Scaling Up SimCalc Project

Generalizability of a Technology-Based Intervention to Enhance Conceptual Understanding in Mathematics

Technical Report 10 | November 2018

Jeremy Roschelle, Digital Promise Global | Elizabeth Tipton, Teachers College, Columbia University | Nicole Shechtman & Philip Vahey, SRI International

Report Series Published by SRI International

SRI Education™

Research collaborators:

University of Massachusetts, Dartmouth; Virginia Polytechnic Institute and State University; The University of Texas at Austin; and the Charles A. Dana Center at the University of Texas at Austin

SRI Education™

SRI Education
333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000

Generalizability of a Technology-Based Intervention to Enhance Conceptual Understanding in Mathematics

Prepared by:

Jeremy Roschelle, Digital Promise Global

Elizabeth Tipton, Teachers College, Columbia University

Nicole Shechtman, SRI International

Philip Vahey, SRI International

Acknowledgments

This report is based on work supported by the National Science Foundation under Grant No. 0437861. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We thank J. Knudsen, K. Rafanan, S. Vasquez, G. Trinidad, L. Gallagher, H. Javitz, T. Lara-Meloy, M. Robidoux, S. Empson, S. Hull, L. Hedges, S. Goldman, H. Becker, J. Sowder, G. Harel, P. Callahan, F. Sloane, B. Fishman, K. Maier, J. Earle, R. Schorr, and B. McNemar for their contributions to this research. We thank the participating teachers and Texas Educational Service Center leaders from regions 1, 6, 9, 10, 11, 13, 17 and 18; this project could not have happened without them. We thank and remember Geneva Haertel for her dedication to fair and equitable assessment and Jim Kaput, who pioneered SimCalc as part of his commitment to democratizing mathematics education.

Generalizability of a Technology-Based Intervention to Enhance Conceptual Understanding in Mathematics

Three previously reported experiments found that a technology-enhanced intervention increased student conceptual understanding of mathematics in Texas. To investigate generalizability to broader populations and settings, we triangulate among three methods. First, we examine interactions between demographic variables and intervention effects. We found that the intervention was not sensitive to typical variations in school populations. Second, we use propensity score methods to measure the match between the sample and a broader population. The sample matches the school population in Texas, with minor exceptions; we report adjusted effect sizes. Third, quasi-experimental research with populations outside of Texas are considered. Results from Florida and England were consistent with Texas findings. Across three methods, the results suggest that the experimental findings generalize across populations and settings. This work also establishes a practical approach to investigating generalizability in experimental research in schools.

Introduction

How can educational researchers establish the degree to which a positive finding is likely to generalize from a sample to a broader inference population? The textbook answer is that the original research should be conducted with a probability sample of the target population (Cochran & Cox, 2002). In practice, however, it is rarely possible to obtain a probability sample of schools (Olsen, Orr, Bell, & Stuart, 2013). Educational researchers, constrained by the pragmatics of recruiting and working with schools, rarely have the ability to conduct their research in a perfect microcosm of the full population of interest (e.g., Tipton et al, 2016).

And yet, considering generalizability of findings is important (Hedges, 2013; Orr, 2015). Even when an intervention is found to work in one school setting, it may not work in other school settings. This may be particularly true when the intervention involves technology, as schools have different capacities to integrate technology into instruction (Means & Haertel, 2004). Thus school leaders rightfully ask of even technologies with a research base "yes, it worked in the schools that participated in the research, but will it work in my school district?" Likewise, to cover the diversity in conditions, practices, and people found in a large country such as the United States, policy makers need to know not only if a new intervention for science or math education "works" but also about variability in the instructional contexts in which it works (Berliner, 2002).

This article considers external validity for a particular intervention aimed at increasing students' conceptual understanding of challenging topics in mathematics. In primary research, a series of experiments was conducted across varied demographic settings to evaluate the efficacy of this intervention and, as will be discussed below, the results were positive.

This article focuses on additional research undertaken to investigate the extent to which these findings might generalize beyond the population and setting of the initial study. Cook and Campbell (1979) argue for considering generalization in two ways: (1) *across* subpopulations in an experimental sample and (2) *to* additional target populations of interest, beyond an experimental sample. To establish generalization across subpopulations of interest, interactions between background factors and the intervention should be evaluated. With regard to external validity beyond the confines of a particular experimental sample, the two key concerns are variation in populations and variation in settings (Bracht & Glass, 1968). To evaluate the first concern, researchers examine the degree to which the experimental sample matches the larger population of interest (e.g., Stuart, Cole, Bradshaw & Leaf, 2011; Tipton, 2014). To evaluate the second concern, additional research can be conducted in additional settings.

This article considers exactly these aspects by triangulating among three possible ways to explore generalizability: (1) by looking for interactions between contextual variables within the original experimental sample; (2) by estimation of a population average treatment effect for a clearly defined target population (by adjusting the experimental sample); and (3) by conducting additional research in new and different settings. We report on the approach to triangulation and its findings for a particular use of technology in mathematics education, considering the strengths and weakness of each method in the triangulation. To establish a context, we first provide background on the nature of the intervention and the previously published results.

Dynamic Representations to Improve Conceptual Understanding

Although using technology is frequently recommended in policy statements regarding the improvement of mathematics and science learning, the different technologies proposed for education do not have equally strong research evidence (U.S. Department of Education, 2010). Technologies have different mechanisms for influencing the learning process – a technology that provides social network among students is different from one that helps students to visualize mathematical concepts and is different again from a technology that "personalizes" learning.

One application of technology that has a relatively strong research base in mathematics education is the use of dynamic representations to enhance visualization of complex ideas and relationships (Kaput, 1992; Heid & Blume, 2008). For example, an algebraic function can be linked to a graph, a table and a simulation of a familiar motion, such that changes in one representation are immediately reflected in the other representations. Although the research base on dynamic representations mostly consists of small studies, reviews suggest that dynamic representations may be particularly effective for increasing students' conceptual understanding (Kaput, Hegedus, & Lesh, 2007). Theorists in mathematics education describe conceptual understanding as connections or relationships among diverse aspects of a concept and its uses (Hiebert & Carpenter, 1992). A particularly important set of relationships is among multiple representations of the concept, for example, the concept of "rate" as represented by the slope of a line in a graph or as the co-variation of pairs of numbers in a table (Brenner et al, 1997). Technology has the potential to help students to make sense of these relationships by revealing how change in one representation (for example, an increased slope) corresponds to change in another representation (for example, more rapid growth of the numeric value of a dependent variable as the numeric value of the independent variable increases) (Kaput, 1992).

To leverage this potential in classrooms, program developers created two interventions, one for 7th grade mathematics and one for 8th grade mathematics. Each was an integrated "curricular activity system" (Roschelle, Knudsen & Hegedus, 2010; Vahey et al., 2013) comprised of technology, workbook, and professional development components to make appropriate use of the dynamic representation capabilities more likely across varied classroom settings. The curriculum workbooks and activities were used as replacement units to tackle difficult concepts in the middle school mathematics curriculum. There were two replacement units developed, each intended for approximately three weeks of instruction, addressing key topics on the path to Algebra. The unit for 7th grade focused on rate and proportionality, and the unit for 8th grade focused on linear function. Both the already-published experimental results and this report on generalizability thus refer to an integration of dynamic representation software, paper curriculum and teacher professional development.

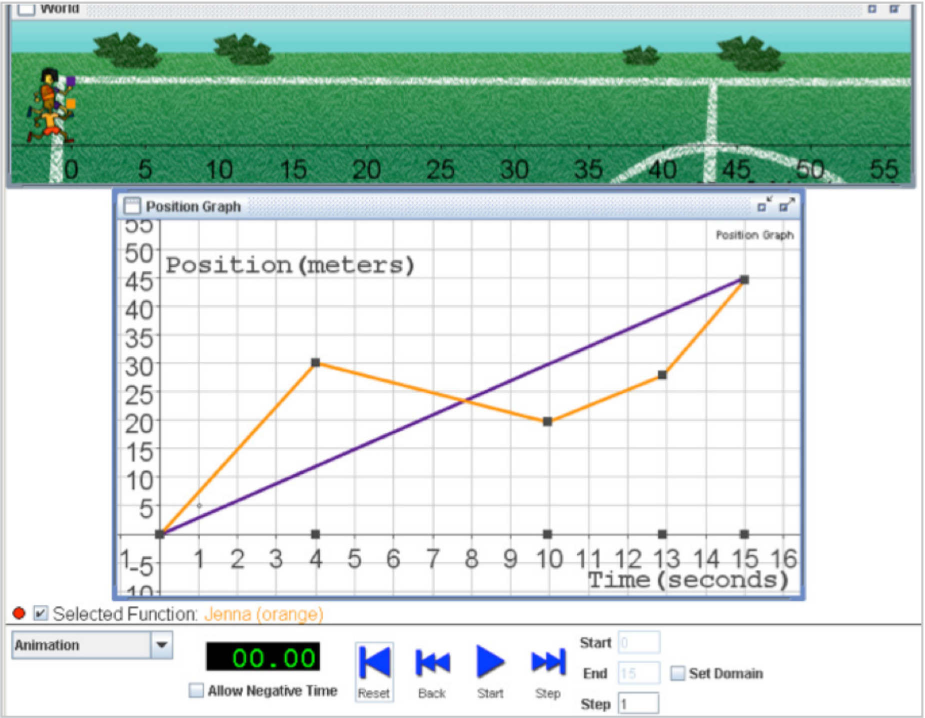
Within these integrated units, the hallmarks of the SimCalc approach to the mathematical topics are the following:

1. Anchoring students' efforts to make sense of conceptually rich mathematics in their experience of familiar motions, which are portrayed as computer animations;
2. Engaging students in activities to 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 and proportionality 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 with mathematical reality, and explain any differences.

Most distinctively, the software presents animations of motion related to position vs. time graphs (see Figure 1). Students can control the motions of animated characters by building and editing mathematical functions in either graphical or algebraic forms. After editing the functions, students can press a play button to see the corresponding animation. Functions can be displayed in algebraic, graphical, and tabular form, and students are often asked to tell stories that correspond to the functions (and animations).

The software is meant to be used in what Dewey described as a cycle of "doing and undergoing" (Dewey, 1938) – wherein students investigate a mathematical phenomenon by changing the variable features of one mathematical representation and then seeing what happens in other representations. The program developers view student use of the software, use of pencil-and-paper workbooks, and teacher explanations and teacher-led discussions as complementary activities; classroom discussion are expected to help students to understand the connections among representation that they are experiencing via the software. The workbooks guide students to conduct investigations in a curricular sequence aimed at gradually building a robust understanding of the relationships entailed in a concept such as rate and the professional development aims to support teachers in using the software and workbooks as fodder for their students' conceptual development.

Figure 1: SimCalc screen showing two representations, the animation of soccer players running on a field and a graph of their motion



Published Experimental Research

Key elements of previously published experimental research from Scaling Up SimCalc form the basis for the investigations of generalization reported below. To set the context, we begin with a summary of this research (Roschelle et al, 2010).

The samples for the experiments reported in the published findings (Roschelle et al, 2010) were classrooms in Texas schools. Researchers chose Texas for this work because it is a large and diverse state and already had a public and fairly comprehensive database of all schools (this is common today but was not common a decade ago when this research began). Access to quantitative data that describes variation in schools throughout a state is essential to examining generalizability, as without population data, there is no means for examining the degree to which a sample is representative of the broader population. Schools in Texas vary notably in poverty (typically measured by the percentage of students enrolled in a federal free or reduced price lunch program), as well as in ethnic composition (e.g., some schools are predominantly White and others are predominantly Hispanic). Many factors can contribute to schools' capacity to effectively integrate technology. While large, the experimental samples included only about 7% of the public middle schools in Texas.

Three studies, conducted in Texas in the 2005-06 and 2006-07 school years, tested the effects of the two 3-week interventions for 7th and 8th grade mathematics compared to traditional instruction addressing the same content. Study 1 was a randomized experiment with 7th-grade teachers in which 48 treatment group teachers (796 students with full data) implemented the intervention and 47 control group teachers (825 students) implemented their usual curriculum. Study 2 was a quasi-experiment in which we followed the control group teachers from Study 1 as they implemented the intervention in the subsequent year. This allowed a quasi-experimental comparison of two cohorts of students taught by the same teacher across years. A total of 30 teachers completed this research in Study 2, with 510 students in year 1 and 538 students in year 2. Study 3 was a randomized experiment in with 8th grade teachers in which 33 treatment teachers (522 students) implemented the intervention and 23 control-group teachers (303 students) implemented their usual curriculum. In addition, the 8th-grade experiment employed a train-the-trainer model of TPD, an important basis for efforts to further scale up the intervention.

The tested interventions addressed rate and proportionality in 7th grade and linear function in 8th grade, foundational topics for Algebra. Both topics are part of the transition to multiplicative reasoning, considered one of the essential difficulties that students experience in middle school mathematics on the path to Algebra and beyond.

Outcomes were measured on researcher-developed assessments addressing conceptual understanding of rate and proportionality (7th grade) and linear function (8th grade). Researcher-developed measures were used rather than the Texas state test because the Texas state test was weak in measuring conceptual understanding. Each of the two researcher-developed assessments had two subscales. One subscale addressed foundational concepts typically covered in the grade-level standards, curricula, and assessments. Many of these items were drawn from released items from the Texas state test, and the subscale was intended to test for any improvement or harm to baseline student performance as already measured by the state. The second subscale addressed the essentials of conceptual understanding that are building blocks for algebra, calculus, and the sciences. Many of these items were drawn from released items from NAEP, TIMSS, and other tests. Each assessment was subjected to a set of validation studies, including expert panel review, cognitive think-alouds, and field-testing. The 7th grade assessment of rate and proportionality had 30 items and an alpha of .86, and the 8th grade assessment of linear function had 36 items and an alpha of .91. More details about the assessment and assessment development process are described elsewhere (Shechtman, Haertel, et al, 2013). In each case, the same test was administered as both a pre-test and a post-test before and after the replacement unit or usual curriculum.

The analysis used hierarchical linear modeling to account for the nesting of students within teachers within schools. Analyses revealed statistically significant main effects of pretest to posttest gain, with student-level effect sizes of .63, .50, and .56 (Roschelle et al, 2010), with greater student learning gains for students in classrooms that used the SimCalc intervention, compared to classrooms that covered the same topic using existing materials. Further, when additional researchers followed up with teachers in participating schools a year after the formal research was complete, more than 50% were continuing to use the intervention without any further incentives (Fishman, Penuel, Hegedus & Roschelle, 2011). Teachers who perceived coherence (or alignment) of the materials with their broader context and who perceived the materials to be uniquely valuable were more likely to report continued use of the materials (Fishman, Penuel, Hegedus & Roschelle, 2011; Hegedus et al, 2009).

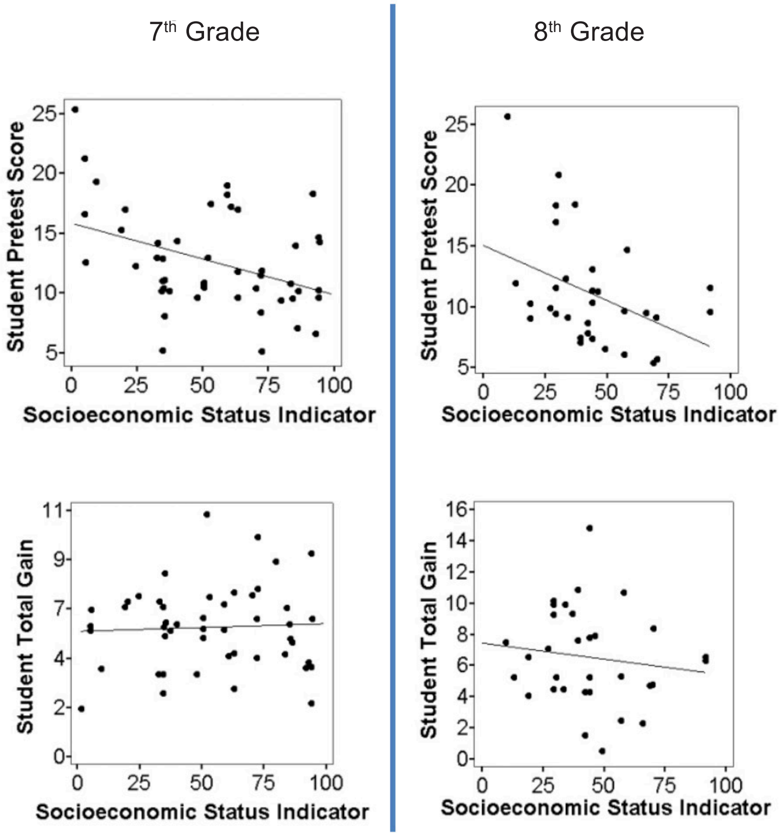
We now proceed to triangulate among three approaches to generalizability.

1. Looking for Interactions of the Treatment Effects

One technique for investigating generalizability is to examine if there are interactions of the treatment effects between experimental group and policy-relevant contextual variables (Cronbach & Snow, 1977). The logic of this type of approach is that if the sample has substantial variability in contextual variables and the treatment effect does not vary significantly across these variables, then variability of the treatment effect across these variables in broader populations is also unlikely to be significant.

In the 7th grade and 8th grade experiments in Texas, we examined the school-level contextual variables of the percentage of students qualifying for free and reduced price lunch programs (FRL), a common indicator of socioeconomic status, and the highly-correlated percentage of Hispanic students in the school. Figure 2 shows the relationships between socioeconomic status of the sample and student achievement in both studies. First, note that the schools in the Texas sample varied substantially in the percentage of students qualifying for FRL. The relevance of this variable to generalizability of the treatment effects is underscored by its correlation with the studies' pretests: mathematics pretest scores were inversely correlated with socioeconomic status of students in the school. However, this variation was weakly correlated to pretest to posttest gains: use of dynamic representations increased learning over

Figure 2: In the treatment classrooms in both 7th and 8th grade, pretest scores correlated with socioeconomic status, but learning gain scores did not correlate with prior status.



traditional materials regardless of socioeconomic status. Findings were similar with the factor of percentage of Hispanic students, potentially important because of the question of whether these English-only materials would help students in regions where English may not be students' only or preferred language.

To further explore this, a post hoc analysis was conducted by configuring a series of regression models to test whether any of the following variables mediated the treatment effect within the treatment group: teacher background, attitudes, and mathematical knowledge; student and school demographics; and qualities of implementation. This analysis found very few and very weak interactions. Among the few noticeable and interpretable interactions, teachers' expectations correlated with student learning gains. Students who were rated "low achieving" by their teacher before the study tended to learn less in the study. Overall, the analysis of these models suggested that effects of the intervention integrating dynamic representations are not likely to vary dramatically across settings. Indeed, 91% of 7th grade classrooms in the treatment group had mean classroom gains exceeding the mean gain in the control group, suggesting that most classrooms in Texas could see a benefit from using dynamic representations to address students' conceptual understanding of mathematics.

While these variables were not statically significant, a set of published case studies provide further information about specific implementation factors that may have contributed to variability. The case studies were selected from the study population in the Texas experiments and were chosen purposefully to examine variation (not at random or representatively). One set of case studies of 7th grade teachers documented variation in pedagogical approaches (e.g., emphasis on full classroom or small group work) across classrooms that implemented the intervention and found that more than one pedagogical style was compatible with high classroom learning gains (Empson et al, 2013). Another set of case studies of 8th grade teachers investigated a train the trainer model and found that implementation varied greatly among teachers with the same trainer (Dunn, 2009). These case studies considered some classrooms with much lower than average learning gains while using the intervention and documented some factors that might explain poor learning results from the unit of instruction. The identified factors were basic: teachers who skipped lessons, did not allow students to use technology, or did not talk about the key mathematical ideas in the unit of study had much poorer outcomes (Dunn, 2009). Nonetheless, most teachers had better classroom learning gains in the treatment condition than did comparable teachers who taught the same mathematical topic without access to the intervention. A third set of case studies examined classroom discourse and found stronger learning gains in classrooms where teachers pressed for understanding and engaged students in extended discourse about mathematical concepts (Pierson, 2008). These case studies suggest the approach is applicable with teachers who vary in pedagogical style, that effects vary with some features of classroom implementation, and that teachers must adhere to some basic implementation requirements in order to obtain learning gains. Subsequently, the development team incorporated these insights into refined teacher professional development materials and workshops.

These secondary analyses, along with the review of published case studies, can provide guidance and support to those who might seek to use the intervention in other settings. For example, there is no suggestion that results should be expected to apply only to particular populations or teachers. The analyses also help to identify key conditions and practices for successful implementation. However, these analyses only report additional information about the study population and does not consider ways in which the study population may not have been a good approximation to a probability sample of schools in Texas or elsewhere. We consider this challenge in the next section.

2. Using Propensity Scores to Estimate a Population Average Treatment Effect

To further examine generalizability, we used propensity score methods (e.g. Rosenbaum and Rubin, 1983, 1984) to first compare the experimental sample to a broader population of schools in Texas (Tipton, 2014) and second, to statistically adjust for any differences that arose using post-stratification (Tipton, 2013; O'Muircheartaigh & Hedges, 2014). In the statistically ideal study, the original research team would have selected the schools using probability sampling and doing so would have resulted in a sample of schools that mimicked the distribution of variation throughout Texas schools. However, as in most experimental research, pragmatic constraints of recruiting and working with schools did not enable precise probability sampling. Propensity score and post-stratification methods allow evaluation of how different the achieved sample was from the population and to statistically adjust for these differences. The goal is to both assess generalizability and to provide an improved estimated of the average treatment effect for the intervention for a well-defined population.

First, we defined the inference population using Texas's state longitudinal data system (i.e., the Academic Excellence Indicator System [AEIS]) for the school year of the study (2006-7). The resulting population included 1,713 non-charter schools, each which contained at least one 7th grade classroom. Each school in the experiment and pilot study (n = 92) was then located in this population data set. We then selected 26 school-level variables to compare the sample and population. Based on guidance provided by Tipton (2013), we chose covariates that could potentially moderate the effect of SimCalc on school average student mathematics knowledge. These variables are listed in Table 1 below.

Table 1: Variables used in the propensity score model

Category	Variables
Test Scores	(1) % 7th grade passing math TAKS (Spring 2007); (2) % 7th grade passing reading TAKS (Spring 2007); (3) % 3-11 grade commended achievement overall; (4) % 3-11 grades commended achievement math; (5) % 3-11 passing all TAKS; (6) % 3-11 passing math TAKS
Student demographics	(1) % students Black; (2) % of students Hispanic; (3) % students mobile; (4) % students on disciplinary action; (5) % students economically disadvantaged; (6) % of students LEP; (7) % of students at-risk; (8) 7th grade retention rate
School structure	(1) number of 7th grade students in school; (2) % of school in 7th grade; (3) total number of teachers in school; (4) student-teacher ratio in school; (5) average teacher tenure at school
Teacher demographics	(1) % teachers with 0 years experience; (2) % of teachers with 1-5 years; (3) % of teachers with 20 or more years; (4) average teacher years experience; (5) % of teachers Black; (6) % of teachers Hispanic
Geography	(1) school in rural county

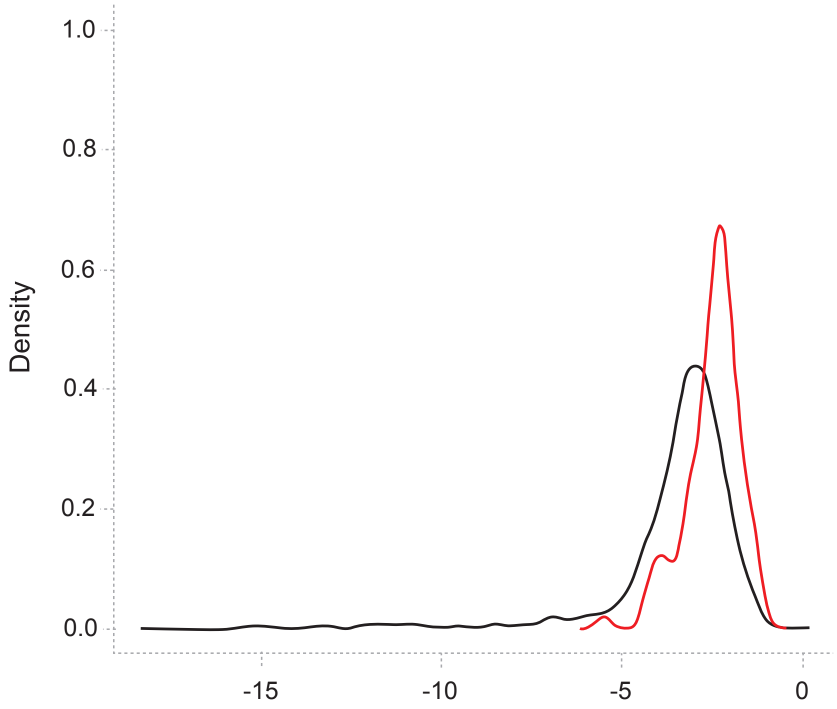
Note: TAKS is the Texas Assessment of Knowledge and Skills.

The virtue of the propensity score approach is that it reduces a difficult multivariate matching problem to a univariate matching problem, making statistical analysis more tractable. The method compares the sample and population via a logistic regression model with the selected covariates. For the purpose of these comparisons, we used an extended sample that also included schools that participated in a pilot study (Tatar et al, 2008) that was conducted before the main experiments and had results similar to the experiments. The advantage of pooling the pilot with the experimental sample was to obtain broader coverage of the population throughout Texas.

A first step in using propensity score methods is to first assess how similar the sample and target population are in terms of the selected covariates. To do so, the distributions of propensity score logits are compared using a generalizability index (Tipton, 2014). This index ranges from 0 to 1, with a 1 indicating that the sample is an exact miniature of the population on the selected covariates. Tipton, Hallberg, Hedges, and Chan (2017) develop an approach for hypothesis testing based upon a transformed value of β; according to this test, in 95% of random samples, with p = 26 covariates and n = 92 schools, we would expect the index to be greater than 0.98. However, Tipton (2014) shows that values of β greater than 0.80 indicate that while the sample differs from that of a random sample, these differences between the sample and population can be easily adjusted for using post-stratification.

Using the data in this study plus the pilot, we estimate the generalizability index to be β = 0.91. This value indicates that while the overall sample is not as representative of the target population on these 26 covariates as a similarly sized random sample would be, that the differences that remain are small and can be easily adjusted for using post-stratification. This degree

Figure 3: Distribution of propensity score logits in experiment and population



Note: In this figure, the distribution with the longer tail corresponds to the population, while the shorter tail corresponds to the sample (n = 92 schools).

of similarity can be seen visually in Figure 3, wherein the distributions of the logits of the probabilities of selection into the study are compared for schools in the study to those in the target population. Note that the distribution for units in the population has both a larger mean and a right skew; the units in the left tail have very small probabilities of being selected into the experiment. Importantly, this degree of generalizability is better than if we exclude the pilot (β = 0.85) or if we only use the pilot data (β = 0.80).

Since the sample was not completely representative of the population – and given the high generalizability index value – the next step was to develop an estimate of the population average treatment effect using post-stratification. The idea of a post-stratification estimator is to reweight the experimental data so that the distribution of covariates in the sample is similar to that in the population. In doing so, the treatment effect estimate and the uncertainty of this estimate are adjusted, both of which are important for school leaders and policy makers. The benefit of this approach in the sample selection context, as compared to other propensity score approaches, is that it requires less exact matches while reducing non-random sample selection bias as much as 50-70% with four strata (Tipton, 2013). For this analysis, we chose an estimator with four strata, defined such that each contains 25% of the population. Table 2 shows that the estimate of the population average treatment effect is about 2% larger while the standard error is about 4% larger.

Because school leaders often ask "did your study consider schools like mine?" we performed a further analysis to determine the range of schools in Texas that have a reasonable match to schools in the available data. There were no charter schools in data set, so we cannot evaluate whether these results generalize to charter schools. By examining the distributions of propensity score logits in Figure 3, we can see that 92% of Texas non-charter middle schools with a 7th grade mathematics class have a good match in the study. (The remaining 8% of schools in the population – those in the long tail of the distribution – do not have any similar schools in the study.) The map in Figure 4a shows the districts containing schools in the experiment or in the pilot study; Figure 4b shows the districts that contain schools for which there is a good match in the experiment or pilot.

Figure 4b reveals that most schools in Texas can be matched to schools within the data set. However, not all regions of Texas participated in the original studies and data was therefore not available for some configurations of school characteristics, such as schools with a high proportion of African American students.

Table 2: Estimates of the population average treatment effect by stratum

Stratum	Estimate	SE	Weights		95% CI	
			w_e	w_p	lower	upper
1	1.42	0.20	0.61	0.25	1.03	1.82
2	1.46	0.29	0.24	0.25	0.90	2.03
3	1.17	0.38	0.08	0.25	0.42	1.91
4	1.81	0.30	0.07	0.25	1.23	2.40
Post-stratification	1.47	0.15	1.00	1.00	1.17	1.76
Conventional	1.44	0.14	1.00	1.00	1.17	1.72

Note: The estimates of the treatment effect given here are effect sizes standardized by the standard deviation of school means in the experiment.

Overall, the advantages of these propensity score-based methods are that they allow an assessment of the degree of mismatch between the available sample data and a well-defined inference population and that they enable quantifying both the degree of bias and uncertainty of generalization. What is particularly useful is that this method requires clear definition of an inference population, thus highlighting that the results of an experiment may generalize well to one population but not to another (such as to charter schools or to those with a high proportion of African American students). A limitation of these methods, however, is that they are limited to the matching variables available for the entire population and these may not include the all of the factors that produce variation in treatment effects. For example, the database we used does include information about each school's technology infrastructure and a school's technology infrastructure is an important aspect of its capacity to implement the intervention. As states build more extensive data systems for their schools – an important current policy trend – researchers abilities to use propensity matching to analyze generalizability will improve.

Figure 4a: Experimental and Pilot Samples: Colored regions show Texas school districts containing schools in which the experiment and pilot took place.

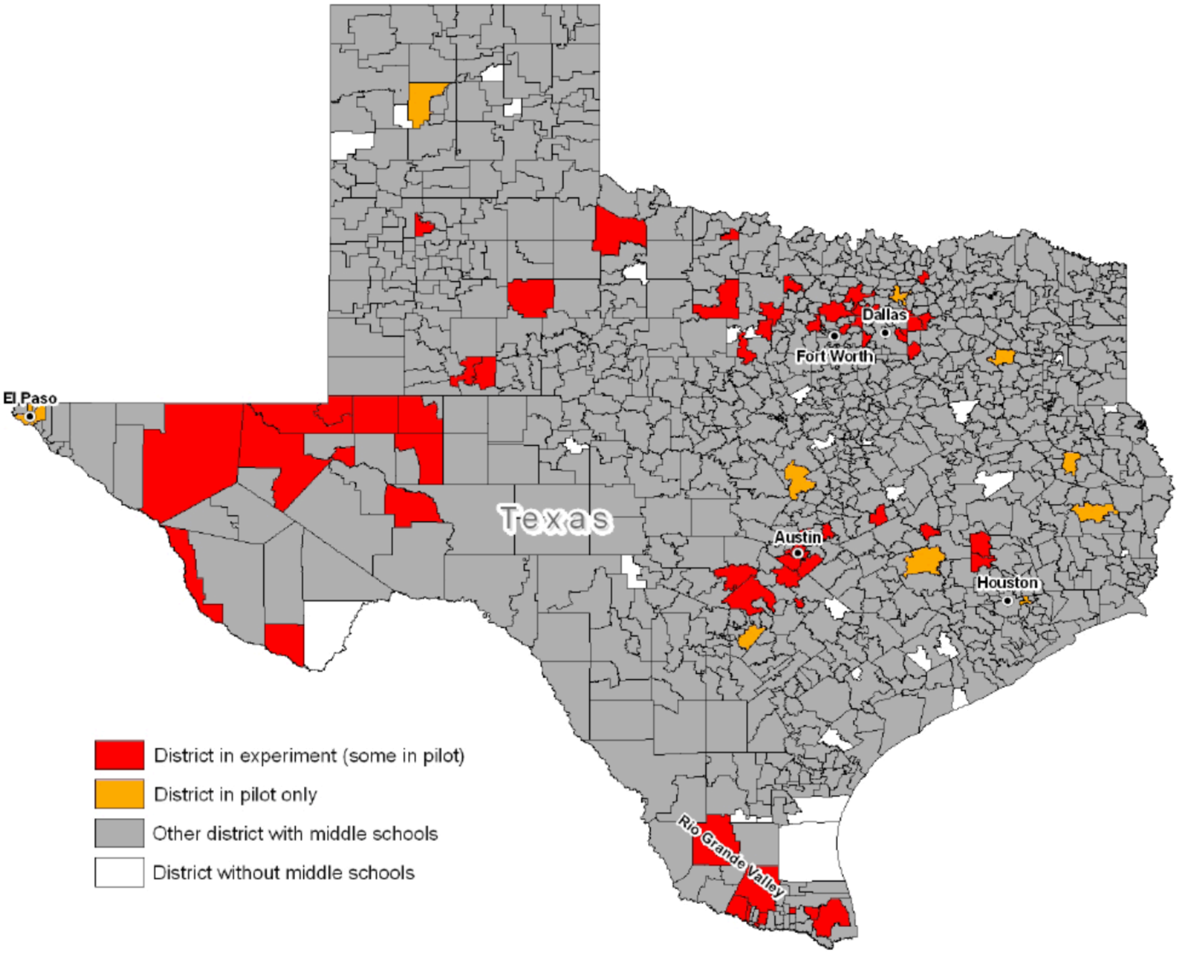
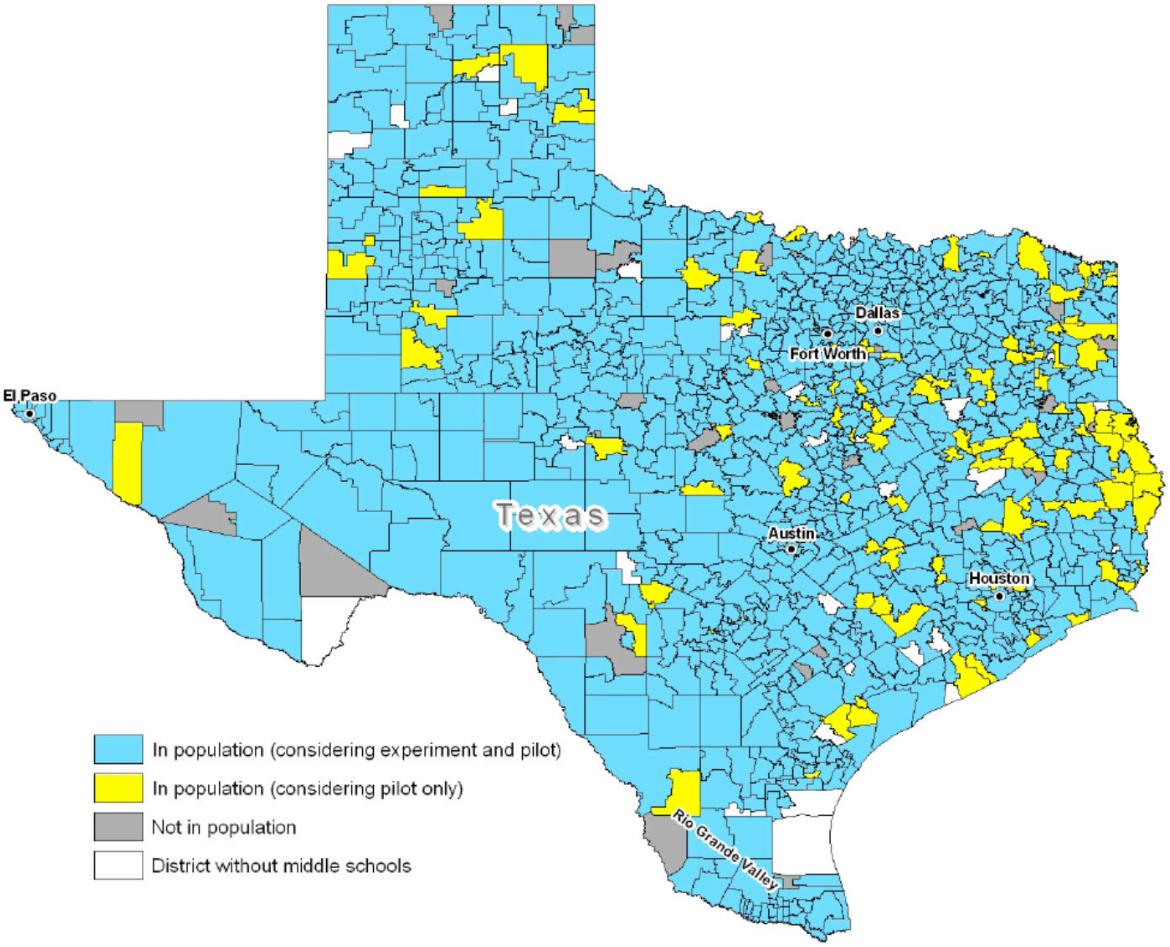

El Paso
Dallas
Fort Worth
Texas
Austin
Houston
Rio Grande Valley
🟥 District in experiment (some in pilot)
🟧 District in pilot only
⬜ Other district with middle schools
□ District without middle schools

Figure 4b. Generalization Regions Where Results Apply: Colored regions show Texas school districts with schools that match the propensity scores of the participating schools.


El Paso
Dallas
Fort Worth
Texas
Austin
Houston
Rio Grande Valley
In population (considering experiment and pilot)
In population (considering pilot only)
Not in population
District without middle schools

3. Extensions with Adaptation to Different Settings

In this section, we consider additional data to shed light on whether the results might generalize to populations and settings outside of Texas. We think of this additional research as an “extension” to the prior experiments, as the intervention was not precisely replicated. Rather it was adapted to suit the new contexts. Each adaptation, however, retained a focus on the same mathematical concepts, the same design in the technology, and on the integration of technological, workbook, and teacher professional development components.

Research found that one level of adaptation occurs as teachers take up the intervention (as previously discussed for Texas case studies above). Indeed, adaptation by teachers to suit their own context is considered important to teacher ownership (Coburn, 2003). The nature of teachers’ adaptations has been thoroughly documented (Hoyles, Noss, Roschelle & Vahey, 2013; Empson et al, 2013) and includes: new introductory activities to transition students into the replacement unit (given their prior learning); varied allocation of activities across full classroom, small group and individual modalities; pace of implementation and further elaboration of classroom lesson plans;

additional teacher-designed discussions, summaries or "consolidations" of the mathematical content.

Adaptations to the intervention as a whole were made by the research partnerships formed to conduct additional implementation research in new settings. One partnership was focused on Florida and another on schools in England. In Florida, the Texas 7th grade unit was the basis of the intervention (Vahey, Roy, & Fueyo, 2013). In England, the Texas 8th grade unit was the basis. The workbooks and technology for Florida and England were based on Texas precedents, and were modified, in consultation with local mathematics education experts, to fit the local contexts while maintaining the same learning goals. In addition, in England, the professional development was substantially adjusted to reflect differences in the strengths of local teachers as well as the national curricular context (Clark-Wilson et al, 2015). The assessments used to measure learning before and after the intervention were very close to those used in Texas, but were lightly adjusted to accommodate linguistic differences (again in consultation with local experts).

The Florida and UK extensions were conducted without local, contemporaneous control groups. From a methodological perspective, this is obviously a weakness. However, the sponsors of this research (a philanthropy in Florida and a different philanthropy in the UK) wanted their funds to go towards use of the materials by large numbers of teachers and students, and did not want to spend money on control groups. The sponsors of the research believed that given a calibrated pretest and posttest, the results could be credibly interpreted by their constituent school decision makers without a control group.

The Florida Study

In Florida, recruitment was led by the mathematics supervisor in one of the largest districts in Florida. Ten schools were invited to participate in the study. Seven of these schools accepted the invitation, and these schools represent the wide variety of schools in the district. Two of the schools were high-poverty (greater than 50% of the students were on the free or reduced lunch program) and the student body of these schools was composed of greater than 50% "minority" students. Two teachers per school were chosen to participate. Due to teacher transfers between the time of recruitment and the beginning of the school year, there were a total of 13 teachers in the study.

This district has a population with more African American students (19%) and substantially fewer Hispanic students (9%) than in the Texas study (4% and 44% in the Treatment group, respectively). The student population in the Florida study closely matched the student population in the district at large generally (see Tables 3 and 4). To quantify the variation in prior achievement in the population, we used the Florida Comprehensive Achievement Test (FCAT). Levels on FCAT range from Level 1 (lowest) to Level 5 (highest). Level 3 indicates that a student's performance is on grade level.

To analyze results in Florida, we conducted a quasi-experimental comparison of the Florida treatment group (there was no Florida control group) to the Texas control group. As Figure 5 shows, learning gains from pretest to posttest were strong in the Florida classrooms using the treatment (ES=1.09). We further analyzed the data by developing a hierarchical linear model. This model (see Table 5) found the posttest scores in Florida classrooms were significantly higher than one would predict on the basis of control group learning gains found in prior research in Texas. One caution in interpreting this data is that the Florida classrooms had a higher pretest score than seen in Texas and therefore it could be that Florida students were better prepared to learn this material from any resource, not just the treatment. Regardless of

Table 3. Student ethnicity in the Florida district and the replication study

Ethnicity	In the Florida district, overall	In the Florida district, in the sample
Caucasian/White	63%	48%
African American/Black	19%	18%
Hispanic/Latino	9%	10%
Asian/Pacific Islander	4%	5%
Native American	<1%	1%
Multiracial	5%	7%
Missing data	--	10%

Table 4. Student 7th grade FCAT scores in the Florida and the replication study.

6th grade mathematics FCAT level	In the Florida district	In the Florida district, in the sample
Level 1	24%	19%
Level 2	21%	26%
Level 3	28%	31%
Level 4	19%	20%
Level 5	8%	5%

Figure 5. Comparison of Texas and Florida Distributions in pretest and posttest

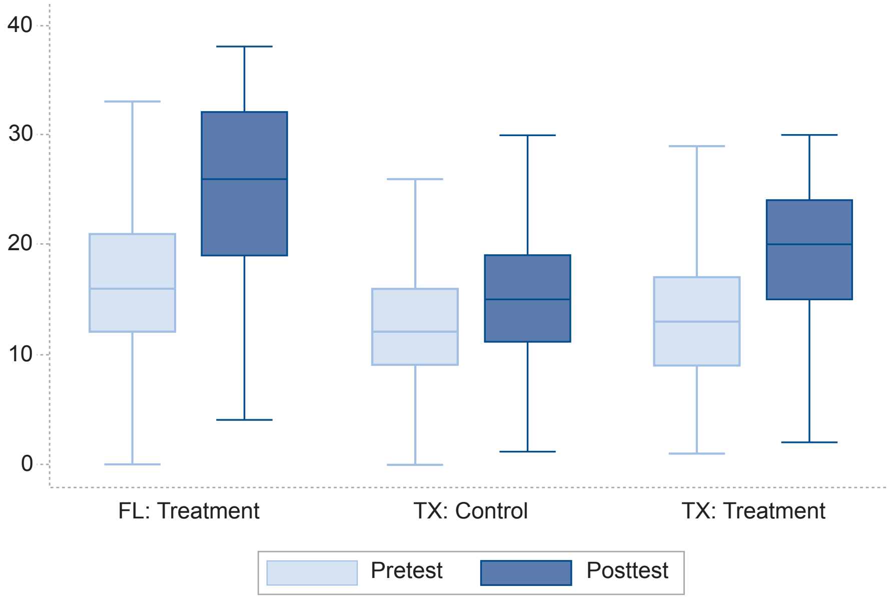


Table 5. HLM model comparing 7th Grade Florida or 8th Grade UK treatment to Texas Control.

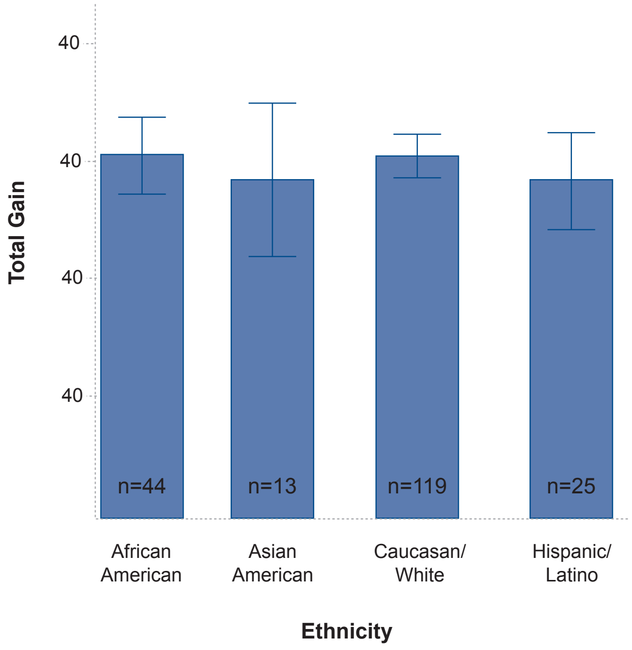
Fixed Effects					
	Predictor	Coeff.	SE	Z Ratio	P
7th Grade model					
Florida (T) Texas (C) Contrast	Intercept	15.59	0.33	47.40	0.000
	Centered Pretest	0.80	0.02	30.14	0.000
	FL/TX Contrast	6.96	0.70	9.91	0.000
8th Grade Model					
UK (T) Texas (C) Contrast	Intercept	14.28	0.67	21.23	0.000
	Centered Pretest	0.80	0.03	24.23	0.000
	UK/TX Contrast	4.62	1.00	4.64	0.000
Random Effects					
				(95% Conf. Interval)	
		Variance	SE	Upper Bound	Lower Bound
7th Grade model					
Florida (T) Texas (C) Contrast	Level 1 (student)	13.69	0.61	12.55	14.94
	Level 2 (teacher)	4.14	0.91	2.69	6.38
8th Grade Model					
UK (T) Texas (C) Contrast	Level 1 (student)	21.71	1.17	19.54	24.12
	Level 2 (teacher)	8.31	2.16	4.99	13.84

whether this resource is uniquely suited to producing high posttest scores among Florida students, the results do indicate that the resource was associated with strong pretest to posttest learning gains in Florida, not just Texas. Likewise, the results increase confidence that the effects seen in Texas were not dependent on factors unique to Texas schools, teachers, or students.

As the Florida population had more African-American students, we disaggregated the data and considered this population specifically. Consistent with achievement data nationwide, the Florida data shows that the mean African American student pretest score is significantly lower than the mean white student pretest score (while the mean for Hispanic students was also lower than Caucasian this was not found to be statistically significant; this may be significant in a larger sample, but our sample of Florida students had relatively few Hispanic students). However, there was no significant difference in mean student gain score across ethnicities (see Figure 6). This finding is again consistent with prior SimCalc studies, which found that the materials are effective across diverse groups of students.

On the basis of this pilot in Florida, the team has since conducted a set of large-scale studies in three other Florida districts, using the Florida state assessment as an outcome measure. Findings will be reported in Sirinides et al (in preparation) and Vahey et al. (in preparation).

Figure 6: Student learning was similar across ethnicities. The error bars, which overlap across all groups, show that the slight difference in gains are not statistically significant.



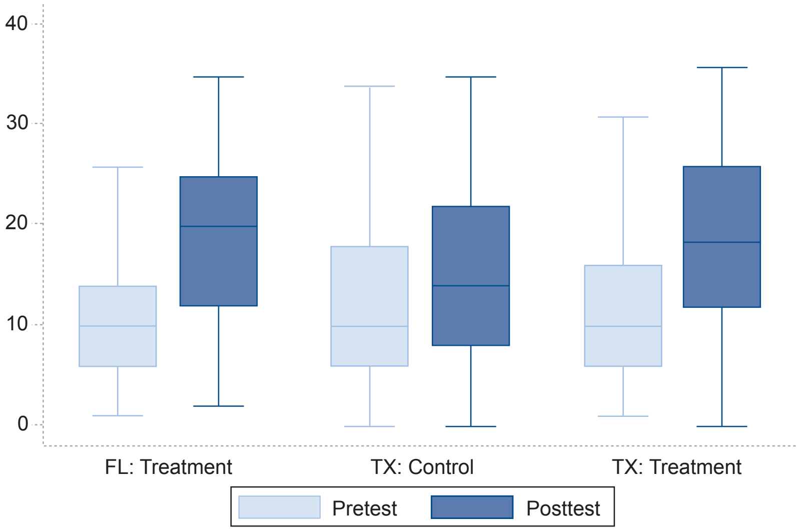
The UK Study

A local UK research team recruited a wide variety of school types within two hours of London, including state schools, Academy schools, and an independent school. Schools represented the wide variety of schools in England, with a range of school-wide achievement as well as students receiving free school meals (FSM). Ten schools participated, with a goal of two teachers per school. One school had only 1 grade-level mathematics teacher, resulting in a total of 19 teachers participating in the professional development and classroom implementation.

Before adapting the materials, team of UK-based researchers as well as two local teacher professional development consultants reviewed the materials in the context of alignment with the English National Curriculum, and decided that the 8th grade Linear Functions unit was the most appropriate for use in their context. These local experts then worked with the original designers to make minor revisions to the materials so that they were adapted to the local context, while ensuring that the mathematics content was not impacted by these superficial changes. Again the procedure for using the materials was similar to the procedure adopted in Texas: teachers attended teacher professional development, and then had access to aligned workbooks and software.

As with the Florida data, we conducted a quasi-experimental comparison of the UK treatment schools to Texas control schools. Figure 7 shows pretest and posttest for three populations of students: UK SimCalc students, Texas Control students (who did not use SimCalc), and Texas SimCalc students. At pretest, all groups had similar mean scores. At posttest, the UK students had gained. We compared the groups via a hierarchical linear model and found that the UK students scored significantly higher than would be

Figure 7. Comparison of Texas and UK Distributions across pretest and postest



predicted without the intervention on the basis of results in the Texas control group (ES=0.57) (see Table 5). Since there was no UK control group, it remains a possibility that UK students would obtained higher posttest scores with another intervention, with no intervention at all, or that the high posttest gains were due to a contemporaneous factor other than the intervention. However, interviews with mathematics education experts and teachers suggest that students in the UK also find learning this material challenging and that strong learning gains on this mathematical content was unusual and noteworthy. Observations in UK classrooms and interviews with teachers suggest the use of the intervention by students and their teachers was plausibly related to the higher posttest scores achieved.

Data on poverty levels of students in the UK sample was available. Consistent with prior achievement data, the UK data shows that students in school with high poverty (as measured by the percent of students receiving free school meals (FSM)) and students in schools with low prior math achievement score lower on the pretest than their counterparts with low poverty and high prior achievement, respectively. However, there was no significant difference in mean student gain score across these demographics. This finding is again consistent with prior SimCalc studies, including the Florida research, which found that the materials are effective across diverse groups of students.

These materials have since been expanded upon and their use continues in England. The research team in the UK have continued to study how these materials, combined with innovative professional development, can impact the teaching of important mathematics (e.g. Clark-Wilson et al., 2015; Clark-Wilson, 2016).

Overall, the data show a very similar pattern of pretest to posttest learning gains in Florida and the UK as in Texas, suggesting that positive effects of the integration of dynamic representations with curriculum workbooks and teacher professional development could be obtained in student populations outside Texas. Further, the Florida data addresses a concern with the Texas data because the Florida sample included a substantial number of African-American students and those students obtained strong learning gains while using the intervention. An important limitation of the Florida and UK studies is that there were no control groups in those studies, and thus, generalization about *causality* based on these data must be cautious. As a practical matter, however, the similar pattern of results beyond Texas enhances confidence in the intervention; indeed, based in part of on these findings, substantial further expansion of the program in both the UK and Florida is underway.

A further benefit of reporting these studies, even though their design is weaker than the original study, is that the studies show the materials can be adapted while maintaining expected pretest to posttest gains. Realistically, many more schools are likely to adopt materials if they can adapt them to better fit their context of use.

Beyond these studies, additional research with the SimCalc software has been conducted in high school settings (Hegedus, Tapper & Dalton, 2014). This further research addresses related but different mathematics, specifically, topics in Algebra 2. As with the Texas, Florida and UK research, the intervention involved an integration of software, workbooks, and teacher professional development. The results from the cluster randomized trial were positive.

Conclusion

The substantive finding across three methods for examining generalizability is that the efficacy of the SimCalc approach for increasing the depth of student' conceptual understanding in middle school mathematics, as found in the Texas experiments, is likely to generalize to broader populations and settings. Research demonstrated that the approach can apply to the rest of Texas, to other states in the United States, and even internationally. This finding is important because educators want to know not only that a statistically significant effect was detected, but also the applicability of the results to varied school settings and populations.

The findings are specific to interventions that use the dynamic representations found in SimCalc in integration with curricular workbooks and teacher professional development, to form a curricular activity system. We expect that the technology alone, or that weakly aligned technology, workbook, and professional development components would not be sufficient to lead to the same treatment effects.

Dynamic representation tools are available for topics of algebra, geometry and data analysis (among others) and both in the form of computer software and cost-effective handheld devices. It would be possible to use these additional tools to address conceptual understanding for a broader range of mathematical topics than considered here. Of course, we do not yet know what the effects of such interventions would be.

From a methods perspective, we note that conducting randomly controlled trials of an educational intervention is an important method of addressing policy questions, such as causal questions regarding how technology in mathematics education to increase student learning. The methods for establishing internal validity of experiments have been well-articulated and reduced to practice, but methods for establishing external validity of findings have not coalesced into a recommended approach.

1. Treatment effect interaction within the experimental sample. This method enabled us to ask "Is the effect different across the varied settings and participants within the existing data set?" and, consequently, how sensitive the intervention might be to variation in the characteristics of the population or setting.

2. Statistical inferences about treatment effects in the larger population. This method enabled us to ask "To what extent do schools in the study sample match schools in the general population, and how might this impact the treatment effect?"

3. Treatment effects in extensions and adaptations in difference settings. This method enabled us to ask "Are similar gains from pre-test to post-test obtainable with participants in very different samples drawn from different populations?" Note, however, that these were not replications, and adaptation to different settings and context were part of the intervention.

A limitation in all the above work is that it used researcher-designed tests to measure outcomes. Initially, researcher-designed tests were deployed because the standardized measures available in Texas at the time were not likely to be sensitive to the learning objective of conceptual understanding. As the work continued in Florida and the UK, a further positive of using the researcher-designed tests became apparent – it allowed continued comparison back to the original study population.

Overall, we conclude that integrating dynamic representations with curricular materials and teacher professional development is a potentially generalizable method for improving students' conceptual understanding and that triangulation among several approaches to examining generalizability can be a reasonable substitute for the impractical technique of probability sampling. We would further urge the field to balance its attention to internal validity and external validity, as both are important to policy audiences.

References

Berliner, D.C. (2002). Comment: Educational research: The hardest science of all. *Educational Researcher, 31*(8), 18-20.

Bohrnstedt, G. W., & Stecher, B. M. (2002). *What we have learned about class size reduction in California [Capstone Report].* Sacramento, CA: California Department of Education.

Bracht, G.H. & Glass, G.V. (1968). The External Validity of Experiments. *American Educational Research Journal, 5*(4), 437-474.

Brenner, M. E., Mayer, R. E., Moseley, B., Brar, T., Duran, R., Reed, B. S., & Webb, D. (1997). Learning by understanding: The role of multiple representations in learning algebra. *American Educational Research Journal, 34*(4), 663-689.

Clark-Wilson, A., Hoyles, C., Noss, R., Vahey, P. & Roschelle, J. (2015). Scaling a technology-based innovation: Windows on the evolution of mathematics teachers' practices. *ZDM Mathematics Education, 47*(1), 79-92.

Clark-Wilson, Alison. (2016). Transforming mathematics teaching with digital technologies – a community of practice approach. In A. Marcus-Quinn & T. Hourigan (Eds.), *Handbook for Digital Learning in K-12 Schools,* pp 45-57. Dordrecht: Springer.

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher, 32*(6), 3–12.

Cook, T. D. & Campbell. D. T (1979). *Quasi-experimentation: Design and analysis issues for field settings.* New York: Houghton-Mifflin.

Cochran, W.G. & Cox, G.M. (2002) *Experimental designs.* New York: Wiley Classics Library.

Cronbach, L. & Snow, R. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Dunn, M. B. (2009). *Investigating variation in teaching with technology-rich interventions: What matters in training and teaching at scale?* Unpublished doctoral dissertation, Rutgers University, New Brunswick, NJ.

Empson, S. B., Greenstein, S., Maldonado, L., & Roschelle, J. (2013). Scaling Up Innovative Mathematics in the Middle Grades: Case Studies of "Good Enough" Enactments. In S. Hegedus & J. Roschelle (Eds.) *The SimCalc Vision and Contributions: Democratizing Access to Important Mathematics, Advances in Mathematics Education Series* (pp. 251 – 270). Heidelberg: Springer.

Fishman, B., Penuel, W. R., Hegedus, S., & Roschelle, J. (2011). What happens when the research ends? Factors related to the sustainability of a technology-infused mathematics curriculum. *Journal of Computers in Mathematics and Science Teaching, 30*(4), 329–353.

Hedges, L. V. (2013). Recommendations for practice: Justifying claims of generalizability. *Educational Psychology Review, 25*(3), 331-337.

Hedges, L. V. & O'Muircheartaigh, C. (2010). *Improving inference for population level treatment effects in social experiments.* Working paper, Institute for Policy Research, Northwestern University.

Hegedus, S., Tapper, J. & Dalton, S. (2014). Exploring how teacher-related factors relate to student achievement in learning advanced algebra in technology-enhanced classrooms. *Journal of Mathematics Teacher Education,* DOI: 10.1007/s10857-014-9292-5

Hegedus, S., Dalton, S., Brookstein, A., Beaton, D., Moniz, R., Fishman, B., & Roschelle, J. (2009). *Scaling up SimCalc project: Diffusion of a research-based innovation in terms of sustainability and spread.* Kaput Center for Research and Innovation in STEM Education, University of Massachusetts, Dartmouth MA.

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65-97). New York: Mcmillan.

Heid, M. K., & Blume, G. W. (Eds.). (2008). *Research on technology and the teaching and learning of mathematics: Research syntheses (Vol. 1).* Charlotte, NC: Information Age Publishing.

Hoyles, C., Noss, R., Vahey, P. & Roschelle, J. (2013). Cornerstone Mathematics: Designing digital technology for teacher adaptation and scaling. *ZDM Mathematics Education, 43*(7), 1057-1070. DOI 10.1007/s11858-013-0540-4

Kaput, J. (1992). Technology and mathematics education. In D. Grouws (Ed.), *A handbook of research on mathematics teaching and learning* (pp. 515-556). New York: Macmillan.

Kaput, J., Hegedus, S., & Lesh, R. (2007). Technology becoming infrastructural in mathematics education. In R. Lesh, E. Hamilton & J. Kaput (Eds.), *Foundations for the future in mathematics education* (pp. 173-192). Mahwah, NJ: Lawrence Erlbaum Associates.

Means, B., & Haertel, G. (2004). *Using technology evaluation to enhance student learning.* New York: Teachers College Press.

Nye, B., Hedges, L.V. & Konstantopolous, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal, 37,* 123-151.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management, 32*(1), 107-121.

O'Muircheartaigh, C., & Hedges, L. V. (2014). *Generalizing from unrepresentative experiments: a stratified propensity score approach.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 63(2), 195-210.

Orr, L. L. (2015). 2014 Rossi award lecture: Beyond internal validity. *Evaluation review, 39*(2), 167-178.

Pierson, J. (2008). *The relationship between patterns of classroom discourse and mathematics learning.* Unpublished doctoral dissertation, University of Texas at Austin.

Roschelle, J., Knudsen, J., & Hegedus, S. (2010). From new technological infrastructures to curricular activity systems: Advanced designs for teaching and learning. In M. J. Jacobson & P. Reimann (Eds.), *Designs for Learning Environments of the Future: International Perspectives from the Learning Sciences*. New York: Springer. 233-262.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J. & Gallagher, L. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal, 47*(4), 833-878.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55. doi: 10.1093/biomet/70.1.41.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20199225.

Shechtman, N., Haertel, G., Roschelle, J., Knudsen, J., & Singleton, C. (2013). Development of student and teacher assessments in the Scaling Up SimCalc Project. In S. J. Hegedus & J. Roschelle (Eds.), *Democratizing access to important mathematics through dynamic representations: Contributions and visions from the SimCalc Research Program.* Dordrecht, Netherlands: Springer.

Sirinides, P., Gray A., Fink, R., Ebby, C., Flack, A., Spillane, M. (in preparation). *The i3 validation of SunBay Digital Mathematics: Impacts and implementation of a technology-based innovation for middle school math.* Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(2), 369-386.

SRI International (2010). *SunBay Digital Mathematics: Pilot Year 1 Final Report Brief: Learning gains in Pinellas County, Florida.* Menlo Park, CA: SRI International.

Tatar, D., Roschelle, J., Knudsen, J., Shechtman, N., Kaput, J., Hopkins, B. (2008). Scaling Up Innovative Technology-Based Math. *Journal of the Learning Sciences, 17*(2), 248-286.

Texas Education Agency (2010). *Academic excellence indicator system reports: 2007-2008.* Retrieved January 10, 2010 from the Texas Education Agency Web site: http:tea.state.tx.us/perfreport/aeis/

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Ruiz de Castilla, V. (2016). Site selection in experiments: An assessment of site recruitment and generalizability in two scale-up studies. *Journal of Research on Educational Effectiveness, 9(*1), 209-228.

Tipton, E. (2014) How generalizable is your experiment? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics, 39*(6): 478 – 501.

Tipton, E. (2013) Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics, 38:* 239-266.

U. S. Department of Education. (2010). *Transforming American education: Learning powered by technology.* Washington DC: U.S. Department of Education.

Vahey, P., Knudsen, J., Rafanan, K. & Lara-Meloy, T. (2013). Curricular Activity Systems Supporting the Use of Dynamic Representations to Foster Students' Deep Understanding of Mathematics. In C. Mouza and N. Lavigne (Eds.). *Emerging Technologies for the Classroom: A learning sciences perspective,* pp. 15 – 30. Springer, NY.

Vahey, P., Roy, G., and Fueyo, V. (2013). Sustainable use of Dynamic Representational Environments: Toward a District-Wide Adoption of SimCalc-based Materials. In S. Hegedus and J. Roschelle (Eds.). *The SimCalc Visions and Contributions: Democratizing Access to Important Mathematics,* pp. 183-202. Springer, NY.

Vahey, P., Wang, S., Lundh, P., Knudsen, J., Kim, H.J.., Lara-Meloy, T., Rafanan, K., and Jackiw, N. (in preparation). *The Design and Implementation of a Dynamic Representation Environment: SunBay Digital Mathematics.*

Sponsor: This report is based on work supported by the National Science Foundation under Grant No. 0437861. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Prime Grantee: SRI International. Center for Technology in Learning

Subgrantees: Virginia Polytechnic Institute and State University; University of Massachusetts, Dartmouth; The University of Texas at Austin; and The Charles A. Dana Center at The University of Texas at Austin

SRI Education
333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000

www.sri.com/education