Title: Preparing Students for Future Learning with Teachable Agents

Author(s):
Doris B. Chin\textsuperscript{1}, Ilsa M. Dohmen\textsuperscript{1}, Britte H. Cheng\textsuperscript{3}, Marily A. Oppezzo\textsuperscript{2}, Catherine C. Chase\textsuperscript{2}, and Daniel L. Schwartz\textsuperscript{2}

\textsuperscript{1} Stanford Center for Innovations in Learning, Stanford University, 450 Serra Mall, Building 160, Stanford, CA, 94305-2055
\textsuperscript{2} School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA, 94305-3096
\textsuperscript{3} Center for Technology in Learning, SRI International, 333 Ravenswood Avenue, Menlo Park, CA, 94025-3493
Background/context:

Over the past several years, we have been developing an instructional technology, called Teachable Agents (TA), which draws on the social metaphor of teaching to help students learn. Students teach a computer character, their “agent,” by creating a concept map of nodes connected by qualitative causal links (e.g., A increases B) (insert Figure 1 here). Figure 1 shows the main interface for the TA software, in which the concept map symbolizes the interior of the agent’s brain. The student has taught her agent, “Dee,” about global warming, including the individual propositions that methane is a type of greenhouse gas, greenhouse gases are a type of insulation, and insulation prevents heat radiation. Artificial intelligence techniques for qualitative reasoning (Biswas, Leelawong, Schwartz, Vye, & TAG-V, 2005) enable the agent to use the concept map to answer questions, thereby providing a model of thinking with high interactivity. In Figure 1, Dee uses her map to answer the query, “What happens to ‘heat radiation’ if ‘methane’ increases?” She highlights successive links in her concept map to show the entire chain of inference. To complement the graphical reasoning, Dee also unfolds her reasoning in text (lower panel). An agent always reasons “logically,” even if the propositions it has been taught are incorrect. Students can trace their agent’s reasoning, and then reorganize or amend its knowledge (and their own) if necessary. To further support students in their learning and “teaching,” a variety of affiliated TA technologies have also been developed that provide students with extensive feedback on their agent’s progress (insert Figure 2 here).

We hypothesize that the metaphor of teaching allows students to use a well-known, and potentially productive, schema for organizing their interactions and for interpreting feedback (for reviews of the benefits of learning-by-teaching, see Annis, 1983; Biswas et al., 2005; Renkl, 1995; Roscoe & Chi, 2008). The teaching metaphor can also enlist fruitful social attitudes including a sense of responsibility towards one’s pupil. In a pair of recent studies with TA, students in one condition were told the character was an agent they were teaching, and in the other condition, students thought the character represented themselves. Students who thought they were teaching the agent exhibited more affect, were more attentive to feedback, and spent more time revising their errors by engaging in learning-relevant behaviors. Consequently, they learned more (Chase, Chin, Oppezzo, & Schwartz, 2009).

We have also conducted studies to examine the nature of the learning benefits from TA. The TA’s knowledge organization and visualizations emphasize reasoning through causal chains. We hypothesize that by making the agent’s thinking visible (Collins, Brown, & Holum, 1991), TA can help scaffold students’ own causal reasoning. In a pilot study with 58 sixth-graders learning about global warming, two classes were randomly assigned to use either the TA software (n = 28), or the widely-used concept mapping software, Inspiration® (n = 30). Logistical constraints necessitated the use of intact classes; however, the school matches classes on ability, and a pre-test on global warming showed no significant differences between treatment groups; p > .25. Over the course of three weeks, students completed a three-unit course on the mechanisms, causes, and effects of global warming. Both classes completed matched instructional activities that included readings, videos, hands-on experiments, and classroom discussion. After each basic unit, students constructed concept maps with the productivity-focused Inspiration® or the interactive, feedback-focused TA system. We assessed the students with three unit tests, each with eight, short-answer, paper-and-pencil questions about causal relations. All tests had questions that required short, medium, and long chains of causal inference.
Figure 3 shows the average score broken out by treatment, lesson unit, and the length of inferential chain needed to answer the question (insert Figure 3 here). After the first unit, the two groups overlap, with the TA students showing a very modest advantage for longer inferences. After the second unit, the TA students show a strong advantage for the medium-length inferences. By the final unit, the TA students show an advantage for short, medium, and long inferences. One interpretation of this pattern is that the TA students were getting progressively better at reasoning about longer and longer chains of inference in the context of global warming.

To further examine if students had internalized ways to integrate and track causal chains, the two groups of students were given an opportunity to learn new content at the end of the study, but without support from the technologies. This is a so-called Preparation for Future Learning (PFL) assessment, because students have an opportunity to learn during the assessment (Bransford & Schwartz, 1999). Students in both conditions made paper-and-pencil concept maps summarizing a new text passage on what they could do to help prevent global warming. Students received four starter nodes to get them started. Students in both conditions added approximately four concepts, with no differences between conditions; $p > .4$. The TA students, however, showed twice as many well-integrated nodes (2.5) compared to the Inspiration® students; $p < .01$.

**Purpose / objective / research question / focus of study:**

If asked, many parents and educators would agree that incorporating technology into the curriculum is a good idea for our schools. However, there are concerns that computer technologies may fail to bring “added-value” to student learning, or worse, they may displace curricula that once provided “basic-value” learning (Clarke & Dede, 2009). Another critique is that technologies may over-scaffold student learning, leaving students overly dependent on their technological scaffolds such that they cannot perform basic procedures on their own. Consider, for instance, the debates over whether students should be allowed to use hand-held calculators in school (Ellington, 2003), or whether word-processing programs and spell-checkers have degraded our nation’s writing skills (Galletta, Dureckova, Everard, & Jones, 2005).

John Dewey (1916) stated, “the aim of education is to enable individuals to continue their education . . . the object and reward of learning is continued capacity for growth” (p. 117). Given this tenet of constructivism, the gold standard for a good instructional technology is one that will not only help students learn the content-at-hand, which is an important outcome, but will also prepare students for future learning (Bransford & Schwartz, 1999).

Our studies on the effectiveness on TA have been, in the past, of relatively short duration, used specially designed content, and were taught under the strict edicts of the research designs. We wanted to see how the technology would fare in the more complex ecology of regular instruction, when teachers could integrate TA as they chose, into the flow of their normal curriculum over a sustained period of time. Of particular interest were 1) whether we could replicate our earlier PFL results showing that learning benefits persisted for students, even when no longer supported by the technology, 2) whether gains would emerge on the standard, basic-value assessments, in addition to our own added-value assessments that focused on the TA’s strength of promoting causal reasoning, and 3) whether the learning benefits would be associated with particular student behaviors in the TA system.

**Setting:**

A small, local school district agreed to use the TA technology as added-value instruction
to complement their regular science curriculum. The district had adopted the Full Option Science System (FOSS), developed by the Lawrence Hall of Science (www.lhsfoss.org). FOSS kits come complete with teacher guides, textbooks, videos, hands-on activities, worksheets, and assessments.

Population / Participants / Subjects:
The study involved six teachers and 134 5th-grade students (104 with permission to analyze their data). The six teachers and their classes had been split into two teaching teams by the school for scheduling purposes. An analysis of students’ math and reading STAR scores (Standardized Testing and Reporting) from the previous year indicated no pre-existing achievement differences between the teams; \( p > .98 \). (STAR does not include a science component in the 4th-grade). We also administered the pretest for the first FOSS unit and found no pre-existing differences; \( p > .68 \).

Intervention / Program / Practice:
Teachers were trained on TA during a full-day, in-service workshop. They were then allowed to integrate the technology as they wanted into their own lesson plans. Teachers tended to implement the TA software differently with their students, e.g. they used TA at different points in their lesson plans, or preferred one feedback tool over another. We provided each teacher with as much technical and curricular support as she wanted.

The timing of state testing plus end-of-year school events yielded different durations for the two FOSS kits, the biology-focused Living Systems (LS) and the earth-science-focused Water Planet (WP). The teachers had approximately 10 weeks for LS, and about 5 weeks for WP. Four expert maps were used for the LS kit (totaling 41 nodes with 42 links), and two maps for the WP kit (21 nodes with 21 links). Overall, teachers averaged two TA sessions per map. These differences had implications for the amount of data we could collect for each unit, as described next.

Research Design:
The study was designed as a simple cross-over. In the winter, one teaching team integrated TA with the first kit, LS. We will refer to these three classes as Cohort 1. The other team served as the control, using the FOSS materials as they normally would. We will refer to this second set of classes as Cohort 2. In the spring, the cohorts crossed over. Cohort 2 used TA for the WP kit, while the Cohort 1 teachers taught WP without TA. This permitted us to examine whether the TA benefits, if any, would continue forward when Cohort 1 students stopped using the technology for the WP kit.

Data Collection and Analysis:
For the LS kit, the teachers covered three sub-units: Human Body, Vascular Plants, and Photosynthesis & Cellular Respiration. For the WP kit the teachers covered the extensive Water Vapor sub-unit. The FOSS kits come with summative assessments for each sub-unit, called I-Checks. They contain multiple-choice, fill-in, and short-answer questions. We sorted the FOSS items into 4 content categories based on their “prompt” word: Why questions asked about causal inferences; How questions probed internal mechanisms; What questions tested declarative factual recall; and, Data questions asked students to interpret charts or tables. These items served as the measure of “basic-value” to determine whether TA displaced or augmented the intended goals of
the original curriculum. To each of the four I-Checks, we appended four “added-value” assessment items that tapped the types of causal reasoning modeled by TA.

All learning assessment items were scored on a scale of 0 to 1. Answers received 0 points (incorrect or no answer), ½ point (partially correct answer), or 1 point (correct answer). Inter-coder reliability, using a random subset of at least 20% of the answers for each item, had correlations of greater than .92 for all tests. In addition, the student scores from the study indicated that the reliability of the 16 added-value items (alpha = 0.83) matched the reliability of the 47 FOSS basic-value items (alpha = 0.86).

We also collected complete log data profiles to see when and how each student used the system and to examine whether working with TA more, or in specific ways, correlated with learning outcomes.

Findings / Results:

The added-value results for the first kit, LS, are shown on the left side of Figure 4. Cohort 1 (using TA) significantly outperformed Cohort 2 (not using TA); \( F_{(1,101)} = 5.2, p < .05 \). Moreover, the class mean for each of the three Cohort 1 teachers was higher than the class mean for each of the three Cohort 2 teachers. Thus, TA provided a significant added-value for the 5th-graders (insert Figure 4 here).

The next question addressed what changes would appear when the cohorts crossed-over in the use of TA for the WP kit. A repeated-measures analysis compared learning for the two kits crossed by condition. The interaction was significant; \( F_{(1,96)} = 4.7, p < .05 \). The interaction addresses two questions. The first was whether the learning of the Cohort 2 students would improve on the added-value questions once they used the software. Figure 4 shows that they did. The second question was whether the Cohort 1 students would continue to perform at the same level once they stopped using TA. Figure 4 shows they did.

There was a concern that the TA lessons might detract from the basic-value of the FOSS kits. To examine this issue, we analyzed the I-Check results from the LS unit (several of the teachers did not give the I-Check for the WP unit, though they did give our added-value questions as requested and analyzed above).

Figure 5 shows that the cohorts did similarly on three question types, but Cohort 1 students did better on the Why questions. A repeated-measures analysis compared question types by cohort. The interaction driven by the Why questions was significant; \( F_{(3, 84)} = 7.0, p < .001 \). Thus, the TA system did not reduce students’ learning of basic FOSS material, and improved it for the Why questions. The benefit for the Why questions fits our general story about TA, because these questions asked students about cause and effect relationships (insert Figure 5 here).

The preceding analyses compared two treatments experimentally to determine the effectiveness of TA. A complementary approach is to look at effects within the TA treatment. If the technology is responsible for improved learning, then we should expect to see “dosing effects” – students who more frequently use productive elements of the software should learn more. The following analyses are exploratory, because the effects are only correlations, and because we were not sure which aspects of the TA system were especially useful for learning if used more frequently.

System-use metrics (e.g., number of mapping sessions, map edits, resource accesses, agent queries, quizzes, etc.) were used to predict outcome performance on the added-value measures (LS for Cohort 1, and WP for Cohort 2). We forced the STAR scores into the
regression equation to control for prior achievement, and then conducted a stepwise regression with the metrics (the sample size reduces because not all students had STAR data). For Cohort 1, the stepwise regression found that number of map edits was most predictive; $F(3, 37) = 10.3, p = .001, R^2 = .45$. For Cohort 2, number of quizzes proved to be most predictive; $F(3, 38) = 6.1, p = .002, R^2 = .33$). Although the specific, predictive actions differed between the cohorts, quizzing and editing are highly correlated ($r = .70$) and indicative of productive behaviors in the system.

Conclusions:

Teachable Agents weathered its first test in the complexity of the real world, where teachers chose how to use the software for several months as an added-value to their normal instruction. Students exhibited a deeper causal understanding of the FOSS material, as measured by the added-value tests and the *Why* questions in FOSS’s own basic-value assessments. The TA activities did not displace basic learning from the FOSS kit. Moreover, the degree to which students used the map editing and feedback features correlated with learning, even after controlling for prior achievement. And finally, perhaps the most exciting result is that experience with TAs supported students’ future learning of new content, even when they no longer used the software.

In conclusion, there has been a good deal of recent discussion about the importance of 21st-century skills and competencies (Banta, 2009; Rotherham, 2008; Silva, 2008). The assumption of these discussions appears to be that times have changed, and they will continue to do so. The latter assumption – that times will continue to change – suggests that no set of static skills or competencies will do; people will need to continue to learn. Therefore, it seems worthwhile to prepare students to continue learning so they can adapt (and contribute) to dynamic times where new technologies come and go. In the current study, we demonstrated that one approach is to provide students with ways of thinking about specific content. We did not teach children how to learn in general, for example, by taking notes or explicitly self-explaining. Instead, we provided them with the powerful, integrative, and visible idea of causal chains in science. We further added interactive feedback, so they could model and refine this type of reasoning. Interactive technologies can provide added-value to many current curricula by targeting critical organizing principles with a range of feedback and visualizations, and in the process, indirectly increase the value of future curricula.
Appendices

Appendix A. References


*2010 SREE Conference Abstract: Preparing Students for Future Learning with Teachable Agents*
Appendix B. Tables and Figures

Figure 1. The Teachable Agent Interface. The student has a) named her agent “Dee,” b) customized Dee’s look, c) taught Dee about global warming, and d) asked her, “What happens to ‘heat radiation’ if ‘methane’ increases?”
Figure 2. TA-Affiliated Feedback Technologies. a) Quiz Feature: students can have their agents take a quiz to test their knowledge and determine if revision is needed. b) All-Possible-Questions (APQ) matrix: automated scoring indicates TA accuracy for all possible questions [Green = correct; Red = incorrect; Yellow = correct but reasoning path is wrong]. c) Front-of-Class (FOC) display: teachers can project and quiz multiple agents simultaneously to provide a visual anchor for classroom discussion. d) Triple-A Game Show: students can chat and have their agents compete in an on-line game for homework.
Figure 3. **Average Item Scores for Global Warming Assessments.** Scores are broken out by unit test, inference length, and treatment. One interpretation of this pattern is that the TA students were getting progressively better at reasoning about longer and longer chains of inference, relative to the Inspiration® students.

Figure 4. **Average Scores on Added-Value Items for FOSS Kits.** Scores are broken out by kit and treatment. Kit 1 results indicate that the TA system provided a significant added-value benefit for Cohort 1 students. Kit 2 results seem to show that Cohort 2 students improved once they used the software, and Cohort 1 students continued to perform at the same level once they stopped using the technology.
Figure 5. Average Scores on Basic-Value Item Types. Scores are for LS unit only and broken out by item types and treatment. Data indicate that the TA system did not reduce Cohort 1 students’ learning of the basic FOSS material, and improved it for the Why questions relative to Cohort 2 students.
Title: The Moment of Learning: Quantitative Analysis of Exemplar Gameplay Supports CyGaMEs Approach to Embedded Assessment

Author(s):

Debbie Denise Reese—Center for Educational Technologies, Wheeling Jesuit University, WV
Barbara G. Tabachnick—California State University, Northridge
The CyGaMEs Group
Background/Context:
We summarize a quantitative analysis demonstrating that the CyGaMEs toolset for embedded assessment of learning within instructional games measures growth in conceptual knowledge by quantifying player behavior. CyGaMEs stands for Cyberlearning through GaME-based, Metaphor Enhanced Learning Objects. Some scientists of learning claim that all cognition is situated, and the only way to effectively study cognition is within authentic contexts (e.g., Brown, Collins, & Duguid, 1989; Greeno, 1997). CyGaMEs assessment does not violate those assumptions. CyGaMEs assessment is authentic because it is embedded in gameplay. The CyGaMEs assessment toolset keeps track of each player’s procedural gameplay activity. That is, CyGaMEs measures learning by tracking player behavior. For the player of a CyGaMEs instructional game, progress toward the procedural game goal requires concurrent discovery and application of the targeted concepts. Thus, the game requires the player to use procedural gameplay to build conceptual knowledge. CyGaMEs assessments are algorithms that quantify player gameplay activity and progress toward the game goal as measures of learning. Within this paper we statistically demonstrate the accuracy and sensitivity of the CyGaMEs assessments. We introduce a moment of learning method, a quantitative methodology conceived by the first author in collaboration with CyGaMEs partner Larry Hedges to quantify the degree to which CyGaMEs tools assess learning. Our findings should help assuage critics who might question our claim that CyGaMEs assessment measures learning. A vision of equity and achievement in 21st century education has motivated federal agencies, national organizations, and private foundations to launch initiatives studying how cyberlearning technologies can enhance learner-centered education through game-based instructional environments and embedded assessments (e.g., Borgman et al., 2008; Laughlin, Roper, & Howell, 2006; The Learning Federation Project, 2003; http://digitallearning.macfound.org/). Those agencies and organizations call for the development of assessment toolsets that can be shared by researchers, developers, and learning environments. This CyGaMEs work supports the development of accurate and authentic instructional game assessment toolsets. Such assessments are essential if education is to enhance its responsiveness to the needs and strengths of each individual learner.

The CyGaMEs method is a theory-based approach to instructional game design, embedded assessment, and research (Reese, 2007b, 2009, in press). The CyGaMEs approach translates targeted abstract concepts into concrete, procedural game worlds. In other words, the CyGaMEs method translates what domain experts think into something domain novices do. The CyGaMEs approach to instructional game design and assessment derives from a synthesis of theories and methods: cognitive science analogical reasoning, game design, instructional design, learning science, and flow. Briefly, CyGaMEs applies structure mapping theory and pragmatic constraints (Gentner, 1983; Holyoak, Gentner, & Kokinov, 2001; Kurtz, Miao, & Gentner, 2001) to specify the design and development of a game world that is relationally isomorphic (consistent and in one-to-one correspondence) with the targeted conceptual domain. The target domain becomes the base for its game world analog. This is possible because game worlds are virtually concrete relational systems, and gameplay is designed to support game goals (Bogost, 2006; Fullerton, 2008; Schell, 2008; Wright, 2003, 2004, 2006). The relational structure of the targeted domain becomes the relational structure game world, the learning goal becomes the game goal, and this makes player progress toward the game goal a quantifiable, behavioral measure of player attainment of the targeted learning goal. The CyGaMEs assessment tool measuring player progress toward the game goal is the timed report. CyGaMEs also captures each player’s
interaction with the game world at the level with which the player changes the game state. This assessment tool is the *gesture report*. Flow theory (Csikszentmihalyi & Csikszentmihalyi, 1988; Csikszentmihalyi & Larson, 1987; Csikszentmihalyi & Schneider, 2000; Hektner, Schmidt, & Csikszentmihalyi, 2007) is an integral component of game design. Every designer attempts to inspire flow—that state in which the player loses all sense of time and self, immersed in the experience of gaming. The first author designed an assessment tool to measure the degree to which CyGaMEs instructional games place players in flow and how flow and the other dimensions (apathy, boredom, routine expertise, control, arousal, anxiety, and worry) interact with learning. The tool is the flowometer, and it produces a *flow report*.

Playing a CyGaMEs game prepares the player to make viable inferences about targeted learning. These inferences serve as prior knowledge. Apt prior knowledge for a targeted domain makes future learning of that domain more intuitive. Learning scientists call this process “preparation for future learning” (Schwartz & Bransford, 1998; Schwartz, Bransford, & Sears, 2005; Schwartz & Martin, 2004). Instructional designers prescribe this process as an event of instruction that activates and/or develops apt prior knoweldge (Gagné, 1965; Gagné, Briggs, & Wager, 1992; Merrill, 2002). Daniel Schwartz and Taylor Martin have developed an experimental design for use in research when interventions are designed to prepare learners for knowledge acquisition (Schwartz et al., 2005; Schwartz & Martin, 2004). CyGaMEs adapted this double transfer paradigm and applied it to design a research environment for studying how game-based learning assists learners to construct preconceptual mental models for targeted concepts (see Figure 1).

(Please insert figure 1 here.)

**Purpose/Objective/Research Question/Focus of Study:**
We wanted to identify a prototypical moment of learning using the gesture report and then confirm that the timed report could identify when people had accomplished that moment of learning. We asked:

- Can the gesture report identify a prototypical learning moment, the players who have achieved it, and the time it occurred?
- Can the timed report also identify if players have achieved the learning moment? That is, if we use the time at which gestures indicate the learning moment occurred, will the timed report identify an increase in player performance after the learning moment?

**Setting:**
The CyGaMEs environment comprises (a) the three embedded assessments, (b) the concrete game analog, and (c) the interstitial research environment, *Selene: A Lunar Construction GaME*. *Selene* is one complete CyGaMEs environment (Reese, 2007a, 2008, 2009, in press). *Selene* is available online to registered players 24/7. The current version of the game is authored in Java and set within a Flash shell that delivers instructional movies and external assessments. *Selene* players slingshot particles to build the Earth’s Moon (accretion), and then change it over time by peppering its surface with impact craters and flooding it with lava. As specified for *Selene*, this domain of lunar science contains 101 interrelated subconcepts.

**Population/Participants/Subjects:**
The first author triangulated video and gameplay data of one female undergraduate from a
Midwest state university psychology pool who self-selected to participate in a collaborative version of the Selene study for research credit. We then used the insights gathered by that triangulation to analyze 22 sets of participant gameplay selected from two phases of data collection (study phase 1 N=554 and study phase 2 N=119) when player data met inclusion criteria. Phase 1 player ages ranged from 13-17 (N=16; female=6, male=10). The typical phase 1 player was 15 years old and attending school grade 9 with a self-reported GPA of B and living in Arkansas (3), Arizona (1), Mississippi (1), Missouri (2), New Jersey (1), New York (2), Ohio (2), Oregon (1), or Pennsylvania (2). The majority of phase 1 players are white=11 (African-American=2, mixed=1, Other=2). About 60 percent of the phase 1 players reported parent’s education ended at high school. The other 40 percent reported their parents had earned college or graduate degrees. The six phase 2 undergraduates had self-selected from the same Midwest university psychology pool for research credit (µ grade = sophomore; µ GPA (self-reported) = C/B; female=4, male=2). Each reported father’s level of education as college; two reported mothers had completed college, and four reported mothers had terminated education at high school.

**Intervention/Program/Practice:**
All participants used an access code to log in and play the Selene game. The collaborative study player was supervised by a researcher in a lab setting and videotaped in a computer lab. Phase 1 participants were recruited by Selene adult volunteers (e.g., educators, parents, club leaders, etc.) who supervised informed consent and issued access codes. Phase 1 and phase 2 participants could play the game 24/7, independently, at any location. Players have taken from 45 minutes to 3-4 hours to complete the entire Selene environment. Data within this analysis are drawn from the first section of accretion module round 1 gameplay (accretion scale 1) and examined before and subsequent to a learning moment. The learning moment occurred at an idiosyncratic time for each player. These players took an average of 9.3 minutes to complete scale 1 (µ prelearning =6.0 and µ postlearning =3.4)

**Research Design:**
The Selene environment is constructed for randomized field trials using an adapted double transfer experimental design (see Figure 1). The design implemented to triangulate video and gameplay data for the single collaborative study participant could be partially characterized as a quantitative case study. This moment of learning analysis uses quantitative repeated measures of around 1 accretion gameplay behavior that occurs before players are differentially routed through the game. Phase 1 players who watch gameplay during round 1 were excluded. Phase 2 players were part of a larger study in which they also completed one of two pregame external assessments. Phase 1 players were not exposed to the external assessments.

**Data Collection and Analysis:**
Selene timestamps all data and sends it to a database. Two Selene embedded assessment tools measure learning:
- Timed report: A timed report is the score of player’s progress toward the game goal calculated every 10 seconds of gameplay. We interpret the scoring as continuous data, calculated for interpretation as “-1” (away from goal), “0” (no progress), or “1” (toward goal).
- Gesture report—slingshot: A gesture is a player- or game-initiated event (behavior) that changes the game state. Each gesture has parameters. During accretion scale 1, the player
initiates the slingshot gesture by selecting a particle from a ring around the early Earth and shooting it into a protomoon. The slingshot velocity parameter reports the speed of the launch.

Accretion is the concept that high kinetic energy collisions cause fragmentation and low kinetic energy collisions cause accretion (particles stick together). Selene players learn to correctly execute accretion via idiosyncratic learning pathways. Using a moment of learning method, the first author reviewed video footage of a single player’s gameplay to identify a prototypical accretion learning moment. This learning moment, accretionLM, is the instant at which a player’s behavior transitions from initiating very high velocity slingshot gestures to sustained low velocity slingshots. The same author triangulated this player’s video corpus with the player’s gesture slingshot report (velocity) and timed report data. Both of these embedded assessments bifurcated at the learning moment as expected (see Figures 2 and 3). Triangulation confirms the existence and characteristics of accretionLM. It demonstrates that timed report accurately reflects accretionLM for this player. The next step in the moment of learning method is to identify the prototypical learning moment in other players. If the timed report does, indeed, describe accretion learning, then we should be able to look at people who have accretionLM and see a change in their progress at that learning moment. The first author ran scatterplots of all players’ velocity data and identified 22 exemplar players who met inclusion criteria for learning moment analysis (i.e., initial high velocity gameplay followed by sustained, low velocity gameplay). She graphed velocity traces for each exemplar and identified the time (in milliseconds) each exemplar’s learning moment occurred. She used each exemplar’s time to split that exemplar’s timed report data into pre/post learning moment. The authors analyzed these data as repeated measures using multilevel modeling on report (trial) level data and using the general linear model on data aggregated within player by pre/post learning moment.

(Please insert figures 2 and 3 here)

Findings/Results:

Multilevel Modeling of Timed Report

A number of preliminary hierarchical models were analyzed through HLM 6.07 software to determine whether factors such as study phase (group types) and slopes of sequence within sets of trials (trials before learning vs. trials at the point of learning and after) added significantly to prediction of timed report changes averaged within each trial set (labeled learning). Full maximum likelihood estimation permitted comparisons among models with and without these factors and showed that neither factor aided the fit of the model to the responses ($p > .05$). Designating timed report changes averaged within each trial set (learning) as a random factor in a three-level model appeared to provide a better-fitting model over one in which learning was designated a fixed factor; however, no significance test comparing the two models is possible, nor was there any difference in interpretation of the learning effect. Therefore, the simpler, 2-level model is reported, based on analyses using restricted maximum likelihood estimation to provide more accurate results. First-level units of the multilevel model were trials for which velocities were measured, a total of 1,232, with the number of trails varying among participants. Second-level units were the 22 participants. A model based on individual differences alone, without predictors, permitted calculation of the variance associated with individual differences. Although there were significant differences among participants (measured as a random effect), $\chi^2 (21) = 51.30, p < .001$, the intraclass correlation was found to be quite small, $p = .023$. This suggests that the multilevel modeling approach may not be necessary, but it does provide some
insights beyond those revealed by repeated-measures ANOVA. Table 1 displays the results of the final two-level model with a single fixed predictor: learning. Table 1 shows that for every one-unit change in learning level (from prelearning trials to trials during and after learning), there is a .42 change in timed report, on a scale of −1 to 1. The random intercepts themselves, i.e., individual differences among participants, have decreased to the point that they are no longer statistically significant ($p > .5$) after accounting for differences due to learning. The statistically significant fixed intercept shows that the grand mean of responses is greater than 0, averaged over all subjects and trials.

(Please insert table 1 here)

**Repeated Measures Analysis of Timed Report**

A 2 x 2 within – between ANOVA (SPSS 15.0.2) evaluated learning (pre vs. post) across the two study phases found the timed report accurately identified learning. A single outlier was retained because corrected results mirror results from the dataset with the outlier, and the outlier dataset provides a more conservative analysis. Alpha was set to .01 to address a variance heterogeneity issue in the postlearning data. The main effect for learning is statistically significant, $F(1,20) = 358.73, p<.001$, partial $\eta^2 = .95$. Learners make little progress before the learning moment (Mean = .054, 99% CI$_{lower} = -.05$, 99% CI$_{upper} = .16$). After the learning moment players make strong progress. Their postlearning moment mean timed report value approaches 1 (Mean = .94, 99% CI$_{lower} = .88$, 99% CI$_{upper} = 1.0$). This indicates their progress is almost always successful. The main effect for study is not, $F(1,20) = .004, p = .95$, partial $\eta^2 <.001$. The interaction between learning and study also fails to reach statistical significance, $F(1,20) = 4.15, p = .055$, partial $\eta^2 = .17$. Although the interaction between study phase and learning is not significant, it does account for a substantial amount of model variance (see Figure 3) but with little statistical power (1−β = .24), suggesting a significant interaction might be expected with a larger sample (see Figure 3b). Study phase 2 player mean timed report scores evidence greater dispersion before learning. After learning there is very little variance in the scores of the six phase 2 players. Additionally their aggregate mean postlearning gameplay is almost perfect (Mean = .99, 99% CI$_{lower} = .89$, 99% CI$_{upper} = 1.1$). This suggests that one or both of the preassessments may act as a prime that enhances achievement after the learning moment.

(Please insert figure 3 here)

**Conclusions:**

*Selene* measures learning as quantified behavior. Different people have learning moments at different times. CyGaMEs identified a moment of learning for the underlying science of accretion, i.e., accretionLM, and used gesture data to identify the time at which each of 22 exemplars players achieved it. The timed report successfully ascertained when people had and had not learned accretionLM. The learning moment, in and of itself, explained 95 percent of the variance in player’s timed report progress. Thus, the timed report can be a strong and accurate measure of learning when games are designed according to the CyGaMEs approach. Future research should explore the interaction between the pretests (external assessments) and *Selene* learning. Future development work should generate a rule and algorithm that will support the *Selene* environment’s backend reporting system to automate discovery, measurement, and reporting of the accretionLM and, eventually, other moments of learning.
Appendices

Appendix A. References


Figure 1. The phase 1 design, an example of one implementation of the double transfer paradigm (Schwartz & Martin, 2004) as adapted for CyGaMEs research. In the phase 1 implementation participants either watched or played the game during round 1. Then all players played round 2. Half the players watched video instruction during round 1. Half watched video instruction after completing the round 2 game. The phase 2 design contained no watcher conditions. Instead, Phase 2 players were assigned to one of two pregame assessments, and then half watched video instruction after round 1 gameplay and half watched video instruction after round 2.

Note: PIP = play-instruction-play, PPI = play-play-instruction, WIP = watch, instruction, play, WPI = watch-play-instruction.
Figure 2. Velocity Bifurcation—Velocity trace for moment of learning, accretionLM, for case study participant. The moment of learning is marked by the orange triangle (velocity=1.5). High velocity collisions cease at the moment of learning, and participant subsequently sustained attenuation of velocity. This graph limits displays to round 1 accretion scale 1 data.
Figure 3. Cumulative timed report trace of moment of learning, accretionLM, for case study participant over two rounds of gameplay, including flowometer reports (skill and challenge) for initial instructional section and subsequent two rounds of gameplay. The dark gray arrow points to the time of the accretionLM, as identified by the velocity data analysis. This player reported a state of worry while watching the solar system accretion gameplay, reported boredom before accretionLM, and reported sustained anxiety for the next half hour of round 1 and round 2 gameplay.

Note: R1SSACC = round 1 solar system accretion, R1Acc = round 1 accretion scale 1, R1Acc2 = round 1 accretion scale 2, R1Acc3 = round 1 accretion scale 3, R1SF = round 1 surfaces features (time periods 1-3), R2Acc = round 2 accretion (scales 1-3), R2SF = round 2 surfaces features (time periods 1-3).
### Table 1. Results of Final Two-level Model of Response Timed Report

(a) Random Effect (Individual Differences, Tau)

<table>
<thead>
<tr>
<th>Effect</th>
<th>Parameter Estimate</th>
<th>Standard Deviation</th>
<th>Chi-square</th>
<th>df</th>
<th>$p$ (1-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercepts</td>
<td>0.0008</td>
<td>0.00869</td>
<td>13.07</td>
<td>21</td>
<td>&gt;.500</td>
</tr>
</tbody>
</table>

(b) Fixed Effects (Averaged over Participants)

<table>
<thead>
<tr>
<th>Effect</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t-ratio</th>
<th>Approx. df</th>
<th>$p$ (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.4944</td>
<td>0.0227</td>
<td>21.73</td>
<td>21</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Learning</td>
<td>0.4222</td>
<td>0.0269</td>
<td>18.62</td>
<td>744</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
Table 2. Means and Standard Deviations for Learning by Study

<table>
<thead>
<tr>
<th>Learning</th>
<th>Study</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>Spring 2007 (13-18)</td>
<td>0.101</td>
<td>0.160</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Fall 2007 (Undergraduate)</td>
<td>0.008</td>
<td>0.164</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0.076</td>
<td>0.162</td>
<td>22</td>
</tr>
<tr>
<td>Post</td>
<td>Spring 2007 (13-18)</td>
<td>0.894</td>
<td>0.105</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Fall 2007 (Undergraduate)</td>
<td>0.992</td>
<td>0.020</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0.921</td>
<td>0.100</td>
<td>22</td>
</tr>
</tbody>
</table>
2010 SREE Conference Abstract: The Moment of Learning: Quantitative Analysis of Exemplar Gameplay Supports CyGaMEs Approach to Embedded Assessment

Figure 3(a). Mean scores and error bars by study phase within learning.

Figure 3(b). The interaction between study phase and learning. Sample size required for $1-\beta = .80$ ($\alpha = .01$) is 50 players per study phase.
Title: Two Perspectives on the Generalizability of Lessons from Scaling Up SimCalc

Author(s):
Jeremy Roschelle, SRI International
Deborah Tatar, Virginia Tech
Larry Hedges, Northwestern
Nicole Shechtman, SRI International
**Background/context:**
At the 2008 SREE conference, we reported the results of two large-scale randomized experiments that addressed the research question “Can a wide variety of teachers use an integration of curriculum, software, and professional development to increase student learning of complex and conceptually difficult mathematics?” In both studies, one targeting 7th-grade mathematics and the other 8th-grade mathematics, the main effect was statistically significant and showed that students using a three-week replacement unit learned more than students in a business-as-usual condition. The student-level effect sizes were 0.63 and 0.50 respectively. These results were robust across the demographic groups and socioeconomic settings that were sampled. This Scaling Up SimCalc research has the potential to influence practice both by identifying a particular genre of software that may be particularly effective – software featuring dynamic mathematical representations – and the importance of integrating software, curriculum, and teacher professional development in order to achieve scalable implementations. However, to have the strongest influence on practice, issues regarding generalizability must be addressed.

**Purpose / objective / research question / focus of study:**
One purpose of educational research is to provide information about the likely impact of interventions or treatments on policy-relevant populations of students. Randomized experiments are useful for estimating the causal effects of interventions on the students in schools that participate in the experiments. Unfortunately, the samples of schools and students participating in experiments, including ours, are typically not probability (random) samples. Thus, even well-conducted experiments may not yield results that generalize to populations of interest. In the Scaling Up SimCalc experiments, one concern about the sample is that teachers were volunteers and potentially not representative of a broader teaching population. Although the volunteer teachers were randomly assigned to condition (reducing the chance that results were due to selection bias), the properties of the volunteer pool as a whole might limit generalizability to broader or differently-selected populations. A second concern is that, because pragmatic issues unrelated to sampling led to recruitment in regions with high proportions of Hispanic and Caucasian students and teachers, other groups of interest, such as African-American students and teachers, were underrepresented in the studies.

In light of these and other concerns, this paper examines generalizability from two complementary perspectives. First, we have conducted detailed analyses of the characteristics of teachers and schools participating in the sample in comparison to others in the state in which the experiments took place. Second, we present findings from a novel statistical method developed to permit principled generalization from research samples to well-defined populations.

**Setting:**
The studies took place during the 2005-06 and 2006-07 school years in 115 middle schools throughout several geographic regions across the state of Texas. Texas is an ideal state for scaling research, as it already has an aligned system of standards, curriculum, assessment, and teacher professional development, as well as a widely diverse student population with respect to ethnicity, socioeconomic status, and urbanicity. Targeted geographic regions included the large metropolitan areas of Austin, Dallas, and Fort Worth; smaller cities and more suburban and rural areas of Midland, Wichita Falls, and Lubbock; and the largely Hispanic and socioeconomically disadvantaged Rio Grande region along the Mexican border.
**Population / Participants / Subjects:**
Teachers were recruited through a statewide network of Educational Service Centers (ESCs) coordinated through the Dana Center at the University of Texas at Austin. The ESCs provide professional development support and training for teachers throughout the state. In order to ensure that the recruitment process was not biased by the types of relationships the ESCs had with particular schools or teachers, the researchers gave each ESC a randomized list of schools in their region and asked the ESCs to contact schools in that order. We believe the ESCs contacted schools in an order that reflected a balance between convenience and the technique we proposed. Nonetheless, schools and teachers were still (paid) volunteers. Thus we would not argue that our samples were random.

Table 1, 2, and 3 show the sample sizes and key characteristics of the teachers, schools and students in the two studies.

(please insert Tables 1, 2, 3 here)

**Intervention / Program / Practice:**
SimCalc interventions are an integration of three elements – software, curriculum, and professional development. In this systems approach, we do not make claims that any one of these elements is more important than the others. Participating teachers used software and curriculum materials, and received professional development designed specially for this experiment. The design goals were to exemplify the SimCalc approach and meet Texas state standards.

Hallmarks of the SimCalc approach to the mathematics of change and variation are:
1. Anchoring students’ efforts to make sense of complex mathematics in their experience of familiar motions, which are portrayed as computer animations.
2. Engaging students in activities in which they make and analyze graphs that control animations.
3. Introducing piecewise linear functions as models of everyday situations with changing rates.
4. Connecting students' mathematical understanding of rate, proportionality, and linear function across key mathematical representations (algebraic expressions, tables, graphs) and familiar representations (narrative stories and animations of motion).
5. Structuring pedagogy around a cycle that asks students to make predictions, compare their predictions to mathematical reality, and explain any differences.
6. Integrating curriculum, software, and teacher professional development as mutually supporting elements of implementation.

Figure 1 shows the key SimCalc MathWorlds software features used in these experiments that allow students to manipulate graphs and algebraic expressions that describe linear and piecewise linear motion.

(please insert Figure 1 here)

Table 4 elaborates the mathematical content covered in the SimCalc interventions. Two curriculum units were designed for these studies. The 7th grade curriculum—Managing the Soccer Team—addresses central concepts of proportionality: linear function in the form $y = kx$, $y = mx + b$, and piecewise linear functions.
and rate. The 8th grade curriculum—Designing Cell Phone Games—addresses linear function and average rate. The materials for both units were designed to be used daily over a 2–3 week period, replacing regular lessons on the same topics. Professional development for teaching these units consisted of a sequence of workshops totaling 6 days in each study. The workshops were training and planning opportunities with mathematical content, SimCalc software, and curriculum materials.

(please insert Table 4 here)

**Research Design:**
Each study was a randomized controlled experiment with pre/post measures. Teachers were randomly assigned to either a Treatment group, which received the SimCalc intervention as outlined above, or a Control group. The counterfactuals were designed such that Control teachers would receive professional development of quality and usefulness similar to the SimCalc intervention, but would not receive the SimCalc intervention and would be asked to teach the parallel content as usual. Random assignment occurred at the school level to avoid contamination between conditions within a school and provide teachers in the same school with a community of practice. Note, however, that most schools had only one participating teacher.

Figure 2 shows the experimental designs and timelines for each study. In the 7th Grade study, both the Treatment and Control groups received the Texas professional development workshop, “TEXTEAMS,” which provides important mathematical foundations for understanding proportionality. Control teachers were then asked to teach rate and proportionality as usual in their classrooms. In addition, the 7th Grade study was a delayed treatment design in which Control teachers were promised and provided the complete SimCalc intervention in a second year. In the 8th Grade study, Control teachers were provided a workshop of equal quality and relevance to their teaching (Teaching Mathematics TEKS Through Technology, “TMT3,” which focused on the content of statistics) and were asked to teach linear function as usual during the school year. In addition, to examine the effects of fading research team support, we implemented a train-the-trainer professional development model in which workshop leaders were trained to deliver the teacher workshops. This is in contrast to the 7th grade studies in which the curriculum designer delivered the teacher workshops.

(please insert Figure 2 here)

**Data Collection and Analysis:**
The measures were as follows. The primary dependent measures in both studies were student learning of relevant mathematical concepts (see Table 4 for content). Pretests were administered pre-unit, and posttests were administered post-unit. Key teacher measures included assessments of teacher mathematical knowledge for teaching; a questionnaire about teacher background, attitudes, and beliefs; a teacher log about the target class; a daily log in which teachers gave a structured report of their implementation of the unit; and a teacher retrospective log about the unit as a whole. In addition, demographic data about each participating school was drawn from the Texas Public Education Information Management System (PEIMS) datasets. PEIMS is maintained and distributed by the state of Texas and reports the results of a complete census of teachers, schools, and districts conducted yearly. Teachers also did a one-hour telephone interview about their experiences in the program.
Hierarchical linear modeling (HLM) was employed to estimate the effects of the treatment (Raudenbush & Bryk, 2002). HLM allows accounting for measurement and sampling error at both the student and classroom level, resulting in correctly adjusted standard errors for the treatment effect. While random assignment occurred at the school level, we used two-level models (students nested within classrooms) because most schools had only one teacher.

Our first additional analysis of generalizability compared teachers, classrooms and schools in the samples to two additional populations of teachers, classrooms and schools. As the study was conducted only in certain regions of Texas, we compared the sample group to the population throughout those regions. Further, we compared the sample group to the population in the entire state of Texas. Population data was extracted from the PEIMS database, which includes measures of school poverty, school ethnicity, school size, teacher gender and ethnicity, teacher certification, teaching experience.

The second additional analysis of generalizability estimates the population average treatment effect, as well as the uncertainty (standard error) of that estimate, providing a quantification of the degree of uncertainty in generalizing from the experiment to the population of interest (see Smith, 1983). It involves first stratifying the population and the experiment on the independent factors of interest and then computing the treatment effect in each stratum of the experimental sample. The population average treatment effect is estimated as a weighted mean of the stratum-specific treatment effects. To avoid the difficulty of evaluating separate treatment effects in the very large numbers of strata that occur when even a small number of contextual factors are considered jointly, this method uses propensity scores to summarize the covariates and stratify on propensity scores, which is virtually as effective in matching the population to an experimental sample (Cochran, 1968).

Findings / Results:
As previously reported, in both studies, the main effect was statistically significant and showed that students in the Treatment group learned more than students in the Control group (see Figure 3). As shown in Figure 3, the effect sizes were large overall. In both studies, the difference between the groups occurred mostly on the complex portions of the assessments. The effect sizes of the treatment on the simple portion were small. This may be because the students had high pretest scores on simple mathematics, suggesting they had already learned it.

(please insert Figure 3)

The focus of this paper is not differences between control and treatment groups; these differences were few and minor; we have argued elsewhere (Tatar & Stroter, 2009) that these differences were not a threat to the validity of the overall experimental results. This paper focuses on potential differences between the sample and larger populations.

The first generalizability analysis found that the sample involved a variety of teachers in terms of age, experience level, attitudes and teaching philosophy and a variety of campus locations, school sizes, and ethnicities. With respect to most variables, we could not detect differences in either the means or range between our sample and the regions or between our sample and the
whole state of Texas. This suggests that the results can be generalized to broader populations and settings. However, there were some key factors that limit generalizability in specific ways. First, neither of the experiments included a large urban campus – mostly likely because large urban campuses tend to have their own professional development centers and thus have weaker ties to ESCs. Second, the African-American population was not well represented in any of our studies, either at the teacher or the student level. These findings suggest that further research is necessary to determine whether the SimCalc interventions would be effective in large urban campuses and schools with large African-American populations.

The second generalizability analysis is not complete as of this writing, but will be complete before the SREE conference. Performing this analysis on the SimCalc dataset will both test the new statistical method and provide valuable information on how we should or should not generalize from our findings to larger populations, such as the state of Texas.

Conclusions:

Although it would be desirable for rigorous experiments to use probability or random samples of the target populations, in practice this is nearly impossible to achieve. Recruiting schools is difficult and under most circumstances, it is not possible to achieve a big enough pool of schools such that they can be selected randomly. Likewise, it is often not possible to sufficiently sample all populations of interest within the scope of a particular experiment, due to pragmatic, logistic, and financial limitations. Hence, the process of translating from research to practice is often limited by questions of generalizability from actual samples to broader populations.

Our research in the context of the Scaling Up SimCalc experiments has led us to propose the use of complementary approaches for examining generalizability. A foundational approach is to start out with the best sampling procedure possible, striving to achieve a broad and representative sample. We provided our recruiters with randomized lists of schools, but found that the actual schools they contacted reflected a tension between random selection and convenience. We complemented this procedure with two additional analyses. The first found some ways in which our samples do not reflect the full diversity of Texas; in particular we did not have large urban districts and did not have many African American participants. However, with regard to many other characteristics, our sample is not systematically different from the full population in the state of Texas. Our hypothesis is that the second analysis will predict positive effects for all populations of interest, but with wider confidence intervals for populations that were undersampled. Thus, we should have good confidence in how our results generalize to Hispanic schools but less confidence as to how the generalize to African American schools. Overall, this affects how we share the results of our research with the practitioner community.

The overall contribution of this research is to suggest a strategy for using experiments that are conducted on nonprobability (nonrandom) samples to draw inferences about the average treatment effects in well-defined, policy-relevant populations, taking full account of the clustered structure of the experimental data. The significance to the field is that our approach permits policy researchers to use the same experiment to evaluate the likely effect of an intervention in different policy-relevant populations (e.g., different cities or states). The method quantifies the uncertainty involved in these inferences, revealing the limits of generalizations that are possible.

2010 SREE Conference Abstract: Two Perspectives on the Generalizability of Lessons from Scaling Up SimCalc
Appendices

Appendix A. References


### Appendix B. Tables and Figures

<table>
<thead>
<tr>
<th>Group</th>
<th>7th Grade</th>
<th></th>
<th>8th Grade</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Teachers</td>
<td>Students</td>
<td>Teachers</td>
<td>Students</td>
</tr>
<tr>
<td>Control</td>
<td>47</td>
<td>825</td>
<td>23</td>
<td>303</td>
</tr>
<tr>
<td>Treatment</td>
<td>48</td>
<td>796</td>
<td>33</td>
<td>522</td>
</tr>
<tr>
<td>Total</td>
<td>95</td>
<td>1,621</td>
<td>56</td>
<td>825</td>
</tr>
</tbody>
</table>

**Table 1.** Sample sizes by study and group
### Table 2. Teacher-level characteristics of the samples.

<table>
<thead>
<tr>
<th>Variable</th>
<th>7th Grade Study</th>
<th>8th Grade Study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Treatment</td>
</tr>
<tr>
<td>Total count</td>
<td>47</td>
<td>48</td>
</tr>
<tr>
<td>Percent female</td>
<td>81</td>
<td>77</td>
</tr>
<tr>
<td>Years teaching total (mean)</td>
<td>10.5</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td>Range: 1 – 29</td>
<td>Range: 1 – 40</td>
</tr>
<tr>
<td>Years teaching mathematics (mean)</td>
<td>9.5</td>
<td>11.0</td>
</tr>
<tr>
<td>Teacher ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent white</td>
<td>70.2</td>
<td>77.1</td>
</tr>
<tr>
<td>Percent Hispanic</td>
<td>25.5</td>
<td>20.8</td>
</tr>
<tr>
<td>Percent Asian</td>
<td>4.2</td>
<td>0</td>
</tr>
<tr>
<td>Percent African-American</td>
<td>0</td>
<td>2.1</td>
</tr>
<tr>
<td>Percent with a master’s degree</td>
<td>17.0</td>
<td>18.8</td>
</tr>
</tbody>
</table>
Two Perspectives on the Generalizability of Lessons from Scaling Up SimCalc

Table 3. School-level characteristics of the samples.

<table>
<thead>
<tr>
<th>Variable</th>
<th>7th Grade Study</th>
<th></th>
<th>8th Grade Study</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Treatment</td>
<td>Control</td>
<td>Treatment</td>
</tr>
<tr>
<td>Total count of schools</td>
<td>37</td>
<td>36</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>Percent free lunch (mean)</td>
<td>53</td>
<td>54</td>
<td>43</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Range: 3 – 99</td>
<td>Range: 2 – 94</td>
<td>Range: 0 – 89</td>
<td>Range: 0 – 92</td>
</tr>
<tr>
<td>Campus ethnicity (mean)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent white</td>
<td>44</td>
<td>47</td>
<td>61</td>
<td>55</td>
</tr>
<tr>
<td>Percent Hispanic</td>
<td>49</td>
<td>45</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td>Percent Asian</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Percent African-American</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>
Two Perspectives on the Generalizability of Lessons from Scaling Up SimCalc

<table>
<thead>
<tr>
<th>Overview of Concepts</th>
<th>M₁ – Conceptually Simple</th>
<th>M₂ – Conceptually Complex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Concepts are typically covered in the grade-level standards, curricula, and assessments</td>
<td>Building on the foundations of M₁ concepts, constitute more complex building blocks of the mathematics of change and variation found in algebra, calculus, and the sciences</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7ᵗʰ Grade Studies (focus on rate and proportionality)</th>
<th>Simple ( \frac{a}{b} = \frac{c}{d} ) or ( y = kx ) problems in which all but one of the values are provided and the last must be calculated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basic graph and table reading without interpretation (e.g., given a particular value, finding the corresponding value in a graph or table of a relationship)</td>
</tr>
<tr>
<td></td>
<td>Reasoning about a representation (e.g., graph, table or ( y = kx ) formula) in which a multiplicative constant “( k )” represents a constant rate, slope, speed, or scaling factor across three of more pairs of values that are given or implied</td>
</tr>
<tr>
<td></td>
<td>Reasoning across two or more representations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8ᵗʰ Grade Study (focus on linear function)</th>
<th>Categorizing functions as linear/nonlinear and proportional/ nonproportional</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within one representation of one linear function (formula, table, graph, narrative), finding an input or output value</td>
</tr>
<tr>
<td></td>
<td>Translating one linear function from one representation to another</td>
</tr>
<tr>
<td></td>
<td>Interpreting two or more functions that represent change over time, including linear functions or segments of piecewise linear functions</td>
</tr>
<tr>
<td></td>
<td>Finding the average rate over a single multi-rate piecewise linear function</td>
</tr>
</tbody>
</table>

Table 4. Core mathematical concepts in the studies’ curricula and assessments
Figure 1. SimCalc MathWorlds software, showing a graph, table, and animation. All mathematical representations are linked so that changes in one representation are immediately reflected in the other representations of the same function.
Figure 2. Experimental designs and timelines for the two studies.
### Two Perspectives on the Generalizability of Lessons from Scaling Up SimCalc

#### 7th Grade Study
*30-item assessment*

<table>
<thead>
<tr>
<th>Gain (Posttest - Pretest)</th>
<th>7th Grade Study (N=95)</th>
<th>8th Grade Study (N=56)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Score</td>
<td>9.04*</td>
<td>5.38*</td>
</tr>
<tr>
<td></td>
<td>0.63</td>
<td>0.56</td>
</tr>
<tr>
<td>M1 Items</td>
<td>1.82</td>
<td>1.61</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.19</td>
</tr>
<tr>
<td>M2 Items</td>
<td>10.03*</td>
<td>7.62*</td>
</tr>
<tr>
<td></td>
<td>0.89</td>
<td>0.81</td>
</tr>
</tbody>
</table>

*p<0.0001

#### 8th Grade Study
*36-item assessment*

<table>
<thead>
<tr>
<th>Gain (Posttest - Pretest)</th>
<th>7th Grade Study (N=95)</th>
<th>8th Grade Study (N=56)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Score</td>
<td>9.04*</td>
<td>5.38*</td>
</tr>
<tr>
<td></td>
<td>0.63</td>
<td>0.56</td>
</tr>
<tr>
<td>M1 Items</td>
<td>1.82</td>
<td>1.61</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.19</td>
</tr>
<tr>
<td>M2 Items</td>
<td>10.03*</td>
<td>7.62*</td>
</tr>
<tr>
<td></td>
<td>0.89</td>
<td>0.81</td>
</tr>
</tbody>
</table>

*p<0.0001

**Figure 3.** Student mean difference scores (± SE of total using HLM) and effect sizes at the student level.