FINAL DRAFT POST-PRINT

Searching for the Two Sigma Advantage: Evaluating Algebra Intelligent Tutors

Kent E. Sabo, Robert K. Atkinson, Angela Barrus, Stacey Joseph and Ray S. Perez

Arizona State University

Office of Naval Research

Citation:

Abstract

This study evaluated 2 off-the-shelf, computer-based, mathematics intelligent-tutoring systems that provide instruction in algebra during a remedial mathematics summer program. The majority of the enrolled high school students failed to pass algebra in the previous semester. Students were randomly assigned in approximately equal proportions to work with the Carnegie Learning Algebra Cognitive Tutor or the ALEKS Algebra Course. Using the tutoring system exclusively, the students completed a 4-hour-a-day, 14-day summer school high school algebra class for credit. The results revealed that both tutoring systems produced statistically and practically meaningful learning gains on measures of arithmetic and algebra knowledge.

Keywords: intelligent tutoring systems, evaluation of CAL systems, applications in subject areas, evaluation methodologies, secondary education.
Searching for the Two Sigma Advantage: Evaluating Algebra Intelligent Tutors

Now more than ever, math skills are fundamental to successful job performance. Although scientific work has always required a high level of mathematical ability, an increasing number of lower level jobs require math skills to operate high-tech equipment (Agondi, Harris, Atkins-Burnett, Heaviside, Novak, & Murphy, 2009). In response to a math achievement gap and the need for math skills in a competitive job market, the No Child Left Behind Act (2001) required schools to make adequate yearly progress in math with the goal that all students meet or exceed proficiency by 2014. Schools across the country are falling far behind this goal. The 2007 National Assessment of Educational Progress showed that many students demonstrate only basic mathematics mastery (Lee, Grigg, & Dion, 2007). Not surprisingly, many curricular approaches are implemented in school math classes, but little rigorous research exists to prove their effectiveness (Slavin & Lake, 2007).

One curricular approach is the use of intelligent tutoring systems in order to leverage advances in artificial intelligence and cognitive science as well as the evolving power of the Internet. Several features differentiate intelligent tutors from more traditional computer-based instruction, including the power to contextually track a student’s performance and carefully adjust the teaching approach based on a student’s learning needs (Woolf, 2009). Intelligent tutors attempt to replicate human one-on-one tutoring which, according to Bloom (1984), enjoys a two sigma advantage over traditional classroom instruction. Numerous studies have concluded that computer-based systems (including intelligent tutoring systems) designed to deliver math instruction and assessments provide positive learning effects, they do not come close to reaching the two sigma advantage that human one-on-one tutoring enjoys (Murphy, Penuel, Means, Korbak, Whaley, & Allen, 2001; Beal, Arroyo, Cohen & Woolf, 2010). Studies of CBI have reported an average effect size of $d = 0.31$ while intelligent tutor studies have reported average effect sizes of $d = .76$ (VanLehn, 2011).
Some researchers have discovered design flaws in many computer-based instruction studies. Few studies employed a randomized, experimental design; many were only descriptive studies, and many lacked relevant data and specificity (Waxman, Lin, & Michko, 2003). Educators and researchers should focus on the effectiveness of algebra curriculum because algebra is a prerequisite for higher-level math and algebra proficiency is correlated with students’ success in college and in obtaining jobs (Adelman, 1999; Carnevale & Desrochers, 2003). However, few independent studies exist that evaluate the effectiveness of algebra intelligent tutoring systems. Many of the algebra intelligent tutoring system studies are conducted by the system developer, which increases the potential for experimenter bias.

**Algebra Intelligent Tutor Studies**

Hannafin and Foshay (2006) evaluated a PLATO Learning's computer-based algebra product as part of a new high school remedial math program. The goal of the larger remedial program was to increase scores on the math portion of the high-stakes state test. The treatment group included 87 students, while 39 students were in the control. The treatment group was scheduled to work with the computer-based system for four of the five instructional days per week. Both the treatment and control made significant gains on the state exam. The mean score for the control was significantly higher than the mean score for the treatment group. However, the treatment group gain scores were significantly higher than the control group's gain scores (Hannafin & Foshay, 2006). This may be a statistical artifact that indicates regression to the mean as the treatment group's scores on the first state exam were significantly lower than the control group's scores.

In an early intelligent tutor study, Koedinger, Anderson, Hadley, and Mark (1997) evaluated a new algebra curriculum called PUMP and an intelligent tutoring system called PAT. Their experimental group included 470 students in 20 algebra classes that worked with the new algebra curriculum and the intelligent tutor. The control group included 120 students
in five algebra classes who received a traditional curriculum and did not use the intelligent tutor. The experimental group worked with the intelligent tutor approximately 25 out of the 180 class meetings, which lasted 44 minutes apiece. The researchers reported a significant difference between the two groups on two standardized tests and two researcher-created tests, with the experimental group performing better on all the tests. The sigma effect on both standardized test was 0.3 and was 0.7 and 1.2 on each researcher-created test. Percent correct on each of the four tests for the experimental group ranged from 52% to 32% (Koedinger et al., 1997).

As part of a large-scale study on reading and mathematics software, Campuzano, Dynarski, Agodini and Rall (2009) evaluated the computer based instructional program Larson Algebra I and the intelligent tutoring system Cognitive Tutor Algebra I. The Larson program was a supplement to traditional instruction while Cognitive Tutor was used as the core algebra I curriculum. Larson students were logged on to the system for an average of 313 minutes per year with the usage occurring over six weeks. Cognitive Tutor students were logged into the system an average of 2,149 minutes a year with usage occurring over 24 weeks. Researchers reported no significant difference between treatment and control groups using computer-based and intelligent tutoring, algebra software products. Overall, the scores for the treatment students on Educational Testing Service's End-of-Course Algebra Assessment was 37.3% correct (Campuzano et al., 2009).

It is important to study intelligent tutors as they attempt to simulate human tutoring and through that simulation, to produce similar learning gains. The reported effect size for human tutoring when compared to classroom teaching is \( d = 2.0 \) (Bloom, 1984), while the effect size of standard computer-based instruction is \( d = 0.31 \) (VanLehn, 2010). If intelligent tutors demonstrate that they produce learning gains close to human tutoring, the implications for education are significant. Our study seeks to add value to the literature investigating the
effectiveness of algebra intelligent tutoring systems through well-designed, independent research. Specifically, we were interested in the effects of algebra intelligent tutoring systems on student learning and attitudes in the intensified and concentrated time period allocated to a summer school session.

Conceptual and Theoretical Frameworks

**Human Tutoring.** Intelligent tutor developers have attempted to replicate human, one-on-one tutoring because it is believed to be the most effective method of instruction (VanLehn, 2011). In his review of intelligent tutoring systems, VanLehn (2011) outlines eight theories of why human, one-on-one tutoring may outperform other types of instruction including traditional computer-based instruction and intelligent tutoring systems. Of the eight theories he reviews, only two are viable explanations: feedback and scaffolding (VanLehn, 2011). Tutors can identify errors in reasoning and help students fix the error in their knowledge during the smallest of steps. Frequent and immediate feedback helps students find their reasoning flaws and fix their knowledge. Scaffolding involves the tutor guiding the student only when the student is working on skills just beyond his or her capability (VanLehn, 2011). Early in learning new skills, the student will require more scaffolding intervention from the tutor, but this guidance should be gradually removed until the student performs a skill independently (Lipscomb, Swanson & West, 2004).

**Computer and Intelligent Tutoring**

VanLehn (2011) identified two types of computer tutors: Computer-Based Instruction (CBI) and Intelligent Tutoring Systems (ITS). CBI provides feedback and hints on student answers. In CBI, a student must solve a problem. He works this out in his head or on paper and then enters the answer. The CBI provides feedback and/or hints based on the student's answer (VanLehn, 2011). The CBI is unaware of any of the student's reasoning or thought processes and thus can be referred to as an answer-based tutor. An ITS provides the student
an interface in which the student enters information for each step of the problem-solving process just as they might if they were solving the problem on paper. The ITS is then able to provide the student feedback and hints based on the analysis of the responses to each step (VanLehn, 2011). This type of ITS is a step-based tutor. A substep-based ITS features a finer-grained level of interactivity by providing feedback and scaffolding on substeps. Substeps are not completed explicitly by students in writing the steps involved in solving a problem but may “correspond to mental inferences” (VanLehn, 2011, p. 12).

Intelligent tutoring systems are architected using information technology systems and student learning models based on advances in cognitive science and artificial intelligence. They have evolved in their ability to customize the learning experience to a student’s ability and simulate the efficiencies of human tutoring since the early 1970s (Corbett, 1998). Currently, there are several commercial software companies and educational research institutions developing intelligent tutoring (Woolf, 2009). A wide-spread adoption of these intelligent tutors in school systems, the military, and other training venues has been slow in the past due to cost and complexity. However, the advances of the Internet now allow intelligent tutoring vendors to provide access to powerful and useable programs via web-based clients anytime, anywhere, at a much lower cost. “The implication is that personalized tutoring can be made widely and inexpensively available just as printed materials and books were made widely and inexpensively available by the invention of the printing press” (Woolf, 2009, p. 20).

**Carnegie Learning’s Adaptive Control of Thought–Rational**

Adaptive control of thought–rational (ACT-R) is a cognitive architecture and theory, the purpose of which is to model the processes of human cognition (Anderson, 1992). The Carnegie Learning Cognitive Tutor is based on ACT-R. In ACT-R theory, procedural knowledge called production rules control human cognition. These production rules take the
form of if-then statements. Learning in this theory is comprised of three learning processes. First, experiences are coded into declarative knowledge called chunks. Second, the chunks are converted into the form of a production rule. Third, the production rules and chunks are strengthened through active repetition (Anderson, 1992).

It is the interaction of the procedural and declarative knowledge that gives rise to complex cognition. A large database of chunks and production rules forms human cognition. Appropriate chunks and rules are selected based on processes that are contextually aware of the environment. Effective human cognition is dependent on the volume of chunks and production rules and the accurate deployment of that knowledge (Anderson, 1996).

In acquiring knowledge, ACT-R assumes that objects in the environment are synthesized and are then available as chunks in the working memory. The theory also assumes that the production rules, which specify the use of the chunks, are encoded by observing the use of the chunks in the environment (Anderson, 1996). For example, students learn to solve math problems by following examples of worked solutions. As defined by the theory, learning is the gradual process of acquiring all the necessary chunks and production rules bit by bit. A process called rational analysis is employed when knowledge needs to be used (Anderson, 1996). This process identifies chunks and production rules that are likely to be needed in a particular context. The contextually appropriate chunks then define the performance in the environment (Anderson, 1996).

ACT-R cognitive tutors are based on the analysis of subject area (e.g., algebra) production rules. By studying the learning rate of these production rules, ACT-R cognitive tutor developers and researchers came to the conclusion that intelligence is explained by the acquisition and tuning of many knowledge units (Anderson, 1996).

ALEKS’ Knowledge Space Theory
Knowledge space theory (KST) attempts to mimic the ability of an expert teacher to assess a student's knowledge state (Doignon & Falmagne, 1985). ALEKS’ intelligent tutoring systems are based on KST. KST is not a theory of human cognition; rather, it is a theory that informed the creation of a computer-based assessment procedure that provides an accurate and continuously updated assessment of student knowledge. KST theorists define a knowledge state as a “particular subset of questions or problems that the subject is capable of solving” (Falmagne, Doignon, Koppen Villano, & Johannesen, 1990, p. 201). A knowledge space for a certain topic is made up of all the knowledge states specific to that topic.

KST posits that given a student response to a problem, whether correct or not, in a certain topic area, inferences can be made about what other questions in the topic area the student could answer. Constructing particular knowledge spaces and their subordinate knowledge states is a labor-intensive undertaking as even a knowledge space defined by tens of questions could result in thousands of possible knowledge states (Doignon & Falmagne, 1985). For example, Falmagne, et al. (1990) asked an expert teacher to evaluate 24 algebra problems. The process involved the teacher evaluating whether given a student fails to answer problem X, would the student also fail to answer problem Y. The teacher responded yes or no to this scenario 196 times to establish all the knowledge sets of the 24-problem knowledge space. To validate the knowledge space defined by the procedure, the knowledge space model is statistically compared to empirical student data (Falmagne, et al. 1990). An assessment system based on a validated knowledge space is then able to assess a student in an efficient manner by using previous student responses to provide future problems. The outcome of a KST assessment is not a numerical score but rather a list of what a student is able to do, represented by the most difficult problem types, and a list of problem types that a student is prepared to learn (Falmagne et al., 2006).

Field-Based Evaluation
Evaluation and research share so many similarities that some researchers and evaluators do not make a distinction between the two. However, there are researchers that have delineated characteristics of evaluations that make them distinct from research. Evaluations focus on the effectiveness of particular interventions and describe the impact of those interventions. Evaluation is a decidedly practical method and is designed to provide potential solutions that allow decisions to be made in practical classroom contexts (Scott & Usher, 2011).

Evaluations are typically conducted in the field, thus providing additional benefits over laboratory-based research. As Cohen, Manion, Morrison, and Morrison (2007) wrote, “Schools and classrooms are not the antiseptic, reductionist, analysed-out or analysable-out world of the laboratory” (p. 278). Since the laboratory is a contrived context, generalizing results from the laboratory to the classroom is inadvisable (Cohen, et al., 2007). It was believed that research conducted by science educators had little impact on classroom practice in the 1980s. One explanation why this occurred was that the research being conducted had such a narrow scope that the results had little application to a real classroom (Brickhouse, 1989). Field-based evaluations also provide opportunities for building connections between schools and universities. The connections can foster valuable collegial networks. Teachers can gain access to curriculum resources previously unavailable due to cost and the students often obtain the benefits of participating in the intervention (Javorsky, Kline & Zentall, 2000).

A specific type of evaluation known as an impact evaluation is used to “assess the effects of a settled program” (Owen, 2006, p. 47) and focuses on what happens to participants as a result of the intervention or program as they typically try to make causal connections between an evaluand and an outcome (Russ-Eft & Preskill, 2009). In order to conduct an impact evaluation of an intelligent tutoring system, an evaluation framework must be
described to outline how the processes and impact will be determined (Russ-Eft & Preskill, 2009). Educational technology field evaluations seek to show that learning outcomes have been achieved and that the software or tool works appropriately, but they usually go beyond laboratory or developer evaluations to demonstrate that the intervention is effective in real classrooms with real students (Woolf, 2009). In order to effectively conduct summative impact evaluations that measure how well intelligent tutoring systems are meeting this goal, a field evaluation framework adapted from Woolf (2009) was used to measure student experiences and achievement using intelligent tutoring systems. The framework included the following processes:

- Establish the goals of the tutor(s)
- Identify the goals of the evaluation
- Develop an evaluation design to test research questions
- Instantiate the evaluation design
- Present results
- Discuss the evaluation results

Using these six stages to design, implement, and document an intelligent tutoring evaluation study increases the validity and rigor of the study and approximates the process of experimental study designs but allows for the fluidity and at times unpredictability of a field-based environment (Woolf, 2009).

Overview of Study

This study was part of larger multi-year research project designed to evaluate the effectiveness of off-the-shelf, algebra tutoring systems. The results of this study were used to inform a series of subsequent studies within the project tasked with determining the most effective intelligent tutoring system for algebra remediation in high school instructional settings and in the military. The two mathematics intelligent tutors selected for evaluation
were the Carnegie Learning Algebra Cognitive Tutor (Carnegie) and the ALEKS Algebra Course (ALEKS). The goal of this field evaluation was to compare the effectiveness of the two intelligent tutors, examine for trends in learning gains, and measure students' experiences with the systems.

**Method**

**Participants and Design**

Participants were 31 remedial high school algebra students (15 male and 16 female, with an average age of 14.7) who chose to enroll in a district-organized 14-day summer school high school algebra class. The majority of these students failed to pass algebra in previous semesters. A total of 41 students enrolled in summer school, with 31 completing the program. English was the first language learned for 77% of the students, and for 97% of the students, English was their preferred language. Students were not enrolled in any other summer courses and did not receive any other formal math instruction other than that provided by the summer school course. Each class meeting was four hours and started at 8AM. Tuition was $175 per session per one-half credit. Three high school math teachers were assigned by the district to supervise the summer school sessions.

The study's design involved a single independent variable with two levels, coupled with pre- and post-tests. The two levels reflected the two tutoring systems used in the evaluation. The participants were randomly assigned in approximately equal numbers to the two tutoring conditions. Three students in the ALEKS condition did not complete the summer school program, resulting in 14 ALEKS students completing the program and 17 Carnegie students completing the program.

**Intelligent Tutors**

This study evaluated two off-the-shelf, intelligent tutoring systems: the ALEKS Algebra Course (ALEKS) and the Carnegie Learning Algebra Cognitive Tutor (Carnegie).
The two systems are comparable intelligent tutors that include similar features. The most important feature, for the purposes of this study, is the ability for both systems to adapt and respond intelligently to student needs based on input from the student. While both systems qualify as intelligent tutors, they do differ in dominant question types, content delivery structure and theoretical bases. Carnegie’s questions are word problems based on real-world scenarios, while ALEKS’ questions are equation-based. Carnegie’s content delivery is linear, while ALEKS allows students to work on any section of the content that the system deems the student is ready to learn. Finally, Carnegie is based on a cognitive architecture and theory that aims to model the processes of human cognition (Anderson, 1992), while ALEKS is based on an assessment theory that attempts to mimic the ability of an expert teacher to assess student knowledge (Doignon & Falmagne, 1985).

**ALEKS Algebra Course.** ALEKS assesses students continuously by using artificial intelligence with questions that require a free response. ALEKS structures questions and requests input from the student such that it makes solving a problem similar to what a student would do on paper. As the first activity for a new student, ALEKS administers a 20-30 question assessment. ALEKS makes a decision on each question delivered based on the student responses to the previous questions. At the completion of the initial assessment, ALEKS determines the topics the student has and has not mastered. ALEKS represents this topic mastery information with a pie chart of topics available in the course.

This pie chart displays the topics ALEKS determines the student is ready to learn based on the initial assessment. Students have the freedom to choose the topic they wish to practice. Once a student chooses a topic, he or she begins working on practice problems. When a student consistently answers the practice problems from a certain topic correctly, ALEKS determines that the student has mastered the topic. ALEKS updates the student’s pie chart and the student can choose the next topic to practice. While practicing, students have
access to explanations of the problems within ALEKS. ALEKS administers assessments periodically throughout the course to test previously mastered topics. If a student no longer demonstrates mastery of a topic, that topic returns to the list of available topics in the student’s pie chart.

**Carnegie Learning Algebra Cognitive Tutor.** Carnegie allows instructors to build a custom curriculum by choosing and excluding topics to meet student, school, and district needs and requirements. Carnegie administers an assessment before and after every unit. Instructors can configure the pretest so that the results of the pretest are used to set pacing for the unit. Students work through the unit in a linear fashion. Carnegie also assesses students continuously by analyzing student responses to problems that require a free response. Like ALEKS, Carnegie attempts to structure questions and provide response options that mimic the problem-solving steps a student would engage in on paper.

Problems are presented with multiple representations and are based on real-world situations. For example a word problem can be displayed along with a graph and a data table. Students can access interactive examples and hints for problems. An interactive example displays the steps required to solve a certain type of problem. Hints provide information to help the student proceed through the steps of a problem. Carnegie requires mastery of a topic within a unit before a student is allowed to progress to the following topic. At the completions of a unit, students are assessed and are allowed to move on to the next unit in the sequence prepared by the instructor.

**Instruments**

**Accuplacer.** The Accuplacer is a computer-adaptive placement test featuring 10 modules, three of which are designed to measure mathematics skills and knowledge. Choosing an appropriate assessment was crucial to evaluating these intelligent tutors. Previous studies have used state high stakes tests (Hannafin & Foshay, 2006), commercially
available standardized tests (Campuzano et al., 2009) and a combination of standardized tests and research-created tests (Koedinger et al., 1997). Accuplacer has characteristics that make it particularly well suited to assess learning gains in an intelligent-tutor context. Primarily, the Accuplacer modules we chose were generally aligned with the content of the two tutoring systems. We say generally aligned because as each student interacts with the systems, the systems deliver content based on the students’ needs. Even within the same tutoring system, no student received exactly the same content or sequence of content. Within a context where students are potentially interacting with and learning different content at different times, a static assessment may not be effective. Thus, Accuplacer’s adaptability was a secondary reason for its selection. The adaptability of the assessment may also reduce potential practice effects from our repeated administrations. Finally, although research-created assessments obtain larger effect sizes (Koedinger et al., 1997), professionally-designed assessments likely provide stronger psychometric properties.

Two modules were used for this study: arithmetic and elementary algebra. Both the arithmetic and elementary algebra modules featured multiple-choice questions. Fundamental arithmetic concepts were measured in the 17-item arithmetic module. Students were tested on basic arithmetic operations (e.g., “Solve the problem. Use the paper you were given for scratch work. A soccer team played 160 games and won 65 percent of them. How many games did it win? A. 94, B. 104, C. 114, D. 124” (College Board, 2011). Questions were divided into three types: (a) “operations with whole numbers and fractions: topics included were addition, subtraction, multiplication, division, recognizing equivalent fractions and mixed numbers, and estimating; (b) operations with decimals and percents: topics included addition, subtraction, multiplication, and division with decimals (percent problems, recognition of decimals, fraction and percent equivalencies, and problems involving estimation were also given); and (c) applications and problem solving: topics included rate,
percent, and measurement problems, simple geometry problems, and distribution of a quantity into its fractional parts” (College Board, 2011).

The 12-item elementary algebra module measured students’ ability to perform basic algebra operations (e.g., “Solve the problem. Use the paper you were given for scratch work. If $2x – 3(x +4) = -5$, then $x = A. 7, B. –7, C. 17, D. -17”$ (College Board, 2011). The module contained three types of questions. The first type of question was on operations with integers and rational numbers, absolute values, and ordering. The second type of question was on operations with algebraic expressions, and monomial and polynomial addition and subtraction. Questions required multiplication and division of monomials and polynomials, positive rational root and exponent evaluation, simplification of algebraic fractions, and factoring. The third type of question required solving equations, inequalities, word problems and linear equations and inequalities. This type also required quadratic equations by factoring, algebraically contextual verbal problems, and translating written descriptions into algebraic expressions (College Board, 2011).

**User experience questionnaire.** Upon completion of the summer school algebra program, students completed a user background and experience questionnaire designed to gather demographic information (e.g., What language do you prefer to communicate in?), comfort level with computers (e.g., Rate your comfort level for locating and opening programs), and their experiences with the intelligent tutors (e.g., My software program helped me understand the math content) and summer school (e.g., I learned a lot from the summer school math class). Students responded to the comfort level questions by selecting from a Likert-type scale with five options ranging from very uncomfortable to very comfortable. They responded to the experience questions by selecting from a Likert-type scale with seven options ranging from not true at all to very true.

**Procedure**
Students completed the 14-day summer school session in two computer labs within the same high school, equipped with PC workstations connected to the Internet. The software programs used in this evaluation were delivered through the Internet using subscription, web-based software clients that do not require individual installations on the computer workstations. Participant and parental consent was obtained through signed assent and consent forms from both the participant and the participant's parents, respectively.

On the first day of the summer school session, each student was randomly assigned to either the ALEKS or Carnegie intelligent tutoring system and given a study booklet that outlined the purpose of the study, the student’s assigned software and directions for its use, and a detailed study protocol. The booklets also included a demographic questionnaire. Students were then given a brief verbal overview of the summer school course and the evaluation study. Students also completed the Accuplacer pretest modules on Day 1. On Day 2, students began working in their assigned intelligent tutoring system and practiced algebra skills and concepts using their tutoring system for the next 10 days for four hours per day. On a typical day, students would arrive at 8AM, log into their assigned tutor and begin working on the algebra content. Along with the system's help and hint functions, the supervising math teachers were available to the students to answer their algebra questions. Students were given regular breaks based on the district’s standard summer school schedule. On Day 7, students completed the adaptive arithmetic reasoning and elementary algebra Accuplacer math tests as a repeated measure assessment intended to capture iterative learning gains. The final two days of the 14-day program consisted of the students completing the Accuplacer module post tests and a demographic and experience questionnaire, as well as completing the district final exam that determined each student’s readiness to progress to the next math course in the high school sequence. By the conclusion of the 14-day summer school course, each student had practiced and studied using his or her assigned intelligent
tutoring system for an average of 35.5 hours. Students completed periodic assessments
provided and used by the intelligent tutoring systems to regularly adapt the practice, pace,
and content to the individual student’s learning needs. The summer school teachers used
these periodic tutor assessments as well as student progress in the tutor curriculum to assign
students their grades for the course.

Results

This study utilized a mixed within- and between-subjects design to evaluate the effect
of the intelligent tutors on algebra and arithmetic test scores. The between-subjects factor was
algebra tutoring groups with two levels (CL and AK) and the within-subjects was Accuplacer
administration time with three levels (Day 1, Day 7 and Day 13). Two separate, two-way
mixed ANOVAs were conducted to evaluate the effect of the two math programs on the
Accuplacer tests.

To evaluate the effect of the tutors on the Accuplacer algebra test, the between-
subjects factor was algebra tutoring with two levels (CL and AK) and the within-subjects was
Accuplacer algebra administration time with three levels (Pretest on Day 1, Day 7, and
Posttest on Day 13). For the dependent measure Accuplacer algebra score, results indicated a
significant effect for time, Wilks’ Λ = .52, F(2, 28) = 12.66, p < .0001, multivariate η² = .48.
The strength of the relationship between the intelligent tutor treatment and the Accuplacer
algebra score was strong, with the intelligent tutor factor accounting for 48% of the variance
in Accuplacer algebra scores. A nonsignificant time X tutor interaction effect was found,
Wilks’ Λ = .99, F(2, 28) = .15, p = .96, multivariate η² = .003. This result suggests that there
were no statistically significant differences between the tutors on Accuplacer algebra scores
over time. Follow-up analyses indicated that at each point in time the Accuplacer algebra
scores differed significantly from one another and increased overtime. Students made
significant gains on the Accuplacer algebra subtest from Day 1 to Day 13.
Table 1

Means and Standard Deviations for Algebra Repeated Measure

<table>
<thead>
<tr>
<th>Tutor</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK(^a)</td>
<td>42.43</td>
<td>18.09</td>
</tr>
<tr>
<td>CL(^b)</td>
<td>27.06</td>
<td>9.86</td>
</tr>
<tr>
<td>Total(^c)</td>
<td>34.00</td>
<td>15.94</td>
</tr>
<tr>
<td>Time 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK</td>
<td>51.00</td>
<td>26.36</td>
</tr>
<tr>
<td>CL</td>
<td>37.94</td>
<td>16.70</td>
</tr>
<tr>
<td>Total</td>
<td>43.84</td>
<td>22.21</td>
</tr>
<tr>
<td>Time 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK</td>
<td>57.71</td>
<td>25.12</td>
</tr>
<tr>
<td>CL</td>
<td>43.65</td>
<td>15.31</td>
</tr>
<tr>
<td>Total</td>
<td>50.00</td>
<td>21.19</td>
</tr>
</tbody>
</table>

Note: \(^a\)n = 15, \(^b\)n = 17, \(^c\)n = 31

Independent samples t-tests were conducted to evaluate the difference in means on the algebra pre-test and post-test between the tutor groups. There was a significant a priori difference in the algebra pre-test scores between the two conditions, t(29) = 3.01, p = .005. There was no significant difference in the algebra post-test scores between the two conditions, t(29) = 1.92, p = .065 (See Figure 1). ALEKS students saw an algebra mean gain of 15.29 while Carnegie students saw an algebra mean gain of 16.59. Combined algebra scores showed an effect size of \(d = .95\) while ALEKS alone showed an effect size of \(d = .95\) and Carnegie showed an effect size of \(d = 1.18\).
Mean Algebra Repeated Measure Scores for ALEKS and Carnegie

To evaluate the effect of the software programs on the Accuplacer arithmetic reasoning test, the between-subjects factor was algebra tutoring condition with two levels (CL and AK) and the within-subjects was Accuplacer arithmetic administration time with three levels (Pretest on Day 1, Day 7, and Posttest on Day 13). For the dependent measure, Accuplacer arithmetic score, results indicated a significant effect for time, Wilks’ $\Lambda = .45$, $F(2, 28) = 17.26$, $p < .0001$, multivariate $\eta^2 = .55$. The strength of the relationship between the intelligent tutor treatment and the Accuplacer arithmetic score was strong, with the intelligent tutor factor accounting for 55% of the variance in Accuplacer arithmetic scores. A nonsignificant time X tutor interaction effect was found, Wilks’ $\Lambda = .98$, $F(2, 28) = .24$, $p = .79$, multivariate $\eta^2 = .02$. This result suggests that there were no statistically significant differences between the tutors on Accuplacer arithmetic scores over time. Follow-up analyses indicated that at each point in time the Accuplacer arithmetic scores differed significantly from one another and increased overtime. Students made significant gains on the Accuplacer arithmetic reasoning subtest from Day 1 to Day 13.

Table 2

Means and Standard Deviations for Arithmetic Repeated Measure
Independent samples t-tests were conducted to evaluate the difference in means on the arithmetic pre-test and post-test between the tutor groups. There was a significant a priori difference in the arithmetic pre-test scores between the two conditions, \( t(29) = 2.40, p = .023 \). There was no significant difference in the arithmetic post-test scores between the two conditions, \( t(29) = 1.48, p = .15 \) (See Figure 2). ALEKS students saw an arithmetic mean gain of 18.92 while Carnegie students saw an arithmetic mean gain of 23.88. Combined arithmetic scores showed an effect size of \( d = 1.10 \) while ALEKS alone showed an effect size of \( d = 1.12 \) and Carnegie showed an effect size of \( d = 1.25 \).

**Figure 2**

*Mean Arithmetic Repeated Measure Scores for ALEKS and Carnegie*
An independent t-test was conducted to evaluate the differences in composite mean scores on the experience questionnaire between the tutor groups. Students rated statements from 1 (Not true at all) to 7 (Very true). There was no significant difference between ALEKS (M = 4.86, SD = 1.09) and Carnegie (M = 4.63, SD = .89), t(25) = .62, p = .54. Because there was no significant difference between the groups, the descriptive statistics from the experience questionnaire are reported in aggregate. Just over half of students (51.6%) responded true to very true that their tutor helped them understand the math content. A majority of the students (61.3%) responded true to very true that they learned a lot in summer school. Only a quarter (25.8%) of students responded true to very true that they enjoyed the activity. Over half of students (54.9%) responded true to very true that they felt competent after working on the activity for a while. Only 38.7% responded true to very true that they were satisfied with their performance. Nearly three quarters (74.2%) of the students responded true to very true that it was important for them to do well on this activity. A majority (61.3%) of students responded true to very true that this activity could be beneficial to them. However, only 51.6% responded true to very true that they would recommend this math class to their classmates.

**Discussion**

This study demonstrates that classroom teachers can implement intelligent tutoring systems to provide an effective learning environment that produces significant learning gains. Both Accuplacer arithmetic and algebra test scores increased significantly over time for students in the two tutoring conditions. In addition, the pre-test to post-test effect sizes for the two tutoring conditions were very large. Interestingly, there was no significant difference for the interaction between tutor conditions across the administration times. These results suggest that both ALEKS and Carnegie are very effective curriculum replacements when implemented in an intensive, short-duration summer school program.
In his meta-analysis, VanLehn (2011) found that intelligent tutor research did not study implementations in which intelligent tutors were used as full curriculum replacement. Since there are many instructional activities that can have an effect on student learning, these studies make it more difficult to attribute the learning gains to only the intelligent tutor (VanLehn, 2011). Our summer school implementation replaced completely the algebra curriculum. Although this study did not compare intelligent tutors to classroom instruction, much of the positive changes in learning over time can be attributed to the intelligent tutors.

The Two Sigma Advantage

According to Bloom (1984), and believed by many intelligent tutor developers and researchers, human tutoring has a two sigma advantage over traditional classroom instruction. Because intelligent tutors are developed to simulate human tutoring, we took our first steps in the search for the same exceptional gains reportedly provided by human tutoring. While the effect sizes we observed were impressive on their own, they did not come close to our a priori expectations. Considering that our study compared pre-test scores to post-test scores rather than comparing differences in scores from a traditional classroom condition to an intelligent tutoring condition, the two sigma gain appears to be an unrealistic goal. Indeed, previous studies evaluating both traditional computer-based instruction and intelligent tutoring systems have failed to find this elusive two sigma gain. VanLehn (2011) suggests that Bloom's assertion inspired the research and development of modern intelligent tutoring systems. However, only two studies of one-on-one tutoring found a two sigma advantage over classroom instruction and the results of these studies can be called into question (VanLehn, 2011). In the Bloom (1984) study there were three conditions: traditional classroom instruction, classroom mastery learning and one-on-one tutoring with mastery learning. The mastery learning students performed better than the traditional classroom students \((d = 1.00)\) and the tutoring students performed better than the mastery learning students \((d = 2.0)\).
However, the mastery level required for the mastery classroom students and the tutoring mastery students were different; 80% mastery versus 90% mastery. The higher mastery threshold for the tutoring students could explain the two sigma advantage (VanLehn, 2011). Evans and Michael (2006) saw a similar two sigma difference in an experiment with a small sample size (N =17), but failed to see as large an advantage (d = .52) in a follow-up experiment with a larger sample size (N = 53). The effect size of their initial experiment was likely due to sampling error.

While the instructional effectiveness of the tutors is of primary importance, student attitudes about their instructional experience are also critical. The results from our student experience questionnaire showed that there was no significant difference between student responses from the two tutor groups. As neither tutor demonstrated a statistically significant advantage in instructional effectiveness, similarly, neither showed a significant advantage in student preference. Results from the experience questionnaire suggest that student attitudes ranged from negative to ambivalent to positive. Unsurprisingly, students did not enjoy summer school but had more positive attitudes about the tutoring systems and the benefits of summer school. Students were ambivalent about their respective tutoring systems and their competence after working in the tutors for a while. Students were equally ambivalent about recommending the summer school program to their classmates. However, many students responded that it was important for them to do well and that learning algebra was beneficial. This suggests students had a desire to learn and understood the relevance of the content. The effect that this desire and understanding may have had on test scores should not be underestimated.

Conclusion

Although Bloom’s two-sigma gain may be unattainable by current intelligent tutoring systems, we intend to continue our investigations in different school contexts with additional
technologies and instruments. Indeed, future studies are needed. Despite the raw number of computer-based math instruction studies and to a lesser extent, math intelligent-tutor studies: 1) few are peer reviewed, 2) many exhibit design deficiencies, 3) most focus on elementary curricula, and 4) recent technologies are not studied (Murphy et al., 2001; Waxman et al., 2003). In our future research, we intend to further address these issues in the computer-based math instruction field, especially as the issues pertain to intelligent tutors.

It is important, as an unbiased third party, to continue our field-based research that studies the learning outcomes that result from use of intelligent tutors. Many intelligent tutoring studies are conducted by the developers of the systems and the companies that market the systems. As such, few of these studies are scrutinized by the peer-review process. The value of peer-reviewed studies produced by those outside the companies should not be underestimated. As academic third-party researchers, we will continue to submit our studies to journals that engage in the rigorous peer-review process. Further, intelligent tutoring studies are often performed in a laboratory or a classroom laboratory-type environment. Our current and future studies have and will continue to take place in real schools and classrooms, with real teachers and students who are engaged in the business of teaching and learning. We believe that studying intelligent tutoring systems in the field as they would be implemented by schools provides additional external validity to our results. Within this context, we will continue to focus on high school algebra content, as success in algebra is the gateway to higher-level mathematics topics. Algebra proficiency also has a strong relationship with college success and in securing a job.

The goal of intelligent tutoring is to produce learning outcomes that far surpass the learning outcomes of traditional instruction. Accordingly, intelligent tutoring research should seek to compare learning outcomes produced by engaging in traditional or business-as-usual instruction and intelligent tutoring. We recognize that the lack of this comparison in our
current study is a limitation. Although our study reported large and significant learning gains from a pre-test to a post-test, we are not able to report that the gains were any better than gains from traditional instruction. A future study will feature a traditional algebra classroom condition that will serve as a control group. In addition to the control group, we will add another answer-based tutor to compare to the tutors we evaluated in this study. The purpose for adding another tutor to our future evaluation is to provide evidence that allows administrators and educators to make informed decisions based on both dollar value and learning outcomes. With this quasi-experimental design, a larger sample size, and the tutoring systems implemented as full curriculum replacement, our results will present a clear picture of the effectiveness of intelligent tutoring systems.

Aside from the summer school context, we plan to examine intelligent tutoring system implementations within constraints of a standard school-year schedule. We intend to study the systems’ use in regular math classrooms as a once-a-week supplement to the traditional curriculum during the standard school day. We will compare classrooms that supplement with intelligent tutors with classrooms that do not. In a third implementation, we will investigate the use of the tutors in an after-school program.

Beyond learning outcomes, we intend to measure student motivation produced by using the tutoring systems as well as students’ metacognitive strategies. Although we’ve found that the tutoring systems can produce large and significant learning gains, it is also important to investigate student strategies and whether they are motivated to use the systems. We hypothesize that students who exhibit higher motivation while using the tutoring systems will perform better on the mathematics and algebra assessments. We also believe that students with more sophisticated metacognitive strategies and self-regulation will perform better on the assessments. Our current and future studies will make for a strong body of work that will produce robust results and provide researchers and educators important information
on the effectiveness of intelligent tutors in a variety of classroom contexts, both as curriculum replacement or as supplemental curriculum.
References


