Computers & Education 76 (2014) 215–224

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Alignment of game design features and state mathematics standards: Do results reflect intentions?☆

Katerina Schenke*, Teomara Rutherford, George Farkas

University of California, Irvine, USA

ARTICLE INFO

Article history:
Received 13 February 2014
Received in revised form
23 March 2014
Accepted 25 March 2014
Available online 3 April 2014

Keywords:
Mathematics education
Learning technology
Standards-based curricula
Assessment

ABSTRACT

This paper describes the results of a randomized control trial of a standards-based mathematics software on elementary school students' (3rd–5th graders; $N = 10{,}860$) mathematics achievement. Spatial Temporal Mathematics (ST Math) engages students by presenting them with a series of game-like activities that are directly tied to the California State Standards for mathematics. We report the effects of the program on students' specific mathematics skills, as well as uncover which elements of the design of the games could be responsible for gains in achievement. We pay particular attention to the alignment of design features of the games, the standards to which developers intended to align their games, and assessments of specific mathematics skills. Results indicate a statistically significant effect of the program on students' basic number sense skills as measured by a standardized measure of mathematics achievement (effect size of 0.14). Subsequent coding of the games for elements that are related to number sense, such as the occurrence of number lines and objects that represent numbers, revealed that these design elements occur throughout the software. We discuss these findings as they relate to the development and design of standards-based mathematics curricula as well as how these features relate to their assessment.

1. Introduction

Computer-supported learning experiences are hypothesized to increase student achievement (Ke, 2008; Kebritchi, Hirumi, & Bai, 2010; Lopez-Morteo & López, 2007). As promising as these technologies may appear, research in the area of computer-supported mathematics instruction has shown mixed results (Dynarski et al., 2007; Ke, 2008; Kebritchi et al., 2010; Slavin & Lake, 2008; Suppes, Liang, Macken, & Flickinger, 2014). Despite these inconsistent findings, mathematics learning technologies still hold promise if they can be designed, implemented, and evaluated in deliberate ways to meet educational goals appropriate and necessary for student mathematics learning. These goals should inform both the content of the curricula and the focus for evaluations of these programs—designers and researchers alike should consider this alignment between goals, content, and outcomes in the creation, implementation, and study of educational technology. The current paper takes this approach in the exploration of specific goals, implementation, and results from the study of one supplementary mathematics software program, Spatial Temporal Mathematics (ST Math), and examines improvement on specific mathematics proficiencies by elementary school students who received ST Math as compared to students who receive mathematics instruction as usual.

1.1. Mixed results of mathematics learning technology

The oft-cited paper by Dynarski et al. (2007) describes disappointing findings for educational technology as shown from the results of a randomized experiment. In a large-scale study of 51 schools, Dynarski and colleagues randomly assigned teachers to utilize different educational technology products or no technology in sixth grade mathematics or Algebra classrooms. A comparison of achievement on the

☆ This research was supported in part by grants from the Institute of Education Sciences (Grant R305A090527). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0808392.

* Corresponding author. School of Education, University of California, Irvine, 3200 Education, Irvine, CA 92697-5500, USA.
E-mail address: kschenke@uci.edu (K. Schenke).

http://dx.doi.org/10.1016/j.compedu.2014.03.019
0360-1315/

Stanford Achievement Test (SAT-10) for sixth graders and the End-of-Course Algebra Assessment for those in Algebra revealed no differences between the treatment and control groups. Similarly, in recent meta-analyses, Slavin and Lake (2008) and Cheung and Slavin (2013) found only small or null effects for mathematics technology in raising student achievement. Nevertheless, there remain individual evaluations presenting large positive effect sizes for mathematics software (e.g., Kebritchi et al., 2010; Page, 2002; Roschelle et al., 2010). Some researchers attribute these mixed findings to the design of the particular intervention or curriculum (e.g., Aleven, Stahl, Schworm, Fischer, & Wallace, 2003; Craig et al., 2013), or the inadequate support provided by teachers to scaffold the use of the technology by students (e.g., Wu & Pedersen, 2011). Looking outside the features of the technology and implementation, Cheung and Slavin (2013) note a variety of criteria that can influence effect sizes: they found that the rigor and design of the study, the sample size, and the measure chosen all influenced the effects seen in evaluations of mathematics technology. Their meta-analysis echoed findings by the National Mathematics Advisory Panel that short, targeted interventions often saw the largest effects (NMAP, 2008, chap. 6). These differences may be explained by one feature common to many modern evaluations of educational technology: the use of broad, standardized tests. Although these assessments are closely tied to policy and may provide useful information about some aspects of a program's effectiveness (see NMAP, 2008, chap. 6), they may be too diffuse to adequately assess the true effects of a program. The view that effect sizes from mathematics technology are "disappointing" can be better understood in light of the metric with which they are measured. The typical effect size benchmarks drawn from Cohen (1988) were broadly derived from effects across the social sciences and may not be applicable to modern educational research, especially interventions measured by existing standardized tests (see Hill, Bloom, Black, & Lipsey, 2008). We offer a few recent examples of studies of educational technology to illustrate the differences that may be brought about by the choice of assessment instrument.

Roschelle et al. (2010) developed an intervention targeting specific mathematics concepts for 7th and 8th graders. Whereas the intervention integrated educational technology, curriculum, and teacher professional development, effects were measured with researcher-developed assessments of specifically targeted mathematics skills (rate, proportionality, and linear functions). Using these researcher-created assessments, Roschelle and colleagues found effect sizes ranging from 0.50 to 0.63 on participants' achievement. Findings from this study might give an inflated view of the true effectiveness of the program for the broad umbrella of mathematics performance. In another example of a study using a narrowly defined assessment, Wilson, Dahaene, Dubois, and Favol (2009) implemented an intervention designed to increase student number sense skills. Wilson and colleagues used narrow measures of number sense such as written and verbal numerical comparisons of numbers, non-symbolic comparison tasks, and verbal counting. Their findings indicated that the intervention had a statistically significant effect on only one of these measures. Both studies indicate statistically significant effects on specific skills but do not explicate the effects of educational technology on broader mathematical competencies, therefore potentially limiting their ability to provide policy-relevant information.

At the same time, other studies have used assessments that are too broad to provide evidence of the evaluated program's effectiveness. For example, an evaluation of a 25-week after school program using the Assessment and Learning in Knowledge Spaces[1] (ALEKS) intelligent tutoring was conducted with a group of sixth grade students (Craig et al., 2013). Craig and colleagues used the Tennessee Comprehensive Assessment Program[2] (TCAP) as their assessment for evaluating learning gains of the two groups (a group using ALEKS and a group with teacher-led instruction). They found that students in the ALEKS condition did not have statistically significantly higher scores on the standardized test than those students in the teacher-led condition. One possible explanation for the lack of findings is that the TCAP was too broad to pick up on the change in skills intended to be improved by ALEKS' designers. Programs that fail to show significant positive effects on broad standardized tests may be dismissed too quickly—although there is value in these assessments, the mathematics learning community acknowledges that they very likely fail to assess all important mathematics skills (see Schoenfeld, 2002). Other evidence can be seen in the research by Hill et al. (2008). They found that interventions measured with broad standardized tests were found to have smaller effects than those measured with narrower measures, more tightly tied to the area of the intervention.

We believe that appropriate assessment instruments should be chosen by reference to the developer's goals and intentions for the program. Many developers are mindful of the pressures placed on districts through government mandates such as in the No Child Left Behind Act (NCLB §1111, 2001[3]) and choose to evaluate their program's effects using the state standardized assessments; however, it is unlikely that all educational technology programs are closely aligned to relevant standards, and therefore such measures may fail to capture otherwise meaningful results. Research from outside of educational technology has brought up the importance of aligning mathematics instruction with valid and meaningful assessment (Krajcik, McNeill, & Reiser, 2008; Martone & Sireci, 2009; Schoenfeld, 2006). Schoenfeld (2002) has suggested that assessment is the key piece to improving mathematics education in the United States, because assessment can help drive the development of better mathematics curricula in the classroom. Moreover, current studies outside of educational technology have lauded efforts to align curriculum and instruction (Krajcik et al., 2008; Martone & Sireci, 2009; Webb, 1997). Our paper extends this work to the field of educational technology and focuses on the importance of alignment between goals, curriculum, and assessment. By way of illustration, we detail the exploration of this alignment in the evaluation of one supplementary mathematics software program for elementary school students, ST Math.

1.2. Context for the current study

The developers of ST Math drew on the California Content Standards for Elementary Mathematics in designing curricular units within the program. Based on documentation provided by the developers, we consider progress under these standards as the goals for ST Math and will discuss the standards, their creation, and prior research on other standards-based curricula in the following section.

[1] ALEKS is web-based tutoring system which adapts to learner behavior. ALEKS has designed curriculum for use in K–12 settings as well as in higher education with courses such as "prep for Precalculus" and "Essential math skills for business". More information can be found at www.aleks.com.

[2] TCAP is the statewide assessment given in the state of Tennessee to measure students' skills and progress in Reading, Language Arts, Mathematics, Science, and Social Studies.

[3] Since then, the Obama administration has passed a bill aimed at greatly narrowing the scope of NCLB and making states, not the federal government, accountable for student achievement (Rich, 2013).