaction effect [$F(1, 106) = 4.25$, $p = .042$], but with a small effect size (η^2 = .02). Specifically, the growth in test scores over time was greater for science majors than for non-science majors (see Figure 6). Also, a statistically significant [$F(3, 106) = 3.41$, $p = .020$] and small-to-moderate ($\eta^2 = .04$) time × science major × semester effect was evident. In semesters 1, 2 and 4, students who were science majors showed greater growth than their non-science major peers, while in semester 3, non-science majors showed greater growth than science majors. Figure 7 illustrates this effect by displaying growth by science major status and semester. Complete descriptive statistics for the pretest and posttest scores of the two treatment groups (paper-based vs. WILSIM-GC) by science major status and by semester can be found in the supplemental material Table S2.

Difference in performance in concept versus application test items

To test whether the two instructional methods have different effects on students' performance in concept-type test items (#1-#5, Supplement A) versus application-type test items (#6-#10,

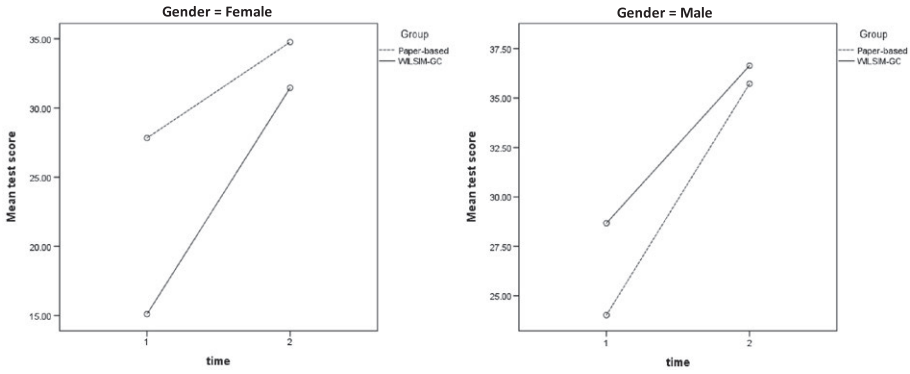

Figure 8: Mean "Application" test score values across time by treatment group and gender

Supplement A), we ran repeated-measures ANOVA using the "concept" and "application" outcomes, and using group, gender, and semester as the between-subjects factors. No significant group difference in "concept" growth occurred. For the "application" outcome, although the WILSIM-GC group showed greater growth than the paper-based group at the sample level, this difference in growth was not statistically significant. A significant three-way time × group × gender interaction effect was evident for "application" questions [$F(1, 106) = 6.86$, $p = .01$], with a small effect size ($\eta^2 = .03$). Here, female students showed greater growth with WILSIM-GC, while males showed greater growth with paper-based approach (Figure 8).

Attitudinal survey results

In addition to comparing test scores of the treatment and control groups, we also administered an attitudinal survey after both groups of students had experienced both methods of teaching (see Figure 2). The survey was based on 6-point Likert items (1 = *Strongly Disagree* to 6 = *Strongly Agree*) and the results are summarized in Table 5. It is clear that, throughout the four semesters, students consistently agreed with the positively worded statements and disagreed with the negatively worded statements. In particular, statements #23-#25 ask students to compare the two instructional methods, and they consistently favored the WILSIM-GC over the paper-based method. To get a sense of the overall result, we calculated a composite score for each student (computed as the mean of the item scores, with item scores for negatively worded statements reverse-coded). The mean composite scores for each semester were greater than or equal to 4 (with 3 of the semesters showing mean values greater than or equal to 4.5), suggesting that students' responses were consistent with our expectation that WILSIM-GC has a positive effect on students' learning.

Interpretation and Discussion

Overall results

The overall results based on four semesters of data are generally consistent with our preliminary findings (Luo *et al*., 2016). These include: (1) for both WILSIM-GC and paper-based groups, the posttest scores were significantly higher than pretest scores; (2) the score growth from pretest to posttest was not significantly different between the two groups; (3) based on attitudinal survey, students favored the WILSIM-GC approach. However, with a larger sample size and replication across four semesters, the results from this study make the findings more robust. In addition, by utilizing a larger sample size and multiple semesters of data, we were able to examine the effects of science major status and gender on score growth across semesters.

Table 5: Attitudinal Survey Results

		Fall 2015		Spring 2016		Fall 2016		Spring 2017	
ID	*Statement*	*Mode*	*(M, SD)*	*Mode*	*(M, SD)*	*Mode*	*(M, SD)*	*Mode*	*(M, SD)*
1	WILSIM-GC helped me understand how landscapes change over millions of years.	++	(5.3, 0.6)	+++	(5.3, 0.9)	++	(4.1, 1.7)	++	(5.1, 0.7)
2	WILSIM-GC makes me feel I can solve the problem based on the information given.	++	(5.0, 0.6)	+++	(5.1, 0.8)	++	(3.9, 1.8)	++	(4.9, 0.7)
3	WILSIM-GC helps me have a clear understanding how I arrive to my final outcomes.	++	(5.1, 0.5)	+++	(4.8, 1.1)	+	(4.0, 1.6)	++	(4.9, 0.7)
4	WILSIM-GC provides a better way to analyze landform evolution.	++	(5.2, 0.7)	+++	(5.4, 0.8)	++	(4.4, 1.7)	+++	(4.9, 1.0)
5	WILSIM-GC encourages me to identify the critical features of landform evolution.	++	(5.0, 0.8)	++	(4.9, 1.0)	++	(3.9, 1.6)	++	(4.9, 0.8)
6	WILSIM-GC helps me apply my understanding of the landform evolution.	++	(5.2, 0.6)	++	(5.1, 0.8)	++	(4.1, 1.6)	++	(4.9, 0.7)
7	Using WILSIM-GC was engaging and interesting.	+++	(5.3, 0.8)	+++	(5.0, 0.9)	++	(4.1, 1.7)	++	(4.9, 1.0)
8	Using WILSIM-GC helped me to think about "how the Grand Canyon formed."	++	(5.0, 0.6)	+++	(5.3, 0.8)	+++	(4.4, 1.5)	+++	(5.2, 0.8)
9^	WILSIM-GC is not compatible with my learning approach.	--	(2.5, 1.3)	--	(3.1, 1.6)	-	(3.3, 1.6)	-	(3.2, 1.5)
10	Watching the landform evolution in WILSIM-GC helped me "see" how the Grand Canyon formed.	+++	(5.5, 0.5)	+++	(5.4, 0.7)	+++	(4.6, 1.5)	+++	(5.2, 0.8)
11	It was easy to navigate among the various features of WILSIM-GC.	++	(5.2, 0.6)	+++	(5.2, 0.8)	++	(4.2, 1.6)	+	(4.4, 1.2)
12	It was easy for me to visualize and compare simulated results to the actual Grand Canyon when using WILSIM-GC.	++	(5.0, 0.9)	+++	(5.2, 0.8)	++	(4.3, 1.5)	++	(5.0, 0.9)
13	It was difficult to use WILSIM-GC.	--	(2.0, 0.9)	---	(2.3, 1.5)	---	(2.7, 1.5)	--	(3.0, 1.4)
14	The WILSIM-GC exercise was about the right length.	++	(4.8, 0.9)	++	(4.8, 1.0)	++	(3.9, 1.6)	++	(4.5, 1.1)
15	I am confident that I understand how to use WILSIM-GC.	++	(4.9, 0.8)	++	(4.9, 1.1)	++	(3.9, 1.6)	+	(4.5, 0.8)
16	I put enough effort into learning WILSIM-GC.	++	(5.1, 0.6)	++	(5.1, 0.7)	++	(4.7, 1.1)	++	(5.0, 0.7)
17^	I feel WILSIM-GC provides inadequate guidelines to help solving problems.	--	(2.8, 1.6)	--	(2.9, 1.5)	--	(3.3, 1.5)	--	(3.4, 1.3)

18^	I feel WILSIM-GC provides inadequate functions to facilitate discussions.	--	(2.7, 1.6)	--	(2.7, 1.5)	--	(3.0, 1.3)	--	(3.3, 1.4)
19	I want more training on WILSIM-GC.	-	(3.7, 1.4)	+++	(3.9, 1.6)	+	(3.6, 1.4)	+	(3.9, 0.9)
20	I would like to continue to use WILSIM-GC.	+	(4.4, 1.0)	+++	(4.7, 1.3)	+	(3.7, 1.4)	+	(3.9, 1.1)
21	I would encourage others to use WILSIM-GC.	++	(4.8, 0.8)	++	(5.0, 0.9)	+	(3.7, 1.5)	+	(4.4, 1.1)
22	Compared to using paper-based self-study material, WILSIM-GC offers me better management of my thinking process toward the inquiry activities.	++	(5.0, 0.8)	+++	(5.0, 1.1)	+	(3.6, 1.6)	+	(4.6, 0.9)
23	Compared to using paper-based self-study material, WILSIM-GC is more time efficient for learning activity.	+++	(4.6, 1.3)	+++	(5.2, 0.9)	++	(3.4, 1.7)	+	(4.4, 0.9)
24	Compared to using paper-based self-study material, WILSIM-GC is more convenient to use.	++	(5.0, 0.9)	+++	(5.1, 1.0)	+	(3.4, 1.6)	++	(4.3, 1.3)
25	Compared to using paper-based self-study material, WILSIM-GC is more fun to use.	+++	(5.2, 1.2)	+++	(5.4, 0.7)	++	(4.1, 1.5)	+	(4.5, 1.0)
26	Another session learning about and using WILSIM-GC would help me better understand erosional processes.	++	(4.6, 1.4)	++	(5.1, 0.8)	+	(3.9, 1.4)	++	(4.9, 0.7)
27^	I had computer or technological issues while trying to use WILSIM-GC.	--	(2.3, 1.5)	---	(2.5, 1.8)	---	(2.8, 1.7)	+	(3.1, 1.4)
	Composite score		(4.9, 0.9)		(4.9, 1.1)		(4.0, 1.5)		(4.5, 1.0)

(+/- signs show the modal or most frequent response: Strongly Agree = +++ (6), Agree = ++ (5), Slightly Agree = + (4), Slightly Disagree = - (3), Disagree = -- (2), Strongly Disagree = --- (1); *M* = mean; *SD* = standard deviation; ^Negatively worded statement)

Major status, background knowledge and scaffolding

Across all four semesters, science major students outperformed non-science major students (Figure 6). In three out of four semesters, pretest to posttest score growth for science major students (STEM fields) was significantly higher than non-science major students (Figure 7 and Table 2). This indicates that the background knowledge or interests that science major students possess (and the non-science major students may lack) may have played a key role in realizing the potential of computer simulation. This finding is consistent with findings of previous studies and suggests that for computer simulations to improve students' learning, proper scaffolding to prepare students with the needed background is necessary (Adams *et al.*, 2008; Bell & Trundle, 2008; eg, Khan, 2011; Schneps *et al.*, 2014). A recent study confirmed that the high cognitive load demands of computer models placed on novice learners such as high school students without enough background knowledge could hinder their understanding of the intended scientific content represented by computer models (Waight & Gillmeister, 2014). The study suggested that the cognitive load reduction should come from the development and scaffolding of adequate background knowledge related to models, modeling, and scientific content (Waight & Gillmeister, 2014). Another recent study that compared the performance and behavior of students learning basic principles of electricity using Augmented Reality Simulation Systems with and without scaffolding support also confirmed that students with scaffolding support showed greater learning achievement than those without such support (Ibanez, Di-Serio, Villaran-Molina, & Delgado-Kloos, 2016).

Gender differences and implications

As a whole, female students responded better than male students to the WILSIM-GC treatment in terms of pretest to posttest score growth, whereas male students responded better to paper-based intervention (Figure 4). In particular, female students' score growth for application-type questions (requiring higher-level thinking) in the WILSIM-GC group was significantly greater than male students (Figure 8). This finding is in contrast to the findings of Kickmeier-Rust, Holzinger, Wassertheurer, Hessinger and Albert (2007), Koh *et al.* (2010) and Mihindo *et al.* (2017), which showed either better male performance or no gender difference. Our finding also appears to be contrary to the stereotypical view that males may be better at computer technology and females may be better at reading and comprehension (Arellano, 2013; Lynn & Mikk, 2009; Shashaani, 1997; Wladis, Conway, & Hachey, 2015; Yau & Cheng, 2012). The significantly higher score growths for female students in WILSIM-GC group and for test items requiring higher-level thinking suggest that the visually-oriented, easy-to-use interactive computer simulation approach has helped female students overcome the perceived gender barrier in technology. Alternatively, this may indicate that the better reading and comprehension skills of female students helped them realize the full potential of the interactive simulation model in enhancing their learning. This interpretation is supported by some recent studies in the literature. For example, Yilmaz, Baydas, Karakus and Goktas (2015) aimed to understand how a learning environment providing rich interactions with technology affects the level of cognitive engagement due to gender difference. Specifically, their research focused on improving what many assume to be female students' lack of confidence and motivation when faced with technology challenges. The study concludes that it is important to provide prompts and visual stimuli in order to increase female students' willingness and ability to identify and deal with technical challenges and problem solving. Another study by Wassenburg, de Koning, de Vries, Boonstra and van der Schoot (2017) provides a similar confirmation of this, suggesting that, when interacting with a computer simulation with either visual or textual representations, female students construct more coherent and vivid mental simulations than male students and rely more heavily on visual and graphical

representations. These results suggest the importance of providing meaningful and visual rich elements and instructional scaffolds in the design of technology integrated learning tools.

Explanation of unexpected results

In two out of the four semesters (semesters 1 and 3), the WILSIM-GC group outperformed the paper-based group in test score growth. The opposite was true for semester 2, while there was no significant group difference for semester 4. The unexpected result for semester 2 may be explained by an examination of the attitudinal survey results. The majority of students in semester 2 "strongly agree" with statement #19 ("I want more training on WILSIM-GC."), whereas the majority of students in the other three semesters only "slightly agree" or "slightly disagreed" with this statement. This suggests that students in semester 2 felt they were not as well prepared and needed more guidance than students in other semesters. The lack of difference in score growth between the two groups for semester 4 may be explained by their responses to statement #27, where most students in semester 4 "slightly agreed" that they encountered some technical issues while using WILSIM-GC, whereas most students in the other three semesters responded "strongly disagree" or "disagree" to this statement.

Limitations and future work

We have addressed most of the limitations inherent in the preliminary study (Luo *et al.*, 2016). For example, the sample size was large and the student groups were randomly assigned. In addition, asking students to indicate if they had guessed the answer for each test item minimized the doubt that they may have guessed the answer correctly, and thus provided an additional level of confidence in the results. However, at least one limitation still remains. The intervention time in this study remained short: approximately 45 minutes for students to complete the intervention, the posttest, and the other lab activity during a single lab period of one hour and 50 minutes. Despite the advantages of WILSIM-GC, the limited exposure time may be challenging for students to fully grasp the concepts and processes behind the model, particularly among the non-science majors who lack the basic background in terms of scientific vocabulary and methodology. This is, in part, dictated by the nature of this lab course that has many topics to cover of which landscape evolution is only one. Using WILSIM-GC in an upper-division geomorphology course over a full semester would give students multiple opportunities to become more familiar with the model, to better understand the meaning of its different parameters, and thus to realize the full potential of dynamic computer simulation. This is corroborated by the survey result for statement #26, which indicated that students agreed another session would help them better understand the erosional processes forming Grand Canyon.

Conclusion

With a true randomized experimental design replicated over four semesters, this study compared students' performance in understanding landform evolution processes of Grand Canyon as measured by the pretest to posttest score growth between two treatment methods: an online simulation tool WILSIM-GC and a paper-based approach. Results show that both methods were effective at teaching students the landform evolution concepts and processes. While there was no statistically significant difference in score growth between the two instructional methods, the attitudinal survey showed that students consistently favored the simulation approach over the paper-based approach over the four semesters. The findings are consistent with the preliminary results from Luo *et al.* (2016). However, the larger sample size and repeated true randomized experiments over four semesters make the findings more robust.

In addition, students' major status also played a key role in the effect of the two different treatment approaches. Science major students generally performed better than non-science major students in terms of pretest to posttest score growth. This suggests that the different background knowledge of science vs. non-science majors played an important role in realizing the potential of using computer simulation in enhancing students' learning. The implication is that adequate scaffolding is needed to provide students with sufficient background knowledge so that they can take the full advantage of interactive simulation tool. Both instructional approaches should be integrated together to maximize the effect of simulation tools in enhancing students' learning, especially for online or hybrid courses and flipped classrooms.

We also found a statistically significant gender difference in the effect of WILSIM-GC. Females showed greater growth from pretest to posttest than males with WILSIM-GC than with the paper-based intervention, while males showed greater growth using the paper-based intervention than with WILSIM-GC. With the WILSIM-GC instructional method, female students also performed significantly better than male students in answering application-type test items, which require higher level thinking. While the reason behind the findings about gender differences may require further study, we speculate that the visually oriented, easy-to-use interactive computer simulation approach has helped female students overcome the perceived gender barrier in technology and/or that female students' better reading and comprehension skills helped them realize the full potential of the interactive simulation in enhancing learning.

Acknowledgements

This research is funded through NSF-TUES program (award DUE-1140375) and further information can be found at https://serc.carleton.edu/landform/index.html. We thank the following teaching assistants for their help throughout the past four semesters: Hannah Eboh, Jacob Strohm, Lily Cobo, Xuezhi Cang. We are grateful for the valuable comments and suggestions from two anonymous reviewers and the editors, which helped to improve the quality of the paper.

Statement on ethics and conflicts of interest

This research was conducted with Internal Review Board's (IRB) human subject research approval ("exempt" status). All participants' identities were removed from analysis and reporting. The authors have no conflicts of interest in this research.

References

Adams, W. K., Paulson, A., Wieman, C. E., Henderson, C., Sabella, M., & Hsu, L. (2008). What levels of guidance promote engaged exploration with interactive simulations? (pp. 59–62). AIP. https://doi.org/10.1063/1.3021273

Arellano, M. D. C. (2013). Gender differences in reading comprehension achievement in English as a foreign language in compulsory secondary education. *Tejuelo*, *17*(1), 67–84.

Bell, R. L., & Trundle, K. C. (2008). The use of a computer simulation to promote scientific conceptions of moon phases. *Journal of Research in Science Teaching*, *45*(3), 346–372. https://doi.org/10.1002/tea.20227

Cai, Z., Fan, X., & Du, J. (2017). Gender and attitudes toward technology use: a meta-analysis. *Computers & Education*, *105*, 1–13. https://doi.org/10.1016/j.compedu.2016.11.003

Darling, A., & Whipple, K. (2015). Geomorphic constraints on the age of the western Grand Canyon. *Geosphere*, *11*(4), 958–976. https://doi.org/10.1130/GES01131.1

Day, T. (2012). Undergraduate teaching and learning in physical geography. *Progress in Physical Geography*, *36*(3), 305–332. https://doi.org/10.1177/0309133312442521

de Jong, T. (2006). Computer simulations: technological advances in inquiry learning. *Science*, *312*(5773), 532–533. https://doi.org/10.1126/science.1127750

Edsall, R., & Wentz, E. (2007). Comparing strategies for presenting concepts in introductory undergraduate geography: Physical models vs. computer visualization. *Journal of Geography in Higher Education*, *31*(3), 427–444. https://doi.org/10.1080/03098260701513993

Gordin, D. N., & Pea, R. D. (1995). Prospects for scientific visualization as an educational technology. *Journal of the Learning Sciences*, *4*(3), 249–279.

Ibanez, M.-B., Di-Serio, A., Villaran-Molina, D., & Delgado-Kloos, C. (2016). Support for augmented reality simulation systems: The effects of scaffolding on learning outcomes and behavior patterns. *IEEE Transactions on Learning Technologies*, *9*(1), 46–56. https://doi.org/10.1109/TLT.2015.2445761

Khan, S. (2011). New pedagogies on teaching science with computer simulations. *Journal of Science Education and Technology*, *20*(3), 215–232. https://doi.org/10.1007/s10956-010-9247-2

Kickmeier-Rust, M. D., Holzinger, A., Wassertheurer, S., Hessinger, M., & Albert, D. (2007). Text-based learning versus learning with computer simulations: Does gender matter? *Gender in E-Learning and Educational Games: A Reader* (pp. 39–52). Innsbruck: Studien Verlag.

Koh, C., Tan, H. S., Tan, K. C., Fang, L., Fong, F. M., Kan, D., … Wee, M. L. (2010). Investigating the effect of 3D simulation based learning on the motivation and performance of engineering students. *Journal of Engineering Education*, *99*(3), 237–251. https://doi.org/10.1002/j.2168-9830.2010.tb01059.x

Luo, W., Pelletier, J., Duffin, K., Ormand, C., Hung, W., Shernoff, D. J., … Furnes, W. (2016). Advantages of computer simulation in enhancing students' learning about landform evolution: A case study using the Grand Canyon. *Journal of Geoscience Education*, *64*(1), 60–73. https://doi.org/10.5408/15-080.1

Lynn, R., & Mikk, J. (2009). Sex differences in reading achievement. *Trames. Journal of the Humanities and Social Sciences*, *13*(1), 3. https://doi.org/10.3176/tr.2009.1.01

Mihindo, W. J., Wachanga, S. W., & Anditi, Z. O. (2017). Effects of computer-based simulations teaching approach on students' achievement in the learning of chemistry among secondary school students in Nakuru Sub County. *Kenya. Journal of Education and Practice*, *8*(5), 65–75.

Pelletier, J. D. (2010). Numerical modeling of the late Cenozoic geomorphic evolution of Grand Canyon. *Arizona. Geological Society of America Bulletin*, *122*(3–4), 595–608. https://doi.org/10.1130/B26403.1

Perkins, K., Moore, E., Podolefsky, N., Lancaster, K., Denison, C., Rebello, N. S., … Singh, C. (2012). *Towards research-based strategies for using PhET simulations in middle school physical science classes*, 295–298. https://doi.org/10.1063/1.3680053

Podolefsky, N. S., Adams, W. K., Lancaster, K., Perkins, K. K., Singh, C., Sabella, M., & Rebello, S. (2010). *Characterizing complexity of computer simulations and implications for student learning*, 257–260. https://doi.org/10.1063/1.3515215

Podolefsky, N. S., Moore, E. B., & Perkins, K. K. (2013). Implicit scaffolding in interactive simulations: design strategies to support multiple educational goals. ArXiv Preprint ArXiv:1306.6544. Retrieved from https://arxiv.org/abs/1306.6544

Rhema, A., & Miliszewska, I. (2014). Analysis of student attitudes towards e-learning: the case of engineering students in Libya. *Issues in Informing Science and Information Technology*, *11*, 169–190.

Sáinz, M., Meneses, J., López, B.-S., & Fàbregues, S. (2016). Gender stereotypes and attitudes towards information and communication technology professionals in a sample of Spanish secondary students. *Sex Roles*, *74*(3–4), 154–168. https://doi.org/10.1007/s11199-014-0424-2

Scalise, K., Timms, M., Moorjani, A., Clark, L., Holtermann, K., & Irvin, P. S. (2011). Student learning in science simulations: Design features that promote learning gains. *Journal of Research in Science Teaching*, *48*(9), 1050–1078. https://doi.org/10.1002/tea.20437

Schneps, M. H., Ruel, J., Sonnert, G., Dussault, M., Griffin, M., & Sadler, P. M. (2014). Conceptualizing astronomical scale: Virtual simulations on handheld tablet computers reverse misconceptions. *Computers & Education*, *70*, 269–280. https://doi.org/10.1016/j.compedu.2013.09.001

Shashaani, L. (1997). Gender differences in computer attitudes and use among college students. *Journal of Educational Cmputing Research*, *16*(1), 37–51.

Smetana, L. K., & Bell, R. L. (2012). Computer simulations to support science instruction and learning: A critical review of the literature. *International Journal of Science Education*, *34*(9), 1337–1370. https://doi.org/10.1080/09500693.2011.605182

Stumpf, R. J., Douglass, J., & Dorn, R. I. (2008). Learning desert geomorphology virtually versus in the field. *Journal of Geography in Higher Education*, *32*(3), 387–399. https://doi.org/10.1080/03098260802221140

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies*, *57*(4), 247–262. https://doi.org/10.1006/ijhc.2002.1017

Waight, N., & Gillmeister, K. (2014). Teachers and students' conceptions of computer-based models in the context of high school chemistry: elicitations at the pre-intervention stage. *Research in Science Education*, *44*(2), 335–361. https://doi.org/10.1007/s11165-013-9385-7

Wassenburg, S. I., de Koning, B. B., de Vries, M. H., Boonstra, A. M., & van der Schoot, M. (2017). Gender differences in mental simulation during sentence and word processing: gender differences in mental simulation. *Journal of Research in Reading*, *40*(3), 274–296. https://doi.org/10.1111/1467-9817.12066

Whipple, K. X., DiBiase, R. A., & Crosby, B. T. (2013). Bedrock Rivers. *In Treatise on Geomorphology* (pp. 550–573). Elsevier. https://doi.org/10.1016/B978-0-12-374739-6.00254-2

Whitley, B. E. (1997). Gender differences in computer-related attitudes and behavior: A meta-analysis. *Computers in Human Behavior*, *13*(1), 1–22. https://doi.org/10.1016/S0747-5632(96)00026-X

Wladis, C., Conway, K. M., & Hachey, A. C. (2015). The online STEM classroom—who succeeds? An exploration of the impact of ethnicity, gender, and non-traditional student characteristics in the community college context. *Community College Review*, *43*(2), 142–164. https://doi.org/10.1177/0091552115571729

Yau, H. K., & Cheng, A. L. F. (2012). Gender difference of confidence in using technology for learning. *The Journal of Technology Studies*, *38*(2). https://doi.org/10.21061/jots.v38i2.a.2

Yilmaz, R. M., Baydas, O., Karakus, T., & Goktas, Y. (2015). An examination of interactions in a three-dimensional virtual world. *Computers & Education*, *88*, 256–267. https://doi.org/10.1016/j.compedu.2015.06.002

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site.