

# Effect of Adaptivity on Learning Outcomes in an Online Intervention for Rational Number Tutoring, “Woot Math,” for Grades 3-6: A Multi-Site Randomized Controlled Trial

R. Brent Milne,<sup>1</sup> Sean A. Kelly,<sup>1</sup> David C. Webb<sup>2</sup>

Personalized learning through adaptive software has widely been identified as a critical enabling technology for education in the twenty-first century (see, e.g.: The Council of Economic Advisors, 2011; Project Tomorrow, 2011). The position that we take on this subject is one that has long been held, that while technology will not replace the teacher, there are great opportunities for adaptive technology to supplement the teacher in beneficial ways:

Will machines replace teachers? On the contrary, they are capital equipment to be used by teachers to save time and labor. In assigning certain mechanizable functions to machines, the teacher emerges in his proper role as an indispensable human being... The role of the teacher may well be changed, for machine instruction will affect several traditional practices. Students may continue to be grouped in grades or classes, but it will be possible for each to proceed at his own level, advancing as rapidly as he can.

The fascinating thing about the above quote is when it was written – by B. F. Skinner in a 1958 issue of *Science* (Skinner, 1958, p. 976). The topic of technology-personalized learning has a surprisingly long history, in fact stretching back much further than the 1950’s. It is a history that has generated periods with great hype and bold projections, interspersed with letdowns and frequent controversy and critique (Benjamin, 1988; Meyer, 2014; Ferster, 2014).

One addressable concern and criticism has been the lack of clear empirical data on the potential benefits of adaptive technology. In a 2013 review of educational technology, the Fraser Institute wrote that (Izumi, Fathers, & Clemens, 2013):

Although informative in terms of how adaptive technology in education has evolved and developed, the review of research undertaken for this paper indicates a vast gap in sound, empirical research to determine and quantify the potential benefits from the adoption of such technology in education. (p. 7)

In the study reported here, which was supported by a grant from the National Science Foundation, we provide sound empirical research that isolates and measures the benefits of just the adaptive aspects of an online intervention for mathematics. We have done so through a well-constructed, triple-blind, randomized controlled trial in which both the test group and the control group were treated using the same online intervention and the same content, with the only difference being that the experimental adaptive capabilities were enabled for the test group but disabled for the control group (i.e., the potential measured benefits in this study can only be attributed to the adaptive technology for personalized learning).

This level of specificity is important because a recent meta-analysis of the effectiveness of “Intelligent Tutoring Systems (ITS)” found only that “overall, ITS had no negative and perhaps a

---

<sup>1</sup> Woot Math, LLC

<sup>2</sup> University of Colorado Boulder, School of Education; Executive Director of the Freudenthal Institute US for science and mathematics education

small positive effect on K-12 students' mathematical learning, as indicated by the average effect sizes ranging from  $g = 0.01$  to  $g = 0.09$ " (Steenbergen-Hu & Cooper, 2013, p. 970). Such small effect sizes are insignificant in the context of meaningful impact on education (Hattie, 2008, pp. 1, 15-17). Moreover, Steenbergen-Hu and Cooper (2013, p. 985) go on to report evidence that existing ITS systems "might have contributed to [widened] achievement gaps between higher and lower achieving students." For these systems to have a meaningful impact in education outcomes and to narrow achievement gaps, it is clear that adaptivity needs to be more carefully studied and honed for effectiveness. As Shute and Zapata-Rivera (2012) observe:

One shift that we see as critically important to the field [computer aided instruction], particularly in the near term, is toward conducting controlled evaluations of adaptive technologies and systems. This will enable the community to gauge the value-added of these often expensive technologies in relation to improving student learning or other valued proficiencies (e.g., self esteem and motivation). Our review has shed light on a range of technologies, but the bottom line has not yet been addressed: What works, for whom, and under which conditions and contexts? (p. 22)

## Executive Summary

We study several forms of software-based adaptivity and demonstrate that they contribute significant effects in learning and retention. While studied in the context of rational number instruction in grades 3-8, this is a result that should generalize to a wide range of software-based education interventions.

In the randomized controlled trial (RCT) we report, both the test and control group used an online intervention for rational number tutoring, "Woot Math." Woot Math is an adaptive learning environment for mathematics, which focuses on helping students in grades 3-8 master core math concepts, beginning with rational numbers. The supplemental software delivers a personalized progression of interleaved video instruction and scaffolded problems to mimic the natural give and take between a student and a tutor. In this trial, the only difference being tested was that the experimental adaptive capabilities were enabled for the test group but disabled for the control group.

The study results show that the adaptive version delivered significantly better learning and retention, with moderate to large effect sizes ( $g = 0.23$  to  $g = 1.50$ ,  $p < .05$ ) in a sample of 350 students completing grades 3-6.<sup>3</sup> Individual students were randomized into the test and control groups in equal numbers, and the trial was fully blinded – students, teachers, and investigators were not aware as to which group each student participant had been assigned. The study was conducted at multiple, demographically diverse sites, and attrition was low (9%). The measurement scale used in the primary outcome analysis was observed to be highly reliable (with  $\alpha = .80$ ,  $\omega = .82$ , and the  $g_{lb} = .87$ ).

Secondary outcome measures of student and teacher sentiment were very positive in terms of the overall experience with the online intervention. We also report statistically significant evidence that the adaptive group more strongly agreed with the statement that "Woot Math

---

<sup>3</sup> The study was conducted during the last two months of the academic school year.

helped them understand some things better,” as well as evidence that the adaptive version may have sustained higher student sentiment throughout the treatment period.

## Research Goals of the Study

The two questions that our study was intended to answer were the following:

- A. Is there evidence that an adaptive version delivers better learning and retention?
- B. Is there evidence that the adaptive version either positively impacts student sentiment or leaves it unchanged?

These questions were investigated with a randomized controlled trial as described below. In the design, analysis, and reporting of the trial, we have followed the standards established by the Department of Education’s What Works Clearinghouse (WWC) and the CONSORT (Consolidated Standards of Reporting Trials) 2010 Statement (WWC, 2014; Schulz, Altman, & Moher, 2010).

We also surveyed students and teachers to determine their sentiments about the underlying intervention (independent of the RCT assignment to the adaptive version or the non-adaptive version). While we will touch on a few findings from the student and teacher surveys in this report, complete details can be found in a separate report at [wootmath.com/research](http://wootmath.com/research).

## Setting

The study was conducted in multiple sites (four) in geographically and demographically diverse settings. One of the sites was in the northeast region of the U.S. and the other three in the southwest region. Two sites were in urban settings in large metropolitan cities; the other two in small cities. At three of the sites, the study was conducted during class at public schools and as implemented by the classroom teacher. At a fourth site, it was conducted in a non-profit-sponsored after school program that offered tutoring sessions, at a public school, with tutors present as well as members of the research team.

For the study, we did not collect information on participants other than gender and English Language Learner (ELL) status (as reported by the teachers) – we did not collect information on race, free or reduced lunch status, or special education status. Demographics for the school populations underlying the samples at each of the four sites is given in Table 1. Note that these sites are listed in no particular order (i.e., randomly relative to the discussion in the preceding paragraph), and these demographics do not necessarily reflect those of the sample.

	Site 1	Site 2	Site 3	Site 4
Free or Reduced Lunch	70%	95%	36%	98%
White	22%	4%	58%	1%
Hispanic	69%	88%	35%	63%
Black	2%	3%	1%	34%
Asian	4%	5%	1%	1%

**Table 1.** Underlying population demographics at study sites.

Within the baseline sample, 53% of participants were female and 40% were reported as English Language Learners by their teachers (see section on Participants below).

## Study Design

Since the enrollment process and all data collection were done during online activities, centralized computer-based randomization procedures and data handling were utilized, allowing protocols for randomization and data collection to be implemented with a high degree of fidelity. The randomization was designed to equally allocate subjects to each group.

To minimize confounding factors a stratified and centralized randomization procedure was used. Our randomization software applied a strategy of per-classroom-section stratification and blocking with a block size of two. This allowed it to randomly enroll all students from any given section into the two treatment groups in equal numbers +/- 1. The implemented procedure was chosen so as to minimize any confounding factors that would result from unequal sampling by grade or any other factor that might vary by classroom section.<sup>4</sup>

The student participants received an average of three hours of the online interventions over several weeks. There was no statistically significant difference in usage by the test and control groups, who generally used the online software during the same class periods and in the same classrooms at each of the study sites. The primary outcome was measured using an online assessment designed by the researchers as a summative assessment of the content covered in the intervention.<sup>5</sup> Because the two treatments were so similar, the designed summative assessment allows us to accurately isolate and identify the effect of the experimental adaptive capabilities from any other effects, without any concern over aligning the assessment to national standards or standardized tests. The summative assessment is discussed in detail in the section on Outcomes below and in the appendices.

The trial was fully blinded – students, teachers, and investigators were not aware of which instructional pathway students had been assigned. We informally looked for any evidence of teachers or students breaking the blind by asking teachers whether they had any indication of which students were receiving which treatment or if any students noticed differences in their own versus others experiences, and we did not find any indication of blind breaking. Our statistical analysis was initially completed using blinded data.

## Participants

The recruitment effort focused on enrolling student subjects from grades 4-7 for the study period (April-May 2014). Subjects from grades 3 and 8 were also allowed to participate when teachers from those grades requested that they be allowed to join the study. Sites were selected for either a commercial pilot program of Woot Math or for the research study described here. This site selection was based on the number of subjects available for the study, the site demographics, and the feasibility of completing the approval process for conducting research in the district in the available timeframe.

---

<sup>4</sup> The procedure also took advantage of the way in which subjects initially enrolled into this study. For each classroom section in the study, all or almost all of the students would initially sign into the online activity nearly simultaneously. The functionally random process given by the order in which students passed the login screen guaranteed that the generated randomization sequence was both unbiased and unguessable.

<sup>5</sup> As further described below, the summative assessment consisted of a set of tasks derived from the common material seen by the participants in both the test and control treatment groups during the intervention.

A total of 16 teachers and 524 student subjects were recruited into the study at the four study sites; 358 of these student subjects met the per-site requirements for assent and parental consent and were enrolled and randomized into the study. All of the seventh and eighth grade participants were from a site that required parental consent be provided through a self-addressed stamped envelope process, which yielded a low permissioning rate (<10%) relative to the other sites. Moreover, the randomization procedure was applied when student participants first signed on to the intervention and both assented and stated that their parents had consented. Parental consent was later verified for each participant against the signed parental consent forms, and participants who had stated they had parental consent but for which there was no matching form were removed from the study. This process caused some per-classroom distributions to become unevenly split between the two treatment groups. This was markedly the case for four classroom sections for which the post-consent-verification sample was only  $n = 8$  students (i.e., 2.0 students per section). These sections contained all of the randomized grade-7 and grade-8 students ( $n = 5$ ), and unfortunately all five had been assigned to the adaptive treatment group, leaving no control for those grades. As including these participants would have biased the results of the study in favor of the adaptive treatment, we have instead excluded these four sections from the study,<sup>6</sup> leaving 350 student subjects in the study in grades 3-6, with identical average grade levels of 4.51 in both the test and control groups. In some of the secondary analysis below where we are not directly comparing test and control groups, we include these eight participants (and will explicitly state when doing so).

	Test	Control	Overall
<i>n</i>	175	175	350
Female	56.6%	48.6%	52.6%
English Language Learner	38.9%	41.7%	40.3%
Grade 3	6.9%	6.9%	6.9%
Grade 4	45.7%	45.1%	45.4%
Grade 5	36.6%	37.7%	37.1%
Grade 6	10.9%	10.3%	10.6%
Average Grade	4.514	4.514	4.514

**Table 2.** Participant background characteristics by treatment group.

The study was conducted during the last six weeks of the school year, so the participating students had nearly completed the grade for which their results are reported. Also, the third grade students were participating in an advanced mathematics intervention program, so that subpopulation is not expected to be representative of the general third grade population – we therefore report our key findings both including and excluding the third grade subgroup. As reported in the Table 2, none of the known participant background characteristics (grade, gender, and ELL status) varied between the test and control groups at statistically significant levels ( $p > .05$ ). The primary outcome measure (summative assessment score) was found to depend ( $p < .05$ ) on grade level but not on gender or ELL status.

<sup>6</sup> Including these eight students would increase the overall effect size reported in Table 4 from  $g = 0.23$  to  $g = 0.24$  under ITT analysis and from  $g = 0.28$  to  $g = 0.29$  under the PP-B analysis. Including only the 4 students from grade 6 (excluding just the grade-7 and grade-8 students) would increase the PP-B effect size for grade 6 from  $g = 0.46$  to  $g = 0.50$ . (And, in all cases reduce the respective  $p$ -values.)

## The Interventions

In this study, the test and control groups both received nearly identical online treatment interventions, each using Woot Math in our two-treatment RCT, with the only difference between the groups being the addition of the experimental adaptive features.

Woot Math is an adaptive learning environment for mathematics, which focuses on helping students in grades 3-8 master core math concepts, beginning with rational numbers. The supplemental software delivers a personalized progression of interleaved video instruction and scaffolded problems to mimic the natural give and take between a student and a tutor.

The baseline treatment for this study consisted of a version of Woot Math with a sequence of 29 modules (or “levels”) that were presented in a predetermined order as established by our content experts (see “Instructional Sequence,” below). This sequence was used in the same order for both the test and the control group, and we refer to these modules as the “mainline modules.” The adaptive treatment version included additional material, which was conditionally sequenced based on probabilistic modeling. Subjects in the two treatment groups received equal amounts of treatment (based on our measurements of time spent in the online activities); the two treatment groups just received different mixes of content. The summative assessment items (see below) were all derived from the common material in the 29 mainline modules seen by subjects in each of the treatment groups.

### Adaptive Treatment Overview

The adaptive treatment included all of the characteristics of the baseline treatment used for the control group and was identical to that treatment except for the addition of the following adaptive, personalized learning capabilities:

1. Adaptive Pacing: determining how fast each student should move along the instructional sequence
2. Adaptive Scaffolding: triggering help based on Bayesian models of each student’s understanding
3. Adaptive Supplemental Content: triggering of supplemental content (additional modules) based on the Bayesian models
4. Adaptive Task Selection: selecting tasks based in part on 3-parameter Rasch models
5. Adaptive Review Scheduling: adaptively inserting levels with review tasks

These categories of adaptivity have an extensive history of academic research. Early efforts for adaptive computer aided instruction such as Atkinson (1974) included adaptivity for sequencing, pacing, task selection and review scheduling. Application of Bayesian inference and probabilistic modeling for such applications dates to the same era (Rothen & Tennyson, 1978), as does related work for adaptive testing leveraging Item Response Theory and Rasch and Bayesian models (Lord, 1971; Owen, 1975; Weiss, 1976; Weiss, 1982). More recent reviews of the evolution of this research can be found in Wainer (2000), Desmarais & Baker (2012), Stone and Davey (2011), van der Linden & Glas (2010), and Shute & Zapata-Rivera (2012).

### Distinguishing Features of our Adaptive Methodology

Since our findings show some moderate to large effects from the mix of adaptive capabilities tested in our study, we highlight how they may differ from previously studied techniques. Two

foci of our research and development of adaptivity were (1) attempting to translate the best practices of expert teachers and tutors into software – best practices both as identified by academic researchers of classroom instruction and mathematics / rational number instruction; and (2) designing the adaptivity not just to measure and accelerate measures of learning but to equally focus on maintaining and improving student engagement and motivation. With regard to the second focus, it is important to note that affective states such as engagement and motivation are known to generate large effect sizes in learning outcomes (Hattie, 2008), so optimizations for affective measures may contribute significantly to the adaptive system’s effect on learning. Both our adaptive pacing systems and our adaptive task selection systems were tuned to optimize engagement at the expense of maximizing information available (ala Item Response Theory) or challenge to the student.

Jameson (2009, p. 119) identifies several challenges that “user-adaptive systems” may face and which may outweigh benefits of the adaptivity including: predictability, comprehensibility, controllability, unobtrusiveness, and breadth of experience. Most of these challenges bear on the affective state of the user (hence our research question B) and in designing our systems we attempted to minimize each of these challenges. Specific examples included:

- Unobtrusiveness: Not using a pre-assessment to establish initial model estimates on student knowledge (etc.) but to rather use adaptive pacing to rapidly move students through already familiar content
- Unobtrusiveness: integrating help at natural flow branch-points and often by recasting tasks rather than presenting explicit help tutorials, etc.
- Comprehensibility and Breadth of Experience: Similarly using rapid pacing rather than topic-pruning to move all students through a baseline instructional sequence
- Controllability: Allowing students to mostly control when they revisit topics to gain or demonstrate higher levels of mastery (through an Angry Birds style level screen and tiered star awarding) – Allowing students to revisit levels to gain mastery in a self-directed way can be expected to increase their intrinsic motivation to engage with the material (Zuckerman, 1978)

Our adaptive systems for scaffolding and supplemental content attempted to translate the best practices of expert teachers into software. We do not believe that expert teacher’s construct elaborate cognitive models about their students’ learning. Rather they are exceptionally good pattern matchers with a wealth of experience, which allows them to identify the key patterns that indicate points at which students need either reinforcement or corrective action in their learning. Cumming and McDougall (2000, p. 201) concluded that “expert teachers use a very wide diversity of types of individual learner information, some specifically acquired by questioning, but it is typically fragmentary and conjectural, and often mistaken. There is a striking contrast with the AIED [Adaptive Intelligence Education Device] strategy of constructing a more or less complete learner model, of domain knowledge only.” While it has been claimed that an advantage of software tutors is that they can model students at a level of fine granularity that no human tutor could match (Graesser, Conley & Olney, 2012), we have instead focused on coarse granularity models of a small number of ideas and misconceptions that expert teachers and researchers working in the subject domain identify as keys for understanding or common pitfalls (Cramer, Post, & delMas, 2002; Cramer & Wyberg, 2009; Daro, Mosher, & Corcoran, 2011; Gersten et al., 2009; Hamilton et al., 2009; National Research Council, 2001; Petit, Laird,

& Marsden, 2010; Seethaler, Fuchs, Star, & Bryant, 2011; Siegler et al., 2010; Webb, Boswinkel, & Dekker, 2008; Woodward et al., 2012).

Our system is concerned mainly with making conjectures about whether it would be appropriate for the software to apply any of a set of strategies derived from the referenced best practices for a given student. Just as with the expert teachers referenced in the earlier quote, we are not particularly concerned that the systems' model of the student is "fragmentary and conjectural, and often mistaken." There is very little risk incurred should the system unnecessarily offer additional scaffolding for a student or reintroduce a topic in an alternative way. We strove to create adaptive systems that were comprehensible and could be informed by classroom practice techniques as opposed to abstract cognitive models. The forms of adaptivity used in this study also are also based on effective techniques that have been used in the context of classroom formative assessment (Black & Wiliam, 1998; Black, 2004; Hattie & Timperley, 2007; Webb, 2004; Wiliam, 2006; Wiliam, 2011)

The present study can say only that the studied set of adaptive features were beneficial as an ensemble; it was not designed to separate out the benefits of the individual features. As this study will establish that the ensemble deliver a clear benefit, future studies could be conducted to establish how each of the various features contribute.

### **Instructional Sequence ("Mainline" Sequence)**

The instructional sequence for both treatments included 29 modules (also called "levels" in this report) of tasks and activities focused on fraction concepts, representations, relationships, and early computation. The modules addressed several topics aligned with the Number and Operations-Fraction sub-strand of the Grade 3 through 5 Common Core State Standards for Mathematics. The instructional sequence and the content for these modules was heavily informed by the seminal research out of the Rational Number Project (Cramer, Behr, Post, & Lesh, 2009; and the lineage of that work). In a review of the modules by the third author, which included listening to students' self-reflection on the reasoning they used in response to tasks completed within each module, it was found that:

- 41% involved comparison and equivalence tasks, focusing on numerators, denominators, and representations;
- 31% focused on modeling fractions by assembling or shading various discrete and continuous representations of part-whole relationships;
- 21% asked students to name fractions for a given representation;
- 7% of the questions involved operations and identification of proper vocabulary.

Most of the tasks required interactive use of the interface, such as drag-and-drop use of representations, sketching a diagram, shading icons and bar models, and inputting the correct fractions for a given representation. When we (the third author) asked students what they were thinking about as they were solving the problems, some of the typical responses included,

- "How many items I am shading"
- "What has the greatest denominator and what has the smallest numerator"
- "What was the total amount of parts and how many are shaded"
- "What one is greater than or less than"
- "How many there are total and how many I should shade"



The instructional sequence across the modules required active consideration of text and visual representations, making connections between numerical and visual representations, use and interpretation of academic language, and relational reasoning.

## Outcome Measures and Analysis

At the end of the Woot Math intervention, all students in the study were given a summative assessment as a Woot Math level – i.e., they completed the assessment within the same Woot Math online environment (iPad or web browser) in which they did the rest of their work. The assessment level was labeled as a ‘final review.’ Unlike other levels, students were not given feedback after each problem as to whether their response was correct or incorrect, and stars were not awarded at the end of the level.

The summative assessment consisted of 14 items designed by the research team and is illustrated in Appendix B. In designing the assessment, a pool of potential assessment items was tested prior to the study with a small group of sixth grade students (untreated – i.e., they had not completed the Woot Math unit) and the assessment items were selected from the test pool for face validity and with an eye toward maximizing our ability to measure positive learning outcomes relative to the performance seen in the pre-treatment study population. The 14 items assessed a range of student reasoning about fractions and required students to recall vocabulary, input the fraction for a given representation, identify accurate representations for a given fraction, draw or shade representations, and justify an answer with a model. The majority of tasks (71%) required reasoning beyond recall, such as making connections between representations, choosing an appropriate strategy to model a situation, or comparative analysis of statements. In addition, for the 14 tasks only 5 required students to select a response, so the majority of tasks (64%) required students to construct a response – e.g., input the correct numerators and denominators, drag a fraction to the correct location on a number line, draw a picture or model of a part-whole situation, model a fraction using a bar model, etc. Considering the grade level of students who completed this assessment, the reasoning required to successfully respond to these tasks was not trivial.

As the assessment was implemented in educational settings, there was no guarantee that a given student would complete the assessment in a single session, and the software allowed students to repeat the assessment level. Each student’s assessment was scored based on their final attempt for each of the 14 assessment items (most students made only one attempt per task). The items were equally weighted and no partial credit was given on any item and the total score was translated onto a 100 point scale. As reported in Appendix A, the cumulative score scale was observed to be highly reliable (with  $\alpha = .80$ ,  $\omega = .82$ , and the glb = .87 over 318 observations), exceeding the WWC standard for reliability (WWC, 2014, p. 15), and all 14 tasks positively correlated with one another and with the cumulative score.

The assessment was not designed to be a standalone measure of student facility with ordering and equivalence of fractions (the subject matter covered in the unit) on any independently reportable scale. Rather it was intended to measure differences in learning and retention between the treatment groups. As such, scale reliability is important but convergent validity to any external measure is not a concern. All assessment tasks were similar to tasks given within the unit and were reviewed by content experts and assessed as having content validity prior to administration. The primary outcome was analyzed through two-tailed *t*-tests on the summative

assessment score distributions for the adaptive and control group to measure differences in means for the aggregate sample and disaggregated subsamples.

### Secondary Outcomes

Following the summative assessment, all student participants were given a student survey as a final Woot Math level. The survey consisted of 12 questions, 11 of them multiple choice, and one asking for a list of grades in which students would enjoy Woot Math the most. In the event that the student did not complete the survey, (or left multiple blank responses at the end of the survey), we counted only the first unanswered question as “no response”. Only one survey response per student was included in the analysis, and surveys were excluded for students who responded to all multiple choice prompts without variation (e.g. A,A,A,...). Seven completed surveys were excluded under this criterion.

As reported in the Findings section below, we analyzed the surveys by subpopulations to see if there were significant differences in sentiment by treatment group. We tested for statistical significance of differences using a nonparametric, rank-based, Mann-Whitney-Wilcoxon test. The student survey questions are briefly summarized in Table 3 below. Full details on the student survey can be found in a separate white paper available at [wootmath.com/research](http://wootmath.com/research). The survey summary results include all students with parental permission who completed a survey in grades 3-8.

1	How helpful did you find Woot Math?	95% Helpful*	67%: “Very helpful”
2	Did you learn new things about fractions?	90% Yes	50%: “Yes, a lot”
3	Did WM help you understand some things better?	90% Yes	47%: “Yes, a lot”
4	Do you feel more confident about math after using WM?	90% Yes	52%: “Yes, much more”
5	Do you feel more confident about fractions after using WM?	88% Yes	48%: “Yes, much more”
6	Did you find Woot Math confusing?	72% A little**	31%: “Not at all”
7	Did you enjoy using Woot Math?	93%: Yes	63%: “Yes, a lot”
8	Did you enjoy learning about fractions more than you expected?	90% Yes	59%: “A lot more”
9	What grades would enjoy using Woot Math most?		see Figure 5
10	Would you say the problems were:		79%: right difficulty
11	Would you say the levels went:		81%: right speed
12	Would you recommend Woot Math to another student?	96% Yes	65%: “Yes, for sure”

**Table 3.** Summary of Student Survey. The questions are listed here in the same order that they were presented to students, and all were Likert type 4-point multiple choice responses (with two gradated positive and negative options) except questions 10 and 11, which were 3-point, and 9, which was a student-selected grade range. \*For question 1, 95% of subjects responded with either “somewhat helpful” or “very helpful.” \*\*For question 6, 71% of subjects responded with “only a little of the time” or “not at all.”

Students participants were also periodically prompted (within the online environment) to answer sentiment and confidence questions about the current instructional material as illustrated in Figure 1. Both were presented with red/yellow/green stoplight iconography (a stoplight being a common formative assessment metaphor).

At the conclusion of the trial, participating teachers were surveyed and interviewed to measure their sentiment about the effectiveness of the interventions, overall. (Because teachers were

blinded to the treatment group assignments, they were not surveyed about the two treatments, only the overall effectiveness of the Woot Math interventions.) One additional teacher was included in the interviews who was not part of the RCT but ran his/her own controlled trial, using Woot Math with one class and comparing outcomes against a second class with a similar student composition. Student and teacher sentiments about the underlying, common, intervention is reported in a separate white paper available at [wootmath.com/research](http://wootmath.com/research)






Do you like this lesson?		How is this lesson going so far?	
	No		Help! I don't get it
	It's okay		I understand it, but I need a little support
	Yes		I understand this very well

Figure 1. In-app prompts for sentiment and confidence.

## Findings

Of the 350 subjects, 318 (91%) completed the summative assessment – 163 of the 175 students randomized into the non-adaptive treatment group, and 155 of the 175 randomized into the adaptive group. This attrition rate (9%) with a differential of 4.6% between the treatment groups meets the WWC’s evidence standards for RCTs (WWC, 2014, pp. 10-12). Furthermore, for this study, the attrition is expected to be due to largely exogenous factors such as student absence on the day particular classroom sections completed the assessment portion and student transfers out of sections. Because the two treatments offered nearly identical experiences with Woot Math (from the students’ perspective and as measured in their sentiment), it is unlikely that any significant attrition was endogenous.

The primary outcome results for the RCT are shown in Table 4. The results demonstrate a causal relationship between the experimental adaptive capabilities and improved learning and retention under the outcome measure. The statistical significance of these results was confirmed using two-tailed *t*-tests and lines are highlighted in bold were statistically significant ( $p < .05$ ) in Table 4. More generally, these results show that the studied class of adaptivity for personalized learning can produce moderate to strong effect sizes in an RCT.

In terms of the three sections of Table 4, we have analyzed the assessment data for the primary outcome hypothesis based on both an intent-to-treat (ITT) principle (including all subject for whom we have assessment data, whether or not they received any of the treatment), as well as two per-protocol (PP) analyses:

- One, per-protocol B (“B” for “began”), excluding subjects who did not begin the treatment
- The second, per-protocol C (“C” for “completed”), excluding subjects who didn’t complete all of the intended treatment, i.e. who did not complete the entire content sequence.

Since this study focused on efficacy of the adaptive treatment versus the non-adaptive (and otherwise identical) treatment, we consider the PP-B results as answering the research question A with the greatest accuracy and least bias. We have included the ITT analysis as a baseline, such as might be used in a single-treatment trial as a more conservative effectiveness analysis

(Lachin, 2000). The PP-C analysis (excluding subject who did not complete all of the treatment) introduces an obvious selection bias in that students with higher pre-existing familiarity or facility with the material covered are assumed to be more likely to complete the unit, but also allows more to be said about the possible effect size of one intervention over the other when they are fully implemented.

	Adaptive Group			Control Group			t-Test		Effect Size	
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>t</i> -stat	<i>p</i> -value	Significance	Hedges' <i>g</i> *
<b>Intent-to-treat:</b>										
<b>All</b>	<b>155</b>	<b>56.7</b>	<b>24.5</b>	<b>163</b>	<b>51.2</b>	<b>24.5</b>	<b>+2.02</b>	<b>.045</b>	<b><i>p</i> &lt; .05</b>	<b>+0.23</b>
<b>Per-protocol B:</b>										
<b>All</b>	<b>147</b>	<b>58.7</b>	<b>23.4</b>	<b>159</b>	<b>52.1</b>	<b>24.1</b>	<b>+2.46</b>	<b>.015</b>	<b><i>p</i> &lt; .05</b>	<b>+0.28</b>
<b>Grades 3-5</b>	<b>130</b>	<b>58.4</b>	<b>23.2</b>	<b>143</b>	<b>52.2</b>	<b>24.6</b>	<b>+2.12</b>	<b>.038</b>	<b><i>p</i> &lt; .05</b>	<b>+0.26</b>
<b>Grades 4-6</b>	<b>139</b>	<b>58.2</b>	<b>23.7</b>	<b>148</b>	<b>52.3</b>	<b>24.3</b>	<b>+2.08</b>	<b>.038</b>	<b><i>p</i> &lt; .05</b>	<b>+0.25</b>
<b>Grade 3</b>	<b>8</b>	<b>67.9</b>	<b>14.3</b>	<b>11</b>	<b>48.7</b>	<b>22.3</b>	<b>+2.12</b>	<b>.049</b>	<b><i>p</i> &lt; .05</b>	<b>+0.94</b>
<b>Grade 4</b>	<b>67</b>	<b>54.6</b>	<b>23.4</b>	<b>73</b>	<b>45.8</b>	<b>22.3</b>	<b>+2.28</b>	<b>.024</b>	<b><i>p</i> &lt; .05</b>	<b>+0.38</b>
Grade 5	55	61.7	23.4	59	60.9	25.5	+0.17	.863	n.s.	+0.03
Grade 6	17	61.3	25.5	16	50.4	19.8	+1.37	.182	n.s.	+0.46
Male	67	58.7	25.0	83	52.5	24.3	+1.55	.124	n.s.	+0.25
Female	80	58.7	22.1	76	51.6	24.1	+1.93	.055	n.s.	+0.31
<b>ELL</b>	<b>59</b>	<b>56.8</b>	<b>22.7</b>	<b>68</b>	<b>48.8</b>	<b>21.7</b>	<b>+2.01</b>	<b>.047</b>	<b><i>p</i> &lt; .05</b>	<b>+0.36</b>
Non-ELL	88	60.1	23.8	91	54.5	25.6	+1.51	.133	n.s.	+0.22
<b>Per-protocol C:</b>										
All	51	74.1	17.8	45	68.3	21.9	+1.44	.150	n.s.	+0.29
<b>Grade 4</b>	<b>15</b>	<b>77.1</b>	<b>12.1</b>	<b>11</b>	<b>55.2</b>	<b>16.6</b>	<b>+3.90</b>	<b>.0007</b>	<b><i>p</i> &lt; .001<sup>†</sup></b>	<b>+1.50</b>

**Table 4.** The RCT Primary Outcome for the aggregated population and disaggregated subsample by: (1) all subjects (intent-to-treat section), (2) all subjects starting treatment (per-protocol B section), and (3) all subjects completing the full treatment (per-protocol C section). In each case we show the sample size for each group, mean and standard deviation of their summative assessment scores, the results of two-tailed *t*-test analyses, and the effect size as standardized mean difference (Hedges' *g*\*). Effects benefiting the adaptive groups are shown as positive (and were positive in all cases reported here). Results with statistical significance (*p* < .05) are highlighted in bold. n.s. = not significant. ELL = English language learner.

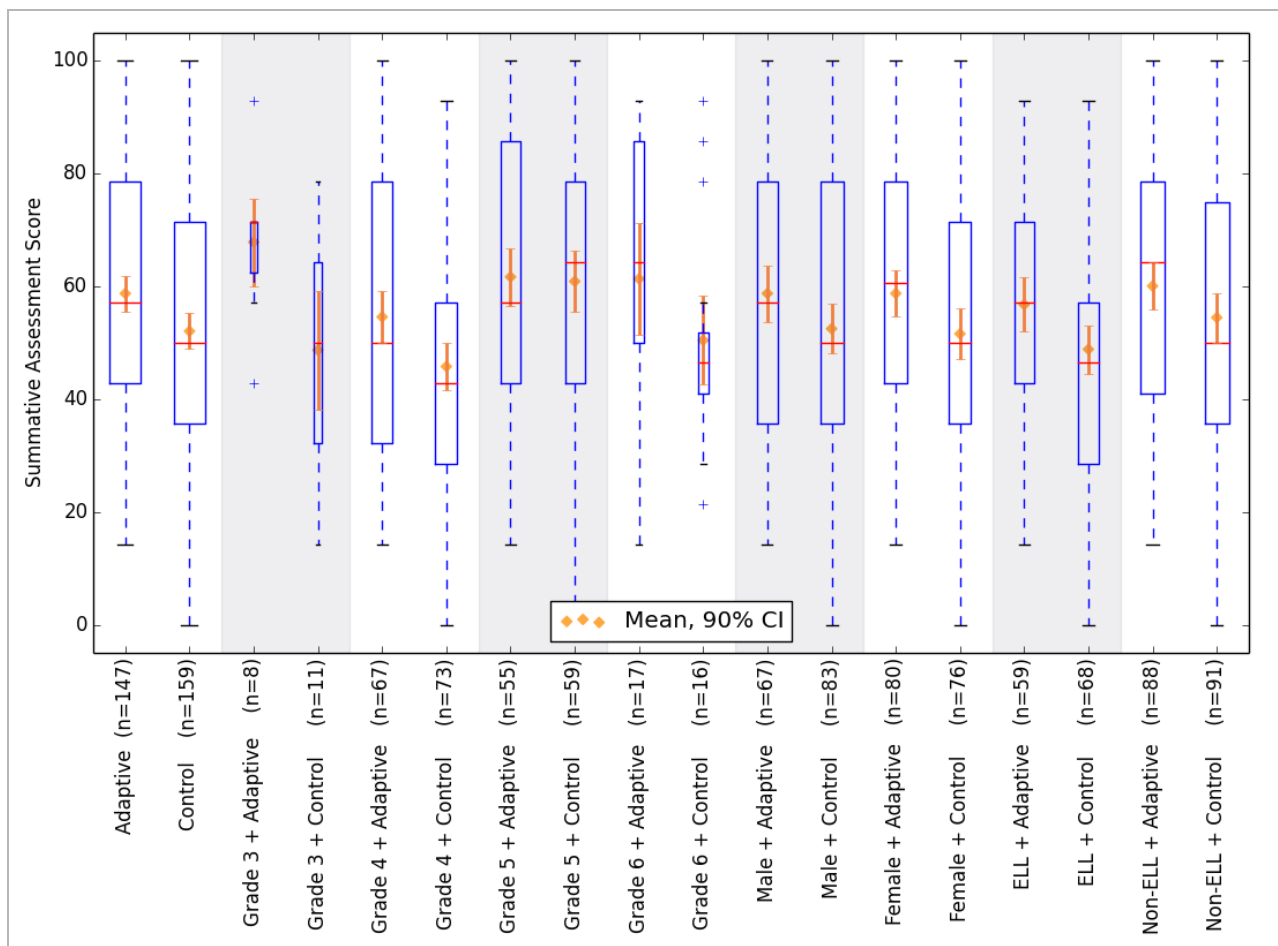
<sup>†</sup>For the Grade 4 per-protocol C result: *p* < .001, or, *p* < .01 after a Benjamini-Hochberg correction for the multiplicity of hypotheses in our subpopulation analysis.

Per-protocol B results exclude students who took a final assessment but either

1. Did none of the content modules or
2. Did only the first 'intro-to-Woot-Math' module, which contained no subject matter content and no adaptivity

There were 4 subjects in the former category and 8 in the latter. As these 12 subjects received no differential treatment, any observed effect can only be at random, so including these subjects adds only error and dilutes any real evidence of efficacy. By chance, the majority (two thirds) of these PP-B-excluded students were in the adaptive group. Therefore, in terms of comparing the two treatment effects, including these students biases the results – i.e., the ITT results are

slightly biased in that a higher percentage of subjects in the control group received some treatment relative to the subjects in the test group. For these reasons, we believe that the PP-B analysis is the least biased and most conservative analysis in terms of measuring the relative efficacy of the two treatments, and that is where we focus our discussion.



**Figure 2.** Summative Assessment Scores (per-protocol B: subjects starting the treatment), illustrated as box plots with median (red line), 25%tile, and 75%tile as well as the mean (orange diamond) and 90% confidence interval for the mean (orange error bar). Adaptive and control treatment groups are shown for the full population and various disaggregated subpopulations by grade, gender, and ELL status. Box widths are scaled as the square root of subgroup sizes.

For per-protocol B, we see effects in favor of the adaptive treatment group for all subpopulations except grade 5 (for which the outcomes were essentially equivalent),<sup>7</sup> and statistically significant improved outcomes for the entire population, and for the disaggregated subpopulations of grade 3 students, grade 4 students and ELL students (all with  $p < .05$ ). The assessment outcome scores for the PP-B subpopulations are also illustrated in Figure 2 with box and whisker plots, in which the relative gains of the adaptive treatment groups is visually apparent.

<sup>7</sup> Regarding the lack of measured effect in grade 5, the most likely explanation is that this was simply a result of chance given the sample size, though it merits further investigation to confirm that there was not some moderator that limited the effectiveness of the adaptivity only in grade 5. The observation of effects in both of the adjacent grades, however, makes such a moderator less likely.

As we noted earlier, the grade-3 sample was drawn from a group of advanced students who used Woot Math as part of their advanced math intervention sessions and therefore not representative of the general grade-3 population. The effect size observed in this subpopulation was particularly large ( $g^* = 0.94$ ), allowing statistical significance to be found even with the small group size. Because of the non-representative population bias in grade 3, for PP-B we also report results for the subpopulation excluding this specialized grade-3 subpopulation, i.e. for grades 4-6. We also report the disaggregated elementary grade (3-5) sample.

In the per-protocol-C analysis (where we are looking for effect size under greater treatment exposure despite selection bias as discussed above), the smaller remaining sample sizes yielded just one statistically significant result, which was for grade 4, where we do see a very large effect size,  $g^* = 1.50$  with  $p < .001$ . Although the other subpopulations did not have statistically significant effects in the smaller PP-C sample, all except grade 5 continued to favor the adaptive groups in effect, and it is likely that a larger study would show effect sizes increasing with treatment exposure across all subpopulations.

### Secondary Outcome Findings

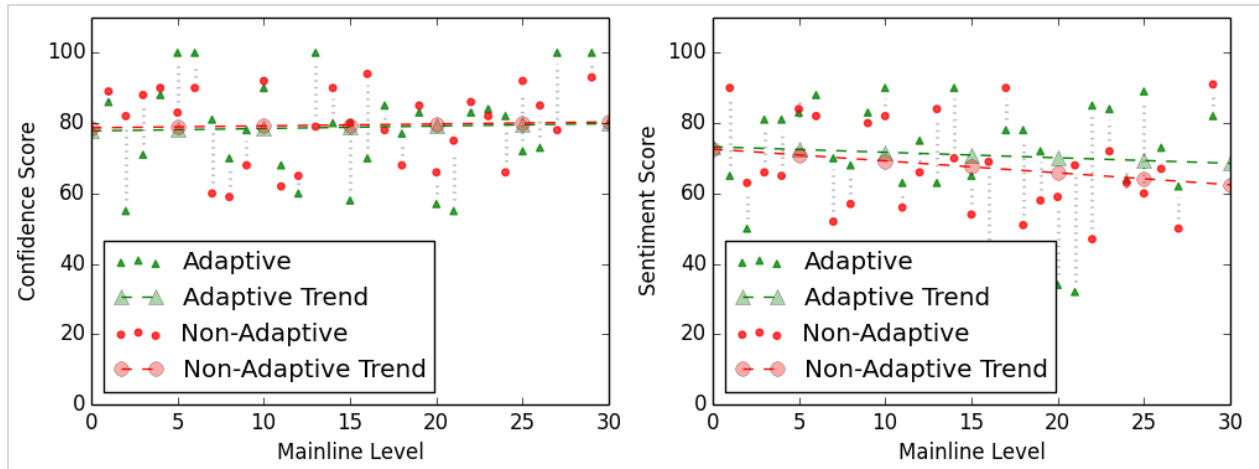
The only statistically significant difference between the adaptive and control treatment group in the survey responses<sup>8</sup> was that the adaptive group was more likely to agree (Q 3) that “Woot Math helped you understand some things better,”  $p < .05$ . Statistical significance was calculated using a nonparametric, rank-based, Mann-Whitney-Wilcoxon test, for which  $U = 17,113$  and  $p = .045$  (de Winter & Dodou, 2010). See Table 5 for details on the student responses to this Likert-type survey item. This result provides clear evidence on our research question B, “that the adaptive sequence either positively impacts student sentiment, or at least leaves them undiminished,” as do the results discussed below pertaining to in-app sentiment measures.

		Adaptive Group		Control Group		Overall	
	Score		<i>n</i>		<i>n</i>		<i>n</i>
“Yes, a lot”	+2	52.3%	79	42.5%	68	47.3%	147
“Yes”	+1	39.1%	59	46.9%	75	43.1%	134
“Not so much”	-1	7.9%	12	8.8%	14	8.4%	26
“Not at all”	-2	0.0%	0	1.2%	2	0.6%	2
<b>No Response</b>	0	1.0%	1	1.0%	1	1.0%	2
<b>Positive Response (“Yes[...]”)</b>		91.4%	138	89.4%	143	90.4%	281
<b>Total Responses</b>			151		160		311
<b>Mean Score</b>		+1.358		+1.206		+1.280	
<b>Effect Size (Hedges’ <math>g^*</math>)</b>		+0.17					

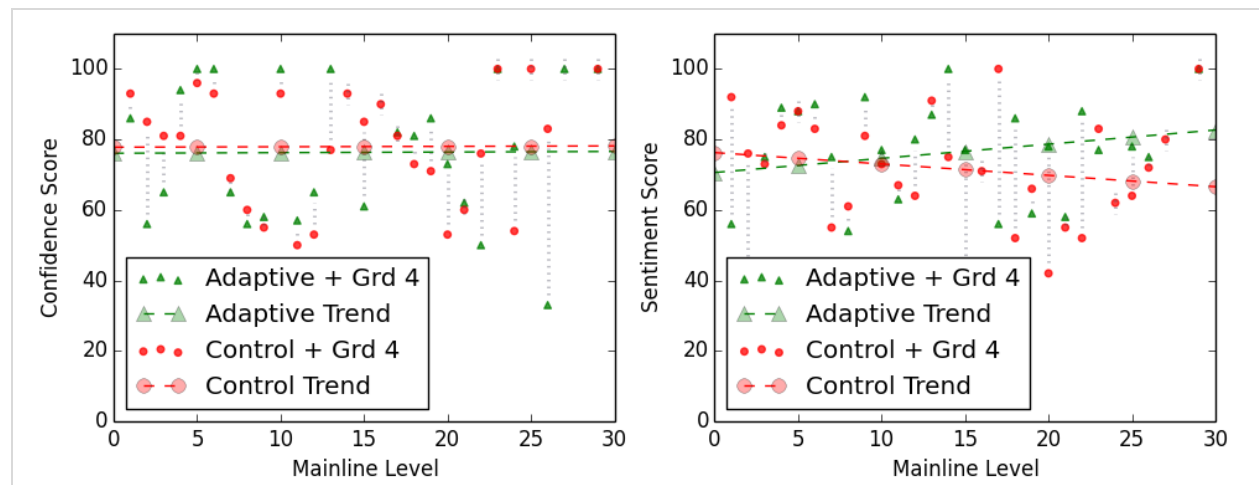
**Table 5.** Student responses to the survey prompt “Did Woot Math help you understand some things better?” The four selectable responses are shown along with the response count and percentages by treatment group. Responses were put onto a numerical scale (scored as shown in the second column) and group statistics and effect size computed as Hedges’  $g^*$ .

<sup>8</sup> The analysis for survey differences between the test and control group was limited to students ( $n=305$ ) included in the PP-B sample (i.e., at least began the treatment)

We also analyzed student sentiment and confidence via the in-app prompting discussed above. Since the two treatment groups followed the same 29 mainline module content arc, we were able to compare the evolution of these sentiment variables (and other factors) over the elapsed treatment time. For Figures 3, we have calculated average sentiment under scores of green = 100, yellow = 0, and red = -100 and plotted how these measures trended (using linear regression) as students went through the material. In terms of our research question B, we again see evidence that affective outcomes are either unchanged or slightly improved under the adaptive treatment condition. In particular there appears to be some evidence that the adaptive treatment resulted in improving sentiment scores over time.



**Figure 3.** Comparison of adaptive and control groups in terms of their in-app expressed confidence and sentiment (see Figure 1) correlated to the 29 mainline levels. Linear regressions are nearly identical, with the sentiment measure indicating the trend for the adaptive group may be slightly better.



**Figure 4.** For the disaggregated, grade 4, subsample, comparison of adaptive and control groups in terms of their in-app expressed confidence and sentiment correlated to the 29 mainline levels. Here the sentiment trend for the adaptive group appears substantially improved over that for the control group.

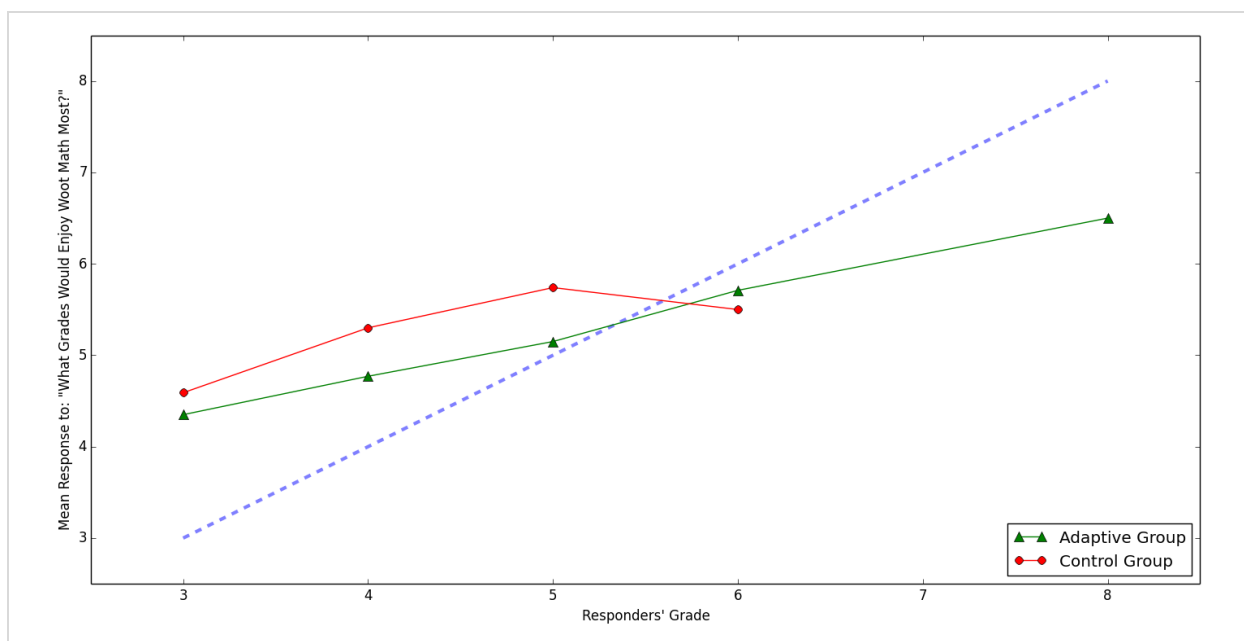
In looking at the disaggregated populations, the tendency towards improving sentiment appears to have been strongest in grade 4 as shown in Figure 4. This trend was offset by a reversed trend in the grade 6 data. The sample size in this study was not large enough to establish any of

these trends with statistical significance, and the observed trends may be simply due to chance. However, the fact that a large, statistically significant effect was seen in the PP-C primary outcome for grade 4 and the sentiment data for grade 4 showed the most significant trend favoring the adaptive group is a coincidence that calls for further investigation.

### Findings about Grade Level Appropriateness

Figure 5 illustrates how participants from the adaptive and control group answered the student survey question: “What grades would enjoy using Woot Math most?” In response to this question the student could select a range of grades (e.g. 3 through 8, or 4 through 5), and in the figure we have plotted the average of all grade ranges selected against the grade of the responding student and the treatment group to which they were assigned. Note that the sample here includes all students who completed the survey, including the additional eight students discussed in the Participants section (N = 358 rather than N = 350) – so, in particular, a grade-8 sample is available for the adaptive group.

Noteworthy is the apparent difference in slope between the plotted adaptive and control group data and that in all cases (i.e., at all responder grades) students in the adaptive group selected grade levels closer to their own than the control group. That is the adaptive group consistently found the software to be more appropriate for their own grade level than did the control group.



**Figure 5.** Student predictions about what grades would most enjoy Woot Math (y axis giving mean of the predicted grades) plotted by their own grade level (x axis) and treatment group (two curves). Blue dashed line plots the identity map (mean prediction matching the responder’s grade).

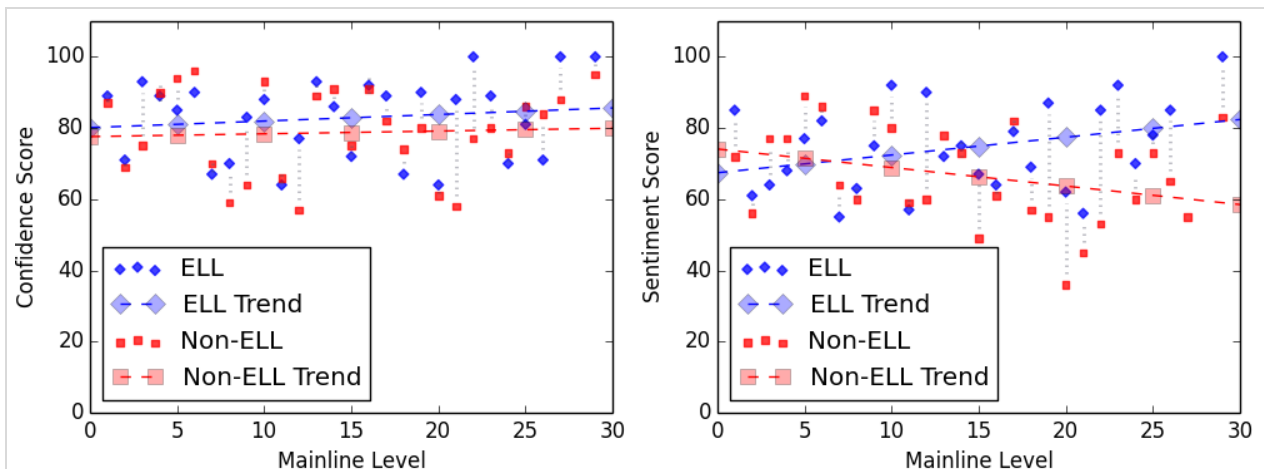
### Achievement Gap Findings

Because it has been found that adaptive learning environments can widen rather than narrow the range of student achievement gaps (Steenbergen-Hu & Cooper, 2013), we felt it important to look for any evidence for similar effects in our dataset, despite the fact that the study was not designed with this purpose in mind. As can be seen in Table 4, the summative assessment



score variances for the adaptive group was smaller than that of the control group for both PP-B and PP-C. This provides at least some anecdotal evidence that the experimental adaptive features did not contribute to widening the achievement gap and may have narrowed it. The differences in variances were not large enough to be statistically significant with the available sample size. We tested for statistical significance (of the population variances being different) using Levene’s test and found no significant results; the overall PP-C measurement came closest to meeting statistical significance with  $p = .08$ . Also visibly apparent in the box and whisker plot of Figure 2 is the trend for the bottom of the range of the score distributions for the adaptive group and subgroups to be pulled up relative to that of the control group and subgroups. This provides some evidence that the adaptivity benefited the lowest performing students in the sample.

We also looked at outcomes for ELL vs. non-ELL students to see if there was any evidence of the adaptivity disadvantaging the ELL group and did not find any, but rather some evidence to the contrary. As shown in Table 4, the difference in summative assessment scores for ELL vs non-ELL students was greater in the control group than it was in the adaptive group (though none of these differences were statistically significant to  $p < .05$ ). Also, as illustrated in Figure 6 is an apparent trend for ELL students to have rising sentiment with more exposure, unlike their peers. All of these anecdotal observations merit further investigation.

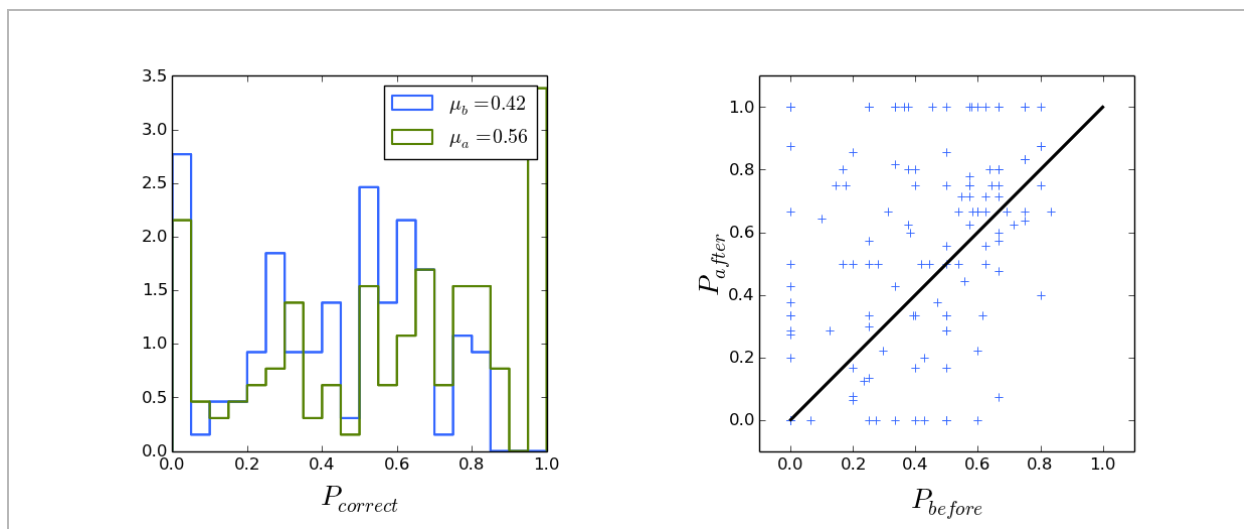


**Figure 6.** In our subpopulation analysis, the most striking variation was the difference in sentiment trends between ELL and non-ELL students illustrated here.

### Findings Regarding the Effectiveness of Scaffolded Help

As we have mentioned, the study was not designed to identify the relative effectiveness of any of the isolated adaptive features that were enabled for the test group. For one of the features, however, there was an opportunity to isolate and measure a localized effect, which was when the adaptive engine triggered “just-in-time” supplementary instructional videos as a form of scaffolding within a problem set. In this case, we have before and after data on how well the student performed comparable tasks. To analyze this data, we partitioned tasks into sets posed before and after the presentation of help and calculated frequencies of correct responses for each. As illustrated in Figure 7, the average probability of a correct response improved from .42 to .56 between the two sets. The null hypothesis of an unchanged probability before and after

help was rejected with  $p = .001$ , thereby demonstrating that the scaffolded help had an immediate positive effect on student performance. The students from the non-adaptive sample had a probability of a correct response that averaged .76 for the same set of problems; that this value exceeds the probabilities in the aforementioned sets indicates that our model is selecting struggling students for scaffolded help, as it should. While the data demonstrate the effectiveness of the scaffolded help within the problem set, we cannot say how this translated to an effect size as measured at the summative assessment.



**Figure 7.** The effectiveness of scaffolded help. The left panel shows the distribution of the frequency of correct responses before and after scaffolded help was shown. The right panel shows the correlation between frequency of correct responses before help (x-axis) and after help (y-axis) for the same tasks (i.e. the same task given to different pools of students); the fact that the data points are centered above the diagonal is indicative of efficacy.

### Post-Hoc Analysis of Outcome vs Treatment Exposure

The choice of the PP-B and PP-C thresholds are natural ones, but these thresholds could also be placed on a continuum and we can investigate the statistical outcomes as a function of how much of the treatment participants received. Since at a fine granule-level the treatment was not homogeneous, but consisted of a sequence of 29 adaptive modules compared against 29 non-adaptive modules, there is some value in examining outcomes relative to both treatment exposure and against this granularity.

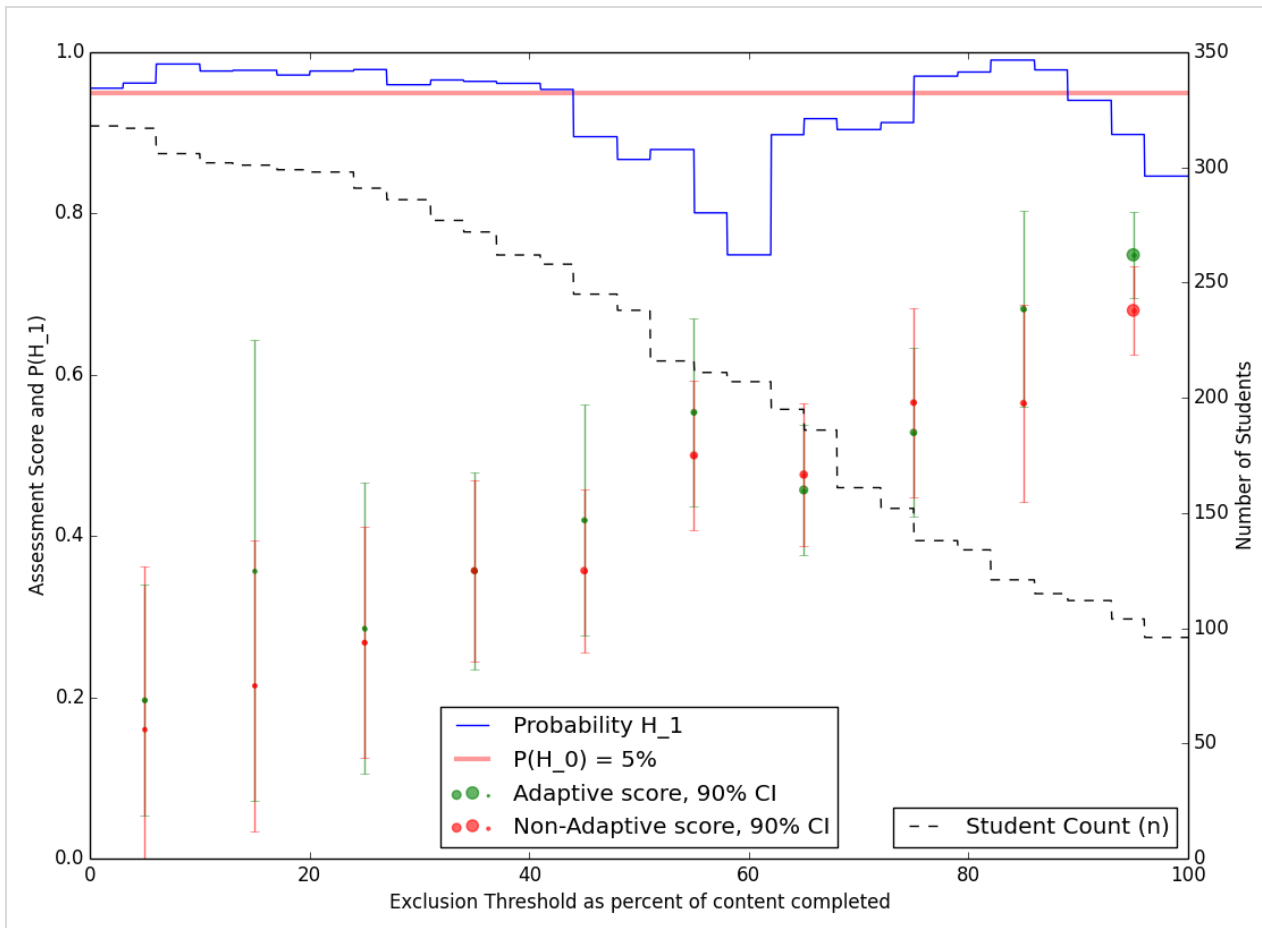
Figure 8 provides a summary view into such an analysis. In this figure multiple data elements are plotted against a horizontal axis representing the subpopulation of students who reached the given percentage of the content set (as measured against the 29 mainline levels) during the study. The black dashed curve gives the number of student participants who reached at least the given percentage of content and took the summative assessment – i.e. at the left hand side curve begins at the ITT outcome observation count ( $n = 308$ ) and at the right hand side drops to the PP-C count ( $n = 96$ ). The blue curve similarly shows the calculated  $p$ -value from t-tests on the outcomes of the two treatment groups excluding all students who did not complete the give percentage of the total content – so, again, at the left hand edge, this curve gives the  $p$ -value for the ITT analysis and at the right edge the PP-C  $p$ -value. (Likewise, the applied PP-B threshold is the first stairstep threshold shown). The figure also illustrates the primary outcome results for

each treatment subgroup of students who completed 0%-10% of the content, 10%-20%, etc.; these are shown as means with 90% confidence intervals (and with the point indicating the mean sized in proportion to each subgroup size).

This figure illustrates many things including the correlation of final assessment score with percentage of content completed. Other elements should be interpreted with the caveat that they are likely due to chance given the relatively small sample sizes involved. For example, the bimodal appearance of the  $p$ -value curve could be:

- attributable to chance
- attributable to a growing effect size vs. a shrinking sample size as the threshold is raised
- attributable to a real effect resulting from the non-homogeneity of the treatment (that is, a weakening or genuine reversal of the adaptive effect for some of the modules around the midpoint of the treatment)

An equivalent observation was that of a reversal of the test vs. control effect for students completing between 60% and 80% of the content. Again, an overall  $p$ -value of only .05 made it likely that such reversals would be found by chance during such subpopulation analysis.



**Figure 8.** Analysis of outcomes as a function of treatment exposure.

That being said, the possibility that a real effect might be attributable to some of the levels around the midpoint of the treatment was worth considering. This possibility was somewhat

corroborated by the fact that in a post-trial content review, several levels in this range were identified as those most in need of improvement. (The review was unrelated to the trial; all of the Woot Math content is iteratively reviewed in its entirety and iteratively improved). This coincidence raises the question: might the adaptive capabilities studied here amplify both the strengths and weaknesses of the underlying content?

## Conclusions

Surveys have found that K-12 teachers, principals, and administrators rank “intelligent adaptive learning software” just behind “1:1 computers” and “Internet access” as the most desirable technologies for improving student achievement (Project Tomorrow, 2011). This desire on the part of educators underscores the appeal of adaptive technology for differentiated learning. It holds promise, despite the fact that such technology has yet to be proven effective. There has been a lack of quality empirical evidence showing that the adaptive learning software available to date delivers any significant effect in terms of student outcomes (Shute & Zapata-Rivera, 2012; Steenbergen-Hu & Cooper, 2013). Ongoing research is therefore called for to continue to evolve the adaptive technology and empirically investigate its effect on educational outcomes so as to close the gap between the promise that such technology holds and the minimal results that it has achieved to date.

The findings in this study demonstrate that “intelligent adaptive learning software” can indeed deliver moderate to large effects in learning and retention as measured in a high-quality randomized controlled trial. This is an important result, providing evidence that the perceived promise of adaptive educational software is likely real, and indicating the need for follow-on research to continue to capture more of the potential that is available.

This study examined the effects of adaptivity with respect to a small, well-defined topic: ordering and equivalence of fractions, a topic that has long been a challenge for students to master and has been the subject of much research. The fact that this topic is challenging to many students allowed us to measure an effect across a wide student population (grades 3-6). This fact may also have contributed to our ability to measure relatively large effect sizes ( $g = 0.23$  to  $g = 1.50$ ,  $p < .05$ ) despite the fact that the intervention was relatively brief (average online time per student was 3.1 hr, typically over 3 to 4 weeks). There was also some anecdotal evidence found that the adaptive treatment did not widen student achievement gaps (a concern that has appeared in studies of earlier ITS's) and possibly narrowed them.

Questions about the generalizability of the results from this study include the following, which could be addressed through follow-on research:

- Do the results generalize to other topics within mathematics, other grade levels, or other subjects?
- Are the results dependent on the style of online instruction and theory of learning used by Woot Math?
- To what extent are the results dependent on the observed levels of student engagement?
- To what extent are the results dependent on context and conditions of use? In particular, what is the impact of longer-term treatment and how well are effect sizes maintained when measured sometime after the treatment is completed?

The adaptive techniques studied should certainly be expected to generalize to other topics, subjects, and grades, as nothing specifically limits them to the contexts chosen for this study. That said, it is certainly the case that the application of those techniques for this study benefited from the measurability of mathematical tasks and from the extensive pedagogical research on the ordering and equivalence of fractions. One noteworthy hypothesis is that adaptive learning software may amplify both strengths and weaknesses of underlying instructional approach, and that in the case of this study, the adaptive effect represents the magnification of a strong underlying effect derived from decades of research on teaching and learning rational number. Should that hypothesis bear out, it would mean that not only does adaptive educational software have the potential for improving student outcomes, it can also become an important tool for studying the effectiveness of curriculum and instruction.

Woot Math, the online intervention to which adaptivity was applied in this study, has various characteristics that could be relevant to the measured adaptive effect. Woot Math takes a constructivist view of learning and includes a strong focus on modelling, approaching mathematical ideas from multiple perspectives, progressive “mathematizing,” and the student’s guided rediscover of mathematics – much in the tradition of Realistic Mathematics Education and its lineage – for an overview of these theories of learning see Fosnot (2005). As such Woot Math has a strong focus on student tasks and the design and modeling forms underlying those tasks. Secondary to the student tasks are brief instructional segments, which help to guide the student’s own discovery through these tasks. Through multiple approaches to mathematical ideas and modelling, it also leverages a “learning landscape” philosophy, which is well suited to adaptive software augmentation.

It is also important to consider student engagement as a possible moderator of adaptive effect sizes. As we have reported, the measured student engagement and motivation during this trial was very high, and we took care in the design of our adaptive systems to avoid negatively impacting such sentiments. It continues to be our opinion that the provision of engaging online experiences and an effective underlying pedagogy are necessary but not sufficient conditions for the creation of effective adaptive instructional software. In our study, we hold these factors constant and examined the potential benefits of adding certain adaptive techniques. We hope that the findings reported here cast some additional light on these adaptive techniques and their considerable potential for improving student outcomes.

## Acknowledgements

This report is based upon work supported by the National Science Foundation under Grant No. IES-1345969. 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 are very grateful for the support of the participating teachers, schools, and districts, the National Science Foundation, and most of all for the participating students, without whom this study would not have been possible.

## Appendix A: Psychometric Analysis of the Summative Assessment (Primary Outcome Measurement Scale)

The study reported here used a researcher-developed summative assessment to measure the primary outcome. Because the measurement device was researcher-developed and new for this study, we have carefully examined it for validity and reliability. Content validity was confirmed through review by multiple content experts prior to administration and face validity was confirmed by classroom teachers and students. Reliability and other important psychometric properties are discussed below. To compute the scale score to measure the primary outcome, the items were equally weighted and no partial credit was given on any item. (I.e., individual item scores were either 0 or 1). For reporting purposes the the total assessment score (the sum of all item scores) was linearly mapped onto a 100 point scale.

The summative assessment consisted of 14 items, and most of the items included a small component of parameter randomization, which was applied on a per-student basis. For the analysis in this appendix, we have ignored this random variation of item parameters across test observations, assuming all item variants measure the same factor(s) for every given item. Appendix B shows example parameterizations for all 14 items along with screenshots of student work. Table A-1 gives descriptive statistics on the 14 items comprising the summative assessment. The analysis reported in this appendix was generally done using R and the ‘psych’ package (Revelle, 2014) where applicable. Additional analysis and verification was done using Python. The analysis in this appendix was based on the 318 study participants completing the assessment, i.e. the 318 students whose assessment scores were used to establish the primary outcome findings for the study.

In this analysis, the cumulative score scale was observed to be reliable under both the traditional Cronbach alpha measure as well as more recent, less biased measures (Revelle & Zinbarg, 2009; Sijtsma, 2009):

Observations	=	318
Coefficient $\alpha$ (Cronbach)	=	.80, 95% CI: [.77, .83]
Coefficient $\omega_t$	=	.82, 95% CI: [.79, .84]
Coefficient $\omega_h$	=	.63
Max split half reliability ( $\lambda_4$ )	=	.85
Min split half reliability ( $\beta$ )	=	.70
“Greatest Lower Bound,” glb	=	.87

Factor analysis (see below) shows that the assessment measured a general factor and that the items are homogenous in that they share a common latent variable. Coefficient  $\omega_h$  estimates the proportion of the assessment score variances attributable to the latent variable common to all items (i.e., the general factor). The factor analysis further indicates that the assessment is not unidimensional, but unidimensionality would not be expected of a valid test of the topics covered (ordering and equivalence of fractions). Both face validity and content validity require that an

assessment for these topics not be unidimensional, but cover multiple related facets of learning and understanding. Coefficient  $\omega_t$  estimates the proportion of test variance attributable to all common factors. The three coefficients  $\omega_t$ ,  $\lambda_4$  and the glb provide the best estimates of internal consistency (Revelle & Zinbarg, 2009), and under these measures the assessment was observed to be highly reliable.

Table A-2 shows the inter-item-correlation matrix. Based on an internal-consistency-with-item-deletion analysis (recalculating  $\omega_t$  with each item deleted from the scale), all items appeared to be worthy of retention; the change in  $\omega_t$  that would result from the deletion of each item is shown in Table A-3. The only increase of  $\omega_t$  would come from discarding Item 2 and that would increase  $\omega_t$  by only .005 but would decrease the greatest lower bound reliability measure by .003. Item 1 (included as “warm-up problem” and confidence builder) conveys very little information and including or deleting it has minimal impact on any of the measures we report.

Item	Mean	Standard Deviation	Standard Error	Corr. Coef. to Score	Reasoning Level	Language Level
1	.896	.305	.0171	.269	1	1
2	.642	.480	.0269	.339	1	1
3	.664	.473	.0265	.613	1-2	2
4	.469	.500	.0280	.609	2	1
5	.686	.465	.0261	.624	2	2
6	.597	.491	.0275	.618	2	2
7	.453	.499	.0280	.566	2	2
8	.830	.376	.0211	.405	1	1
9	.409	.492	.0276	.593	2	1
10	.164	.370	.0208	.533	2	1
11	.469	.500	.0280	.540	2	1
12	.381	.486	.0273	.446	3	3
13	.544	.499	.0280	.605	2	1
14	.343	.475	.0267	.635	2	1

**Table A-1.** Descriptive statistics for the 14 individual summative assessment items showing mean score, standard deviation, standard error, and the Pearson’s correlation coefficient between the item and the cumulative assessment score. Response-outlying items for which 80% or more of tested subjects provided correct or incorrect answers are highlighted in red – of these, Item 1 was included as a confidence builder, Item 8 was a true/false style multiple choice item, and Item 10 was a challenging number line problem. Reasoning levels are language levels are only intended as a rough references (the values were not used in any of the analysis). The language levels are defined as 1: minimal english language demands, 2: moderately low demands, and 3: moderately high demands. The reasoning levels correspond to the taxonomy of Marzano and Kendall (2006) as estimated by a domain expert for each item.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	–													
2	.004	–												
3	.172	.217	–											
4	.113	.111	.349	–										
5	.080	.087	.406	.364	–									
6	.183	.162	.338	.372	.314	–								
7	.102	.140	.273	.222	.358	.322	–							
8	.038	.081	.192	.206	.181	.192	.210	–						
9	.157	.181	.308	.309	.398	.252	.233	.206	–					
10	.123	.082	.225	.300	.299	.259	.264	.155	.341	–				
11	.175	.032	.229	.318	.256	.244	.222	.190	.181	.386	–			
12	.033	.086	.147	.186	.126	.181	.237	.113	.178	.161	.238	–		
13	.082	.145	.350	.328	.359	.330	.237	.191	.299	.200	.240	.249	–	
14	.050	.125	.318	.278	.361	.390	.328	.185	.343	.290	.291	.294	.382	–
Score	.269	.339	.613	.609	.624	.618	.566	.405	.593	.533	.540	.446	.605	.635

**Table A-2.** Correlation matrix (inter-item matrix of Pearson’s correlation coefficients) along with item-to-score correlation coefficients.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$\Delta(\omega_i)$	+0.000	+0.005	-0.018	-0.018	-0.019	-0.019	-0.013	-0.004	-0.016	-0.012	-0.010	-0.003	-0.017	-0.021

**Table A-3.** Item deletion analysis of reliability – the change in coefficient  $\omega_i$  that would result from the deletion of each of the summative assessment items from the cumulative scale.

### Factor Analysis

We also conducted various factor analyses of the summative assessment. A principal component analysis yielded eigenvalues of:

4.12, 1.10, 1.04, 0.93, 0.89, 0.87, 0.82, 0.78, 0.69, 0.66, 0.59, 0.55, 0.49, 0.46.

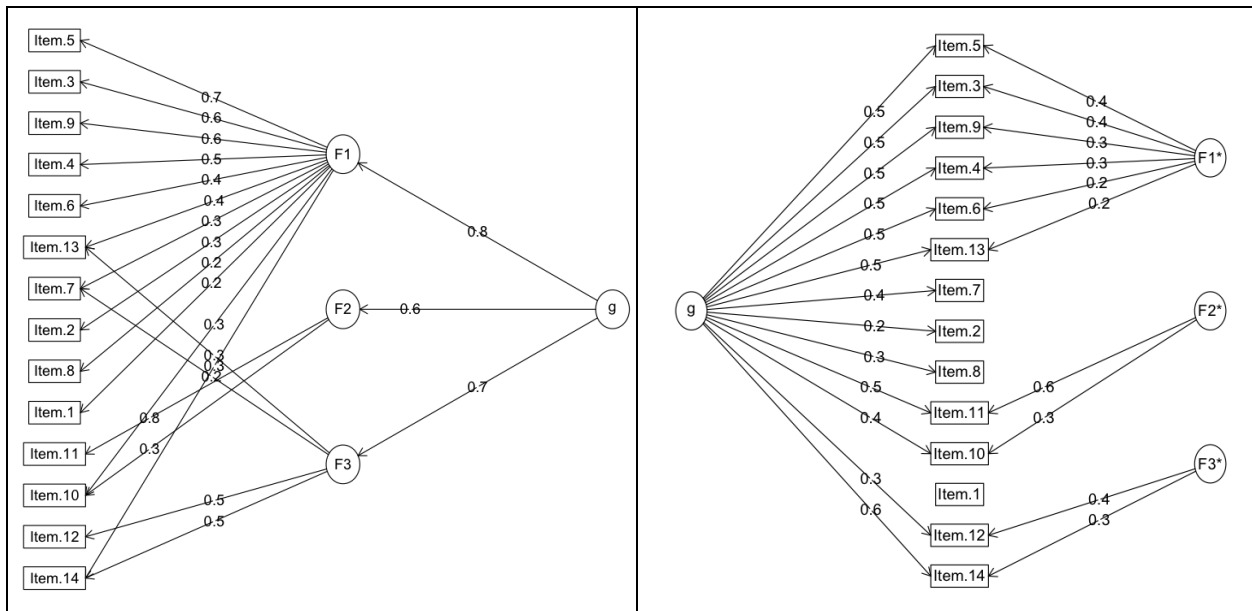
And the factor loadings for the first three components under this analysis are shown in Table A-4

The results of a hierarchical factor analysis are illustrated in Figure A-1 along with the Schmid-Leiman solution, for which factor loadings of the latter are shown in Table A-5. Note that this analysis identifies the two number line tasks (Items 10 and 11) as a subscale (factor F2 / F2\*). The hierarchical factor analysis did not identify a factor correlated to the higher-language-level items (nor did a repeated analysis restricted to the ELL subsample), so whereas any mathematical assessment of this sort also to some extent measures language, that effect appears to have been minimized in the case of this assessment.



Item	PC1 ( $\lambda=4.12$ )	PC2 ( $\lambda=1.10$ )	PC3 ( $\lambda=1.04$ )	$h^2$	$u^2$
1	+.25	+.52	+.56	.64	.36
2	+.27	-.61	+.26	.52	.48
3	+.62	-.17	+.32	.52	.48
4	+.62			.40	.60
5	+.65		+.12	.45	.55
6	+.63		+.12	.41	.59
7	+.56		-.13	.34	.66
8	+.39		-.10	.17	.83
9	+.60		+.18	.40	.60
10	+.56	+.35		.44	.56
11	+.53	+.48	-.22	.57	.43
12	+.41		-.59	.52	.48
13	+.61	-.19		.41	.59
14	+.65	-.11	-.25	.50	.50

**Table A-4.** Per-item factor loadings (standardized, pattern matrix, based on correlation matrix) for the first three principal component analysis factors. Loadings smaller than .10 are not shown. The leftmost columns give the the communality and the unique variance for each of the items.



**Figure A-1.** Graphs illustrating the results of the hierarchical factor analysis (left) and the Schmid-Leiman solution (right).

Item	g ( $\lambda=2.67$ )	F1* ( $\lambda=0.72$ )	F2* ( $\lambda=0.56$ )	F3* ( $\lambda=0.34$ )
1				
2	+.20			
3	+.49	+.36		
4	+.49	+.28		
5	+.53	+.38		
6	+.50	+.25		
7	+.44			
8	+.30			
9	+.47	+.32		
10	+.44		+.28	
11	+.47		+.64	
12	+.34			+.36
13	+.50	+.23		
14	+.56			+.33

**Table A-5.** Schmid-Leiman factor loadings (where greater than .20) for all items.

## Appendix B: Summative Assessment Items

Each of the 14 items from the summative assessment are described below with two sample parameterizations, screen captures of participant student work, and the item design goals.

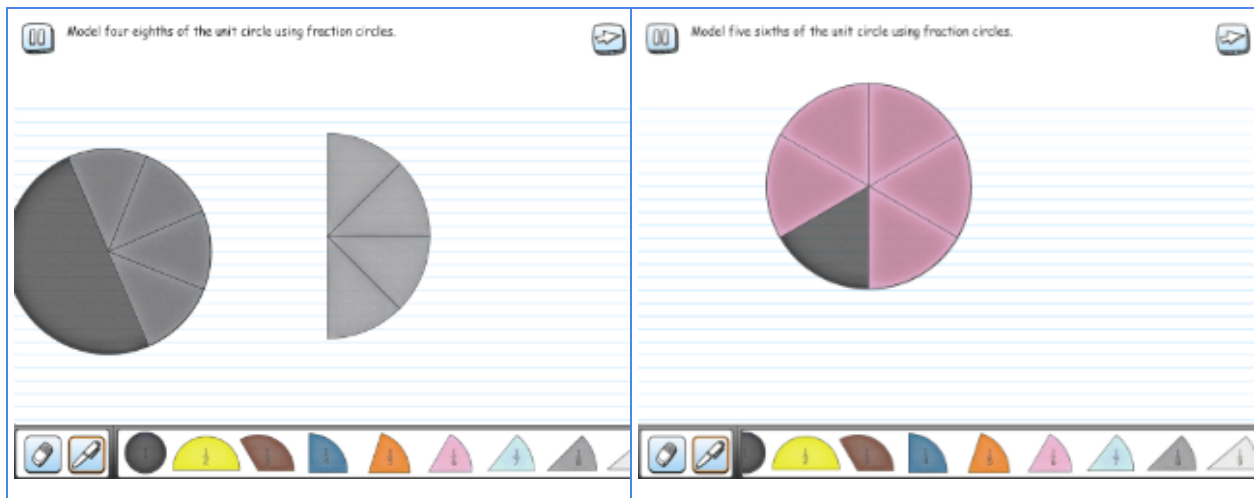
### Item 1: What fraction of the bar is shaded?

**Goal:** Build the student's confidence with a warm-up task; assess student's understanding of part-whole construct using area model; fraction notation; and answer entry.

### Item 2: Do these shapes have $\frac{1}{4}$ shaded? or: Do these shapes have $\frac{3}{4}$ shaded?

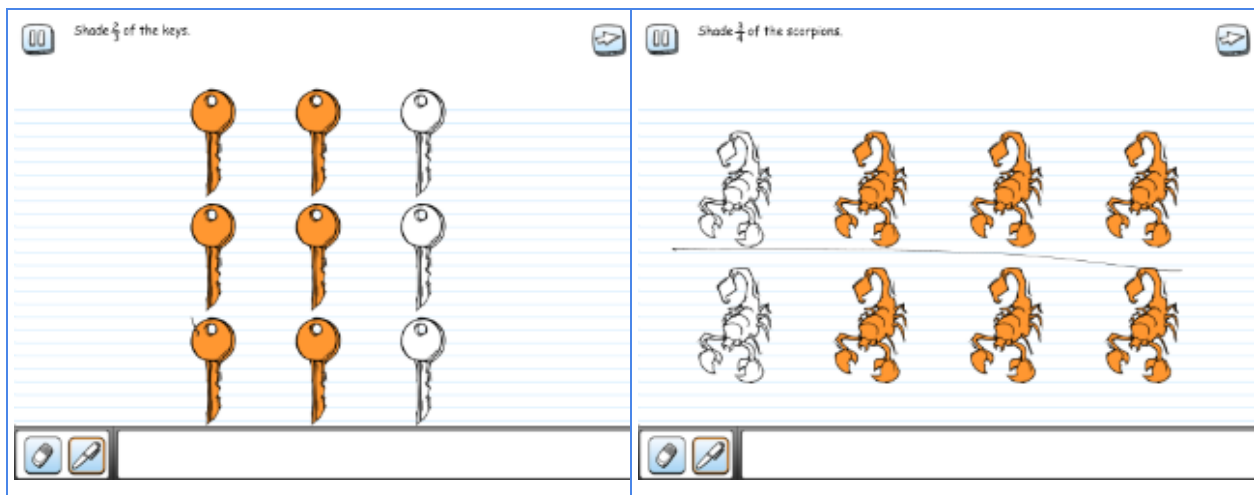
**Goal:** Assess student's understanding of "equal" parts of the whole - equal area partitions with two differently-shaped area models.

**Item 3:** Model four eighths of the unit circle using fraction circles.  
 or: Model five sixths of the unit circle using fraction circles.  
 etc.



**Goal:** Assess student's understanding of part-whole construct and ability to model a fraction given a fraction word name.

**Item 4:** Shade  $\frac{2}{3}$  of the keys.  
 or: Shade  $\frac{3}{4}$  of the scorpions.  
 etc.



**Goal:** Assess student's ability to effectively partition sets of objects where the number of objects is a multiple of the specified denominator.

**Item 5:** Eight [five, six, etc.] people want to share an apple pie [pizza, etc] equally.

- (1) Draw a picture of the apple pie showing how you would divide it up equally.
- (2) What fraction of the apple pie will each person get?

The image shows two side-by-side screenshots of a digital learning interface. Each screenshot contains a math problem and a student's handwritten response.

**Left Screenshot:**  
 Problem: "Eight people want to share an apple pie equally. (1) Draw a picture of the apple pie showing how you would divide it up to share equally. (2) What fraction of the apple pie will each person get?"  
 Answer: The student has drawn a circle representing a pie, divided into 8 equal sectors. Next to the drawing, the student has written "my Pie 1/8!". The interface shows the answer  $\frac{1}{8}$  in a box.

**Right Screenshot:**  
 Problem: "Five people want to share a pizza equally. (1) Draw a picture of the pizza showing how you would divide it up to share equally. (2) What fraction of the pizza will each person get?"  
 Answer: The student has drawn a circle representing a pizza, divided into 5 equal sectors. The interface shows the answer  $\frac{1}{5}$  in a box.

**Goal:** Assess a student ability to model and partitioning object in real world scenario; effective use of halving strategy

**Item 6:** What fraction of the word "INSPIRE" is made up of the letter I?

or: What fraction of the word "FANTASTIC" is made up of the letter A?

etc.

Multiple choice among (e.g.):  $\frac{2}{5}$ , AA,  $\frac{2}{7}$ , 2

The image shows two side-by-side screenshots of a digital learning interface. Each screenshot contains a multiple-choice math problem and a student's handwritten response.

**Left Screenshot:**  
 Problem: "What fraction of the word 'INSPIRE' is made up of the letter I?"  
 Multiple choice options:  $\frac{1}{8}$ ,  $\frac{1}{11}$ ,  $\frac{2}{7}$ , 2.  
 Answer: The student has selected  $\frac{2}{7}$ .


**Right Screenshot:**  
 Problem: "What fraction of the word 'FANTASTIC' is made up of the letter A?"  
 Multiple choice options: 2,  $\frac{2}{7}$ ,  $\frac{2}{9}$ , AA.  
 Answer: The student has written "FANTASTIC" and  $\frac{2}{9}$ .

**Goal:** Understanding part-whole construct using sets; abstraction of set model in a word problem.

**Item 7:** The picture shows two fifths [two thirds, etc.] of a candy bar.


- (1) Draw a picture of the whole candy bar.
- (2) Enter the number of pieces that make up the whole candy bar.

**Item 7:** The picture shows two fifths of a candy bar.  
 (1) Draw a picture of the whole candy bar.  
 (2) Enter the number of pieces that make up the whole candy bar.



Answer:

**Item 7:** The picture shows three fifths of a protein bar.  
 (1) Draw a picture of the whole protein bar.  
 (2) Enter the number of pieces that make up the whole protein bar.

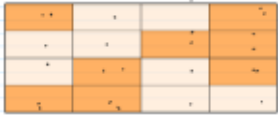


Answer:

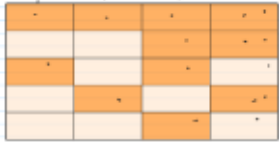
**Goal:** Assess student's ability to recreate the whole from fraction unit.

**Item 8:** Is  $\frac{1}{2}$  of the rectangle shaded?

**Item 8:** Is  $\frac{1}{2}$  of the rectangle shaded?  
 Yes  
 No



**Item 8:** Is  $\frac{1}{2}$  of the rectangle shaded?  
 Yes  
 No



11

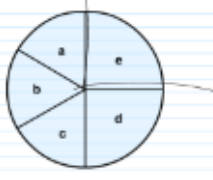
$4 \times 5 = 20$

**Goal:** Assess student's understanding equal parts - part of whole don't have to be contiguous, and their understanding of equivalence.

**Item 9:** What fraction of the circle is “e” [“a”, etc.]?

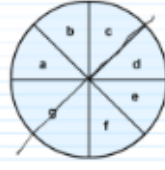
What fraction of the circle is 'e'?

Answer:  $\frac{1}{4}$



What fraction of the circle is 'g'?

Answer:  $\frac{2}{8}$

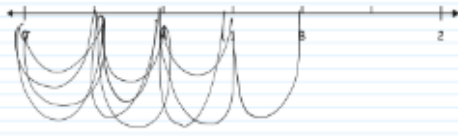


**Goal:** Assess student’s understanding equal parts of the whole and identifying different partitioning of area model. Some students see number of parts - not equal parts.

**Item 10:** Write the fraction represented by each letter on the number line.

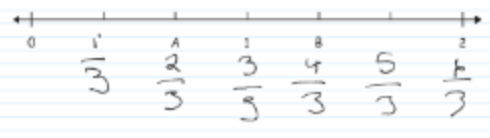
Write the fraction represented by each letter on the number line.

A =  $\frac{2}{3}$    B =  $\frac{4}{3}$







Write the fraction represented by each letter on the number line.

A =  $\frac{2}{3}$    B =  $\frac{4}{3}$



**Goal:** Assess student’s ability to extend part-whole thinking to a linear model. Do they understand how to use the tick marks and equal partitioning.



**Item 11:** Drag the fraction to the correct location on the number line.

<p>Drag the fraction to the correct location on the number line.</p> 	<p>Drag the fraction to the correct location on the number line.</p> 
	

**Goal:** Assess student's ability to extend part-whole thinking to a linear model. Do they understand how to use the tick marks and equal partitioning.

**Item 12:** Check the statement that is true:

- $\frac{3}{8}$  is less than  $\frac{1}{2}$  because the difference between 1 and 3 is two while the difference between 3 and 8 is five
- $\frac{3}{8}$  is greater than  $\frac{1}{2}$  because 3 is greater than 1 and 8 is greater than 2
- $\frac{3}{8}$  is less than  $\frac{1}{2}$  because to be a half you need 4 eighths and 3 is less than 4
- $\frac{3}{8}$  is less than  $\frac{1}{2}$  because eighths are smaller parts than halves

<p>Check the statement that is true:</p> <p><input type="radio"/> <math>\frac{3}{8}</math> is less than <math>\frac{1}{2}</math> because the difference between 1 and 3 is two while the difference between 3 and 8 is five</p> <p><input type="radio"/> <math>\frac{3}{8}</math> is greater than <math>\frac{1}{2}</math> because 3 is greater than 1 and 8 is greater than 2</p> <p><input checked="" type="radio"/> <math>\frac{3}{8}</math> is less than <math>\frac{1}{2}</math> because to be a half you need 4 eighths and 3 is less than 4</p> <p><input type="radio"/> <math>\frac{3}{8}</math> is less than <math>\frac{1}{2}</math> because eighths are smaller parts than halves</p>	<p>Check the statement that is true:</p> <p><input type="radio"/> <math>\frac{4}{10}</math> is less than <math>\frac{1}{2}</math> because tenths are smaller parts than halves</p> <p><input checked="" type="radio"/> <math>\frac{4}{10}</math> is less than <math>\frac{1}{2}</math> because to be a half you need 5 tenths and 4 is less than 5</p> <p><input type="radio"/> <math>\frac{4}{10}</math> is less than <math>\frac{1}{2}</math> because the difference between 1 and 4 is three while the difference between 4 and 10 is six</p> <p><input type="radio"/> <math>\frac{4}{10}</math> is greater than <math>\frac{1}{2}</math> because 4 is greater than 1 and 10 is greater than 2</p>
	

**Goal:** Assess student's ability to compare a fraction to the benchmark fraction  $\frac{1}{2}$ . An understanding of equivalence is needed.



**Item 13:** Enter the smallest [greatest] fraction.

**Goal:** Assess student's ability to compare three fractions using two strategies for ordering fractions – common numerator and common denominator – transitively.

**Item 14:** Enter an equivalent fraction to  $\frac{2}{3}$ . Justify your answer below.

**Goal:** Assess student's ability to calculate an equivalent fraction and model equivalent fractions using area models.

## References

- Atkinson, R. C. (1974). Adaptive instructional systems: Some attempts to optimize the learning process. *Cognition and Instruction*, Klahr, D. ed., Erlbaum Associates, 335-362.
- Benjamin, L. T. (1988). A history of teaching machines. *American psychologist*, 43(9), 703.
- Black, P., & William, D. (1998). *Inside the black box: Raising standards through classroom assessment*. Granada Learning.
- Black, P. (2004). *Working inside the black box: Assessment for learning in the classroom*. Granada Learning.
- Cramer, K. A., Post, T. R., & delMas, R. C. (2002). Initial fraction learning by fourth-and fifth-grade students: A comparison of the effects of using commercial curricula with the effects of using the rational number project curriculum. *Journal for Research in Mathematics Education*, 111-144.
- Cramer, K., Behr, M., Post T., Lesh, R., (2009) Rational Number Project: Initial Fraction Ideas. Originally published in 1997 as Rational Number Project: Fraction Lessons for the Middle Grades - Level 1, Kendall/Hunt Publishing, Dubuque Iowa.
- Cramer, K., & Wyberg, T. (2009). Efficacy of different concrete models for teaching the part-whole construct for fractions. *Mathematical thinking and learning*, 11(4), 226-257.
- Cumming, G., & McDougall, A. (2000). Mainstreaming AIED into education? *International Journal of Artificial Intelligence in Education (IJAIED)*, 11, 197-207.
- Daro, P., Mosher, F. A., & Corcoran, T. (2011). Learning Trajectories in Mathematics: A Foundation for Standards, Curriculum, Assessment, and Instruction. CPRE Research Report# RR-68. *Consortium for Policy Research in Education*.
- de Winter, J. C., & Dodou, D. (2010). Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11), 1-12.
- Desmarais, M. C., & d Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1-2), 9-38.
- Ferster, B. (2014). *Teaching Machines: Learning from the Intersection of Education and Technology*. Tech.Edu: a Hopkins Series on Education and Technology. In Press.
- Fosnot, C. T. (2005). Constructivism revisited: Implications and reflections. *The Constructivist*, 16(1), 1-17.
- Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). Assisting Students Struggling with Mathematics: Response to Intervention (RtI) for Elementary and Middle Schools. NCEE 2009-4060. Institute of Education Sciences, U.S. Department of Education.
- Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. *APA educational psychology handbook: Vol. 3. Application to learning and teaching*, American Psychological Association, 451-473.
- Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., & Wayman, J. C. (2009). Using Student Achievement Data to Support Instructional Decision Making. IES Practice Guide. NCEE 2009-4067. *National Center for Education Evaluation and Regional Assistance*.
- Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1), 81-112.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, 6(2), 107-128.
- Izumi, L., Fathers, F. and Clemens, J (2013). *Technology and Education: A Primer*. The Fraser Institute, Retrieved from <http://www.fraserinstitute.org/research-news/display.aspx?id=20268>
- Jameson, A. (2009). Adaptive interfaces and agents. *Human-Computer Interaction: Design Issues, Solutions, and Applications*, Sears, A. & Jacko, J. A. (Eds.). CRC Press, 105-130.

- Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled clinical trials*, 21(3), 167-189.
- Lord, F. M. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement*, 31, 3-31.
- Meyer, D. (2014, June 30). Personalized Learning Software: Fun Like Choosing Your Own Ad Experience. [Blog post]. Retrieved from <http://blog.mrmeyer.com>.
- Mislevy, R. J., Senturk, D., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., & Yan, D. (2002). *Modeling conditional probabilities in complex educational assessments*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.
- National Mathematics Advisory Panel. (2008). Foundations for Success: The Final Report of the National Mathematics Advisory Panel, U.S. Department of Education: Washington, DC.
- National Research Council. (2001). Adding it up: Helping children learn mathematics. J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70(350), 351-356.
- Petit, M. M., Laird, R. E., & Marsden, E. L. (2010). *A focus on fractions*. Taylor & Francis.
- Project Tomorrow. (2011). Speak Up: Learning in the 21st Century: Mobile Devices + Social Media = Personalized Learning.
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74(1), 145-154.
- Revelle, W. (2014, May 14). Manual for R package 'psych': A package for personality, psychometric, and psychological research (ver. 1.4.5). Retrieved from <http://www.personality-project.org/r/psych/psych-manual.pdf>
- Rothen, W., & Tennyson, R. D. (1978). Application of Bayes' theory in designing computer-based adaptive instructional strategies. *Educational Psychologist*, 12, 317-323.
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*, 8(1), 18.
- Seethaler, P. M., Fuchs, L. S., Star, J. R., & Bryant, J. (2011). The cognitive predictors of computational skill with whole versus rational numbers: An exploratory study. *Learning and individual differences*, 21(5), 536-542.
- Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. *Adaptive technologies for training and education*, 7-27.
- Siegler, R., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., Thompson, L., & Wray, J. (2010). Developing effective fractions instruction for kindergarten through 8th grade: A practice guide (NCEE #2010-4039). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from [whatworks.ed.gov/publications/practiceguides](http://whatworks.ed.gov/publications/practiceguides).
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107-120.
- Skinner, B. F. (1958). Teaching Machines. *Science*, 128, 969-977.
- Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K-12 students' mathematical learning. *Journal of Educational Psychology*, 105(4), 970-987.

- Stone, E., & Davey, T. (2011). *Computer-adaptive testing for students with disabilities: A review of the literature*. Research Report 11–31). Princeton, NJ: Educational Testing Service.
- van der Linden, W. J. & Glas C.A.W. (Eds.). (2010). *Elements of adaptive testing* (pp. 3-30). New York, NY: Springer.
- Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd Edition). New York, NY: Routledge.
- Webb, D. C. (2004). Enriching Assessment Opportunities Through Classroom Discourse. In T. A. Romberg (Ed.), *Standards-Based Mathematics Assessment in Middle School: Rethinking Classroom Practice* (pp. 169-187). New York: Teachers College Press.
- Webb, D. C., Boswinkel, N., & Dekker, T. (2008). Beneath the Tip of the Iceberg: Using Representations to Support Student Understanding. *Mathematics teaching in the middle school*, 14(2), 110-113.
- West, P., Rutstein, D. W., Mislavy, R. J., Liu, J., Choi, Y., Levy, R., ... & Behrens, J. T. (2010). A Bayesian Network Approach to Modeling Learning Progressions and Task Performance. CRESST Report 776. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.
- Weiss, D. J. (1976). Adaptive testing research in Minnesota: Overview, recent results, and future directions. In *Proceedings of the first conference on computerized adaptive testing*, 24-35.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied psychological measurement*, 6(4), 473-492.
- What Works Clearinghouse (2014). What Works Clearinghouse procedures and standards handbook (ver. 3.0.1). *Washington, DC*: Institute of Education Sciences, U.S. Department of Education. Retrieved August 26, 2014.
- The Council of Economic Advisors. (2011). *Unleashing the Potential of Education Technology*. Washington DC, The White House.
- William, D. (2006). *Mathematics inside the black box: Assessment for learning in the mathematics classroom*. Granada Learning.
- William, D. (2011). *Embedded Formative Assessment*. Solution Tree.
- Woodward, J., Beckmann, S., Driscoll, M., Franke, M., Herzig, P., Jitendra, A., ... & Ogbuehi, P. (2012). *Improving mathematical problem solving in grades 4 through 8: A practice guide* (NCEE 2012-4055). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Zuckerman, M., Porac, J., Lathin, D., & Deci, E. L. (1978). On the importance of self-determination for intrinsically-motivated behavior. *Personality and Social Psychology Bulletin*, 4(3), 443-446.