(2017) In (Eds.) A. Weinberg, C. Rasmussen, J. Rabin, M. Wawro, and S. Brown (Eds.), Proceedings of the 20th Annual Conference on Research in Undergraduate Mathematics Education (pp. 638-645). San Diego, CA.

Exploration of the Factors that Support Learning: Web-based Activity and Testing Systems in Community College Algebra

[Contributed Report]

Shandy Hauk, WestEd
Bryan Matlen, WestEd
Larry Thomas, WestEd

A variety of computerized interactive learning platforms exist. Most include instructional supports in the form of problem sets. Feedback to users ranges from a single word like "Correct!" to offers of hints and partially- to fully-worked examples. Behind-the-scenes design of systems varies as well – from static dictionaries of problems to "intelligent" and responsive programming that adapts assignments to users' demonstrated skills within the computerized environment. This report presents background on digital learning contexts and early results of a cluster-randomized controlled trial study in community college elementary algebra classes where the intervention was a particular type of web-based activity and testing system.

Key words: Adaptive Tutoring System, College Algebra, Multi-site Cluster Randomized Controlled Trial

Many students arrive in college underprepared for college level algebra, despite its importance for future success in mathematics (Long, Iatarola, & Conger, 2009; Porter & Polikoff, 2012). Web-based Activity and Testing Systems (WATS) are one approach to supporting equity and excellence in mathematics learning in colleges. When it comes to technology and algebra learning in college, what works? For whom? Under what conditions? These ubiquitous questions plague educational researchers who are assessing the whats, whys, and hows of a technology intervention or addition to a course. Did the instructors have enough support to adequately implement the technology tool? Were the materials adequate to provide enough practice hours for students? Was instruction sufficient to prepare students to pass the final exam?

This preliminary report offers early results from a large project investigating relationships among student achievement and varying conditions of implementation for a web-based activity and testing system (WATS) used in community college algebra. Implementing a particular WATS constitutes the "treatment" condition in this cluster randomized controlled trial study. As described below, there are several ways to distinguish WATS tools. Some systems, like the one at the heart of our study, include adaptive problem sets, instructional videos, and data-driven tools for instructors to use to monitor and scaffold student learning.

Research Questions

Funded by the U.S. Department of Education, we are conducting a large-scale mixed methods study in over 30 community colleges. The study is driven by two research questions:

Research Question 1: What is the impact of a particular digital learning platform on students' algebraic knowledge after instructors have implemented the platform for two semesters?

Research Question 2: What challenges to use-as-intended (by developers) are faculty encountering and how are they responding to the challenges as they implement the learning tool?

Background and Conceptual Framing

First, there are distinctions among cognitive, dynamic, and static learning environments (see Table 1). Web-based Activity and Testing System (WATS) learning environments can vary along at least two dimensions: (1) the extent to which they adaptively respond to student behavior and (2) the extent to which they are based on a careful cognitive model.

Table 1. *Conceptual Framework of the Types of Instruction Based on Adaptability and Basis in a Theory of Learning*

		Static	Dynamic
Is a particular model of learning explicit in design and implementation (structure and processes)?	No	Text and tasks with instructional adaptation external to the materials	Adaptive tutoring systems (Khan Academy, ALEKS, ActiveMath)
	Yes	Textbook design and use driven by fidelity to an explicit theory of learning	"Intelligent" tutoring systems (Cognitive Tutor)

Static learning environments are those that are non-adaptive without reliance on an underlying cognitive model – they deliver content in a fixed order and contain scaffolds or feedback that are identical for all users. The design may be based on intuition, convenience, or aesthetic appeal. An example of this type of environment might be online problem sets from a textbook that give immediate feedback on accuracy to students (e.g., "Correct" or "Incorrect").

Dynamic learning environments keep track of student behavior (e.g., errors, error rates, or time-on-problem) and use this information in a programmed decision tree that selects problem sets and/or feedback based on students' estimated mastery of specific skills. An example of a dynamic environment might be a system such as ALEKS or the "mastery challenge" approach now used at the online Khan Academy. For example, at khanacadmy.org a behind-the-scenes data analyzer captures student performance on a "mastery challenge" set of items. Once a student gets six items in a row correct, the next level set of items in a programmed target learning trajectory is offered. Depending on the number and type of items the particular user answers incorrectly (e.g., on the path to six items in a row done correctly), the analyzer program identifies target content and assembles the next "mastery challenge" set of items.

Above and beyond such responsive assignment generation, programming in a "cognitively-based" dynamic environment is informed by a theoretical model that asserts the cognitive processing necessary for acquiring skills (Anderson et al. 1995; Koedinger & Corbett, 2006). For example, instead of specifying only that graphing is important and should be practiced, a cognitively-based environment also will specify the student thinking and skills needed to comprehend graphing (e.g., connecting spatial and verbal information), and provide feedback and scaffolds that support these cognitive processes (e.g., visuo-spatial feedback and graphics that are integrated with text). In cognitively-based environments, scaffolds themselves can also be adaptive (e.g., more scaffolding through examples can be provided early in learning and scaffolding can be faded as a student acquires expertise; Ritter et al., 2007). Like other dynamic systems, cognitively-based systems can also provide summaries of student progress, which better enable teachers to support struggling students.

No fully tested cognitively-based system currently exists for college students learning algebra. As mentioned, several dynamic systems do exist (e.g., ALEKS, Khan Academy "Missions"). The particular WATS investigated in our study is accessed on the internet and is

designed primarily for use as replacement for some in-class individual seatwork and some homework. ***Note***: We report here on data collected from the first of two years. The second year of the study – which repeats the design of the first – is currently underway. Hence, we purposefully under-report some details.

Method

The study we report here is a multi-site cluster randomized trial. Half of instructors at each community college site are assigned to use a particular WATS in their instruction (treatment condition), the other half teach as they usually would, barring the use of the Treatment WATS tool (control condition). In addition, faculty participate for two semesters in order to allow instructors to familiarize themselves with implementing the WATS with their local algebra curriculum. Specifically, the Fall semester is a "field" semester to field-test the intervention and the Spring semester of the same academic year is the full "efficacy" study.

Using a stratified sampling approach to recruitment, we first conducted a cluster analysis on all 113 community college sites eligible to participate in the study (e.g., those offering semester-long courses in elementary algebra that met at least some of the time in a physical classroom or learning/computer lab). The cluster analysis was based on college-level characteristics that may be related to student learning (e.g., average age of students at the college, the proportion of adjunct faculty, etc.). This analysis led to five clusters of colleges. Our recruitment efforts then aimed to include a proportionate number of colleges within each cluster. The primary value of this approach is that it allows more appropriate generalization of study findings to the target population (Tipton, 2014). The first cohort of participants was a sample of 38 colleges similar to the overall distribution across clusters that was the target for the sample (see Figure 1).

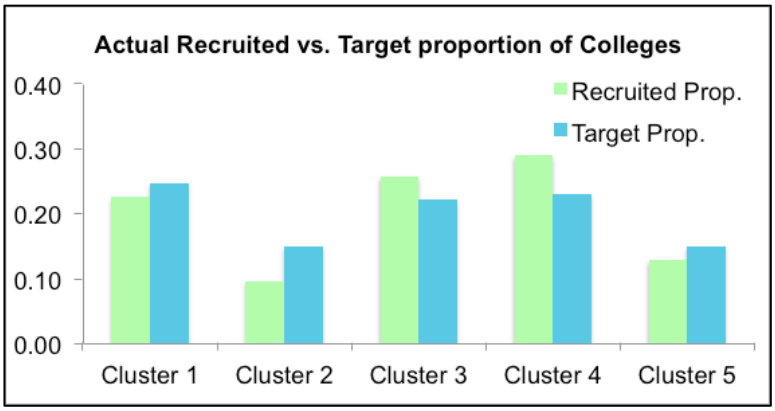

Figure 1. Recruited sample proportions and target sample proportions across clusters.

Sample for this Report

Initial enrollment in the study included 89 teachers across the 38 college sites. For this report on early results, we have used the data from 510 students of 29 instructors across 18 colleges. Student and teacher numbers related to the data reported on here are shown in Table 2.

Table 2. *Counts of Teachers, Students, and Colleges in the Study*

Condition	*Teachers*	*Students*	*Colleges*
Control	17	328	13
Treatment	12	182	11
Total	*29*	*510*	*18*

Results

The primary outcome measure for students' performance is an assessment from the Mathematics Diagnostic Testing Program (MDTP), which is a valid and reliable assessment of students' algebraic knowledge (Gerachis & Manaster, 1995). The primary aim of the quantitative analysis was to address Research Question 1, what is the impact of WATS use on students' outcomes? To this end, we employed Hierarchical Linear Modeling (HLM) (Raudenbush & Bryk, 1998) to predict students' end of semester MDTP scores. The HLM model includes a random effect of teacher to account for the nesting of students within instructors, and covariates that account for students' pretest MDTP scores at both student and teacher levels (i.e., student scores are aggregated at the teacher level; covariates were grand mean centered to achieve the intended covariate-adjustment). Importantly, in the model below, $WATS_j$ represents a dichotomous variable (dummy coded) indicating treatment assignment, and the main effect of the intervention is captured by β_{01}.

Model

$$\gamma_{ij} = \beta_{00} + \beta_{01}\boldsymbol{WATS}_j + \beta_{10}\boldsymbol{StuPre}_{ij} + \beta_{02}\boldsymbol{TeaPre}_j + \xi_{0J}\boldsymbol{Tea}_j + \varepsilon_{I0}\boldsymbol{Stu}_{ij}$$

The random and fixed effects for the model presented above are displayed in Tables 3 and 4, respectively.

Table 3. *Random Effects of the Model*

	Variance	*Standard Deviation*
Teacher ξ_{0J}	6.95	2.64
Level-1 Error ε_{I0}	37.69	6.14

Table 4. *Fixed Effect Results of the Model*

Variable	*B*	*Standard Error*	*p-value*
Intercept β_{00}	21.98	0.74	< .001
WATS β_{01}	2.59	1.17	.04
StuPre β_{10}	0.54	0.04	< .001
TeaPre β_{02}	0.35	0.17	.05

Controlling for students' pretest scores, we found that using WATS corresponded to a 2.59 point increase in students' post-test scores, a statistically significant positive effect ($p < .05$). Since the post-test is out of a 50 point total, the estimate corresponds to about 5 percentage points greater post-test score, on average, for treatment group students (2.59/50). The Hedges *g* value for this effect is 0.32, which is considered a small but noteworthy effect in educational research for studies of this size (Cheung & Slavin, 2015; Hill et al. 2008). The 95% confidence interval of the Hedges *g* value is .14 - .50.

We note this study suffered from high instructor attrition, which may bias the outcome of results. To investigate the robustness of the findings above, we are in the process of repeating this study with a second cohort of participants during the current (2016-17) academic year. Pooling the results of these two studies will help to determine the extent to which study results replicate with different populations. In this same vein, we plan to reanalyze the results using

post-test scores that are estimated using item response theory (IRT). IRT is a measurement approach that takes into consideration potential differences in item characteristics when scoring individuals and places scores on a continuous metric. The use of IRT will allow us to take into consideration the difficulty and discriminability of items and represent these in the calculation of post-test scores, which can then be analyzed using the model presented above.

To address Research Question 2, a great deal of textual, observational, and interview data were gathered last year (and will be gathered again for the second iteration of the study). These data allow careful analysis of the intended and actual use of the learning environment and the classroom contexts in which it is enacted – an examination of implementation structures and processes. Indices of specific and generic fidelity derived from this work also will play a role in HLM generation and interpretation in the coming year.

As in many curricular projects, developers of the WATS in our study paid attention to learning theory in determining the content in the web-based system, but the same was not true for determining implementation processes and structures. The pragmatic details of large-scale classroom use were under-specified. Developers articulated their assumptions about what students learned as they completed activities, but the roles of specific components, including the instructor role in the mediation of learning, were not clearly defined. Thus, there was an under-determined "it" to which developers expected implementers (instructors and students) to be faithful.

Fidelity of implementation is the degree to which an intervention or program is delivered as intended (Dusenbury, Brannigan, Falco, & Hansen, 2003). Do implementers understand the trade-offs in the daily decisions they must make "in the wild" and the short and long-term consequences on student learning as a result of compromises in fidelity? As Munter and colleagues (2014) have pointed out, there is no agreement on how to assess fidelity of implementation. However, there is a growing consensus on a component-based approach to measuring its structure and processes (Century & Cassata, 2014). Century and Cassata's summary of research offers five components to consider in fidelity of implementation: Diagnostic, Procedural, Educative, Pedagogical, and Student Engagement (see Table 5).

Table 5. *Components and Focus in a Fidelity of Implementation Study*

Components	*Focus*
Diagnostic	These factors say what the "it" is that is being implemented (e.g., what makes this particular WATS distinct from other activities).
Structural-Procedural	These components tell the user (in this case, the instructor) what to do (e.g., assign intervention *x* times/week, *y* minutes/use). These are aspects of the *expected* curriculum.
Structural-Educative	These state the developers' expectations for what the user needs to know relative to the intervention (e.g., types of technological, content, and pedagogical knowledge needed by an instructor).
Interaction-Pedagogical	These capture the actions, behaviors, and interactions users are expected to engage in when using the intervention (e.g., intervention is at least *x* % of assignments, counts for at least *y* % of student grade). These are aspects of the *intended* curriculum.
Interaction-Engagement	These components delineate the actions, behaviors, and interactions that students are expected to engage in for successful implementation. These are aspects of the *achieved* curriculum.

The components in Table 5 are operationalized through a rubric, the guide for collecting and reporting data in our implementation study. A rubric articulates the expectations for a category by listing the criteria, or what counts, and describes the levels of quality from low to high. Each component has several factors that define the component. The research team has developed a rubric for fidelity of implementation that identify measurable attributes for each component (for example, see Table 6 for some detail on the "educative" component).

Table 6. *Example Rubric Descriptors for Levels of Fidelity, Structural-Educative Component.*

Educative: These components state the developers' expectations for what the user (instructor) needs to know relative to the intervention.

	High Level of Fidelity	*Moderate Fidelity*	*Low Level of Fidelity*
Users' proficiency in math content	Instructor is proficient to highly proficient in the subject matter.	Instructor has some gaps in proficiency in the subject matter.	Instructor does not have basic knowledge and/or skills in the subject area.
Users' proficiency in TPCK	Instructor regularly integrates content, pedagogical, and technological knowledge in classroom instruction. Communicates with students through WATS.	Instructor struggles to integrate CK, PK, and TK in instruction. Occasionally sends digital messages to students using WATS tools.	Instructor CK, PK, and/or TK sparse or applied in a haphazard manner in classroom instruction. Rarely uses WATS tools to communicate with students.
Users' knowledge of requirements of the intervention	Instructor understands philosophy of WATS resources (practice items, "mastery mechanics," analytics, and coaching tools),	Instructor understanding of the philosophy of WATS tool has some gaps. NOTE: Disagreeing is okay, this is about instructor knowledge of it.	Instructor does not understand philosophy of WATS resources. NOTE: Disagreeing is okay, this is about instructor knowledge of it.
Users' knowledge of requirements of the intervention	Instructor understands the purpose, procedures, and/or the desired outcomes of the project (i.e., "mastery")	Instructor understanding of project has some gaps (e.g., may know purpose, but not all procedures, or desired outcomes).	Instructor does not understand the purpose, procedures, and/or desired outcomes. Problems are typical.

Defining and Refining Measures for the Fidelity of Implementation Rubric

The ultimate purpose of a fidelity of implementation rubric is to articulate how to determine what works, for whom, under what conditions. In addition to allowing identification of alignment between developer expectations and classroom enactment, it provides the opportunity to discover where productive adaptations may be made by instructors, adaptations that boost student achievement beyond that associated with an implementation faithful to the developers' view.

In using the rubric, we assign a number to each level of fidelity. This can be as simple as a 3 for a high level of fidelity, 2 for a moderate level of fidelity, or a 1 for a low level; or the items can be weighted. The general score for the intervention will be the total number of points assigned in completing the rubric as a ratio of the total possible, across all instructors. It will also be possible to create a fidelity of implementation score on each row for each instructor – these data will be used in statistical modeling of the impact of the intervention as part of a "specific

fidelity index" (Hulleman & Cordray, 2009). We first total points for the item, then the component, and finally all components for a single score as an index of implementation.

Next Steps

In upcoming work, we will analyze a host of data on students, teachers, and colleges that may influence learning with WATS, including issues of feasibility of use in differing contexts, and measures associated with the nature of alignment or "fidelity" of implementation to WATS developers' expectations. Such analysis will help to inform important questions such as how and for whom WATS are most effective.

As indicated above, we will continue this study with a second cohort of new participants who will repeat the year-long study in the 2016-2017 academic year. Also, between now and the conference we will do more complex modeling of the data, with the introduction of IRT-informed scores and specific fidelity indices. Our specific objectives in the coming six months are to (1) continue analyses from the Spring 2016 efficacy study, and (2) conduct the field-test semester of the study with second cohort of participants.

Acknowledgement

This project is supported by a grant from the U.S. Department of Education, Institute of Education Sciences (IES-R305A-140340). Any opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the views of the Federal Government.

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*(2), 167-207.

Cheung, A., & Slavin, R.E. (2015, September). *How methodological features affect effect sizes in education.* Baltimore, MD: Johns Hopkins University, Center for Research and Reform in Education. Source: http://www.bestevidence.org/methods/methods.html

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research*, *18*(2), 237-256.

Gerachis, C., & Manaster, A. (1995). *User manual*. Mathematics Diagnostic Testing Project, California State University/University of California. Retrieved from http://mdtp.ucsd.edu/approvalstatus.shtml

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2(3),* 172- 177.

Hulleman, C. S., & Cordray, D. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Intervention Effectiveness*, *2*(1), 88-110.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *Cambridge Handbook of the Learning Sciences* (pp. 61-78). New York: Cambridge University Press.

Koedinger, K.R., & Sueker, E.L.F. (1996). PAT goes to college: Evaluating a cognitive tutor for developmental mathematics. In *Proceedings of the Second International Conference on the*

Learning Sciences (pp. 180–87). Charlottesville, VA: Association for the Advancement of Computing in Education.

Long, M. C., Iatarola, P., & Conger, D. (2009). Explaining gaps in readiness for college-level math: The role of high school courses. *Education Finance and Policy, 4*(1), 1-33. Retrieved from http://www.mitpressjournals.org/doi/pdf/10.1162/edfp.2009.4.1.1

Munter, C., Garrison Wilhelm, A., Cobb, P., & Cordray, D. S. (2014). Assessing fidelity of implementation of an unprescribed, diagnostic mathematics intervention. *Journal of Research on Educational Effectiveness*, *7*(1), 83-113.

Porter, A. C., & Polikoff, M. S. (2012). Measuring academic readiness for college. *Educational Policy, 26*(3), 394-417.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249-255.

Tipton, E. (2014). Stratified sampling using cluster analysis: A sample selection strategy for improved generalization from experiments. *Evaluation Review*, *37*(2), 109-139.