(2018). In (Eds.) A.Weinberg, C. Rasmussen, J. Rabin, M.Wawro, and S. Brown, Proceedings of the 21st Annual Conference on Research in Undergraduate Mathematics Education, San Diego, California.

Implementation and Impact of a Web-based Activity and Testing System in Community College Algebra

Shandy Hauk
WestEd

Bryan Matlen
WestEd

Most community college students in the U.S. must complete at least one developmental class, such as elementary algebra, before they can enroll in a college-level mathematics course. Increasingly common in such courses is the use of a web-based activity and testing system (WATS). This report presents initial results of a mixed-methods study of elementary algebra learning among 510 students in the classes of 29 instructors across 18 community colleges. Instructors were randomly assigned to use a particular WATS (treatment condition) or their usual approach (control condition). The focal WATS had adaptive problem sets, hints, and videos. Treatment group instructors had access to online support for implementation. For the study, students completed common pre- and post-tests and instructors regularly provided information about their teaching practices. The early results reported here indicate that greater instructor fidelity to developer intentions regarding frequency of assignments are positively associated with greater student learning.

Key words: College Algebra, Multi-site Cluster Randomized Controlled Trial, Fidelity of Implementation

More than 14 million students are enrolled in community college in the United States. Each is seeking an educational path to a better life. Community college students are more likely to be low-income, the first in their family to attend college, from a group under-served by status quo K-12 education (e.g., from an ethnic, racial, or linguistic minority group; Bailey, Jeong, & Cho, 2012). Most must take at least one developmental class, such as elementary algebra, before they can enroll in a college-level course (Porter & Polikoff, 2012). When it comes to technology and early algebra learning in college, what works? For whom? Under what conditions? When instructors implement technology tools, how are they used? In ways aligned with developer intentions? To what degree? Several web-based activity and testing system (WATS) have emerged for use in college developmental mathematics (e.g., ALEKS®, Khan Academy). Some WATS, like the one at the heart of this study, include adaptive problem sets, videos, and tools for instructors to monitor student learning. Though some research on the efficacy of WATS exists (e.g., Gardenhire, Diamond, Headlam, &Weiss, 2016 and references therein), the study reported here is the first large scale, multi-institution, mixed-methods experimental study of a WATS in community college developmental algebra of which we are aware.

Research Questions

Funded by the U.S. Department of Education, we are conducting a multi-year large-scale mixed methods study in over 30 community colleges in one U.S. state. We report here on the first year. The study is driven by two research questions:

Research Question 1: What is the impact of a particular WATS 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

Regardless of how they might be used, WATS environments vary along at least two dimensions: (1) the extent to which they adaptively respond to user behavior (e.g., static vs. dynamic) and (2) how they are informed by a model of cognition or learning. Static WATS are non-adaptive – 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, and/or a hypothesized common learning trajectory. An example of this type of environment might be online problem sets from a textbook that give immediate feedback on accuracy (e.g., “Correct” or “Incorrect”).

Dynamic WATS environments keep track of learner behavior (e.g., errors, error rates, time-on-problem) and use this information in a programmed decision tree that selects problem sets or feedback based on estimates of student learning. An example of a dynamic environment might be a system such as ALEKS® or the approach now used in Khan Academy Missions. 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 target learning trajectory is offered. Depending on the number and type of items the particular user answers correctly (e.g., on the path to six in a row correct), the analyzer program identifies and assembles the next “mastery challenge” set of items.

Above and beyond responsive assignment generation, programming in a dynamic environment that is also cognitively-based is informed by a theoretical model that asserts the cognitive processing necessary for acquiring skills (Anderson, Corbett, Koedinger, & Pelletier, 1995; Koedinger & Corbett, 2006). For example, instead of specifying only that graphing should be practiced, a cognitively-based environment also will specify skills needed to comprehend graphing (e.g., connecting spatial and verbal information), and provide feedback and scaffolds that support these (e.g., visuo-spatial feedback and graphics that are integrated with text). In cognitively-based environments, scaffolds themselves can be adaptive (e.g., more scaffolding through examples can be provided early in learning and scaffolding faded as a student acquires expertise; Ritter, Anderson, Koedinger, & Corbett, 2007). Like other dynamic WATS, such systems can provide summaries of student progress. No fully operational cognitively-based WATS currently exists for college students learning algebra. Several dynamic systems do exist (e.g., ALEKS®, Khan Academy Missions). The particular WATS investigated as the treatment condition in our study was designed primarily for use by learners as replacement or supplement to homework or in-class individual seatwork.

The theoretical basis for our approach to examining the instructional implementation of a WATS lies in program theory, “the construction of a plausible and sensible model of how a program is supposed to work” (Bickman, 1987, p. 5). As in many curricular projects, developers of the WATS in our study paid attention to learning theory inasmuch as it shaped the content in standard algebra texts upon which the WATS content was based. Developers articulated their assumptions about what students learned in completing WATS activities, but the roles of specific components, including the instructor role in the mediation of learning, were not clearly defined.

Munter and colleagues (2014) have pointed out that there is no agreement on how to assess fidelity of implementation (how close implementers come to realizing developer intentions; Dusenbury, Brannigan, Falco, & Hansen, 2003). However, there is a growing consensus on a component-based approach to measuring the structure and processes of implementation (Century & Cassata, 2014). Five core components are key in examining implementation: Diagnostic, Procedural, Educative, Pedagogical, and Engagement (see Table 1).

Table 1. Components and focus in examining implementation.

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, 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.

Method

The study we report here used a mixed methods approach combining a multi-site cluster randomized trial with an exploration of instructor and student experiences. Half of instructors at each community college site were assigned to use a particular WATS (treatment condition), the other half taught as they usually would, barring the use of the treatment group’s focal WATS tool, though other WATS might be used (control condition). Faculty participated for two semesters so treatment instructors could familiarize themselves with implementing the WATS with their local algebra curriculum. ***Note***: We report here on data collected from the first of two years. Hence, we purposefully under-report some details.

Sample for this Report

Initial enrollment in the study included 89 instructors across 38 college sites. Attrition of instructors by the end of the year was significant (68%). In the end, 29 instructors at 18 colleges finished the study (i.e., we had sufficient data from them to include them in analyses). This report is based on data from these instructors and their 510 students (see 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*

Measures

A great deal of textual, observational, and interview data were gathered. These data allow analysis of impact (Research Question 1) and an examination of implementation structures and processes (Research Question 2). Initial indices of implementation fidelity were based on instructor weekly self-reports of WATS use and, for the treatment group, on the WATS audit trail of student use. The primary outcome measure for students' performance was an assessment from the Mathematics Diagnostic Testing Program (MDTP), a valid and reliable test of students' algebraic knowledge (Gerachis & Manaster, 1995).

Student Mathematics Performance. One way to estimate student achievement on the MDTP tests is the *raw* score (i.e., proportion of correct answers as a percentage). However, such a calculation does not take into consideration other parameters of interest, such as item difficulty. To address this, in a second analysis we used a multilevel extension of two-parameter logistic item response theory to compute student pre- and post-test *scale* scores (Birnbaum, 1968). Specifically, we computed response-pattern *expected a posteriori* estimates (EAP scores; Thissen & Orlando, 2001) for each student. Also, we created EAP average scores for each classroom (a teacher-level score).

Instructor Implementation Processes. The components in Table 1 were operationalized through a rubric, a guide for collecting and reporting data on implementation. Each component has several factors. The research team developed a rubric for fidelity of implementation that identified measurable attributes for each component (Hauk, Salguero, & Kaser, 2016). For this report, we focus on the first element in the "procedural" component (see Table 3). The values and proportions (e.g., at least 2/3 of weeks) were specified by the developer. Data on the aspects in Table 3 was collected from weekly logs in which instructors indicated (a) WATS assignments made, (b) encouragement to students to complete assignments, (c) use of recommended mindset lessons, and (d) nature of attention in-class to student experiences with the WATS.

Table 3. Example of rubric descriptors for levels of fidelity, Structural-Procedural component.

Procedural: These components tell the user (instructor) what to do regarding instruction. Scaling within unit of instructor..			
	Low Level of Fidelity	**Moderate Fidelity**	**High Level of Fidelity**
Assigned WATS	Instructor rarely or never assigns WATS activities (2 or fewer times per semester).	Instructor sometimes assigns WATS activities (between 3 and 8 times per semester).	Instructor regularly assigns WATS activities (at least 8 times per semester).
Value of WATS	Instructor rarely or never encourages students to complete assignments (less than 1/3 of weeks/term).	Instructor sometimes encourages students to complete assignments (1/3 to 2/3 of weeks/term).	Instructor regularly encourages students to complete assignments (at least 2/3 of weeks/term).
Effort-based mindset	Instructor conducts at most 1 session of mindset training.	Instructor conducts 2 sessions of mindset training.	Instructor conducts recommended 3 sessions of mindset training.
Intensity of in-class supports for WATS use	Explicit mention or attention in class to content in WATS in fewer than 50% of weeks in term.	Explicit mention or attention in class to content in WATS from 50% to 80% of weeks in term.	Explicit mention or attention in class to content in/from WATS at least 80% of weeks in term.

Results

The study employed Hierarchical Linear Modeling (HLM), controlling for students' pre-test MDTP scores, to estimate the impact of WATS use on student achievement. The hierarchical modeling approach accounts for the nested structure of the sample (Raudenbush & Bryk, 2002), specifically the nesting of students within instructors. Preliminary analysis indicated that such a hierarchical model was justified: the intra-class correlation in the unconditional model was 0.36, suggesting that the observations were not independent (i.e., scores varied based on classroom – statistically, the teacher mattered – so single-level regression was not appropriate). The exact model and random and fixed effects for it are reported elsewhere (Hauk & Matlen, 2017).

Intervention Impact

Baseline equivalence. The What Works Clearinghouse (2014) considers baseline differences with a Hedges $g > .25$ not to be amenable to statistical correction. The raw pre-test scores were higher in the treatment group, with a marginal effect size ($g = 0.25$) for the difference between groups; however, the difference between treatment and control student pre-test EAP scores was substantive ($g = 0.30$). The EAP pre-test difference is large enough that the analytic sample might be considered non-equivalent at baseline on this variable (below, we discuss this fact).

Impact analysis. The aim of impact analysis was to address the question: What is the impact of the WATS intervention on students' elementary algebra knowledge, as measured by the MDTP? Controlling for students' pre-test scores, we found that using WATS corresponded to, on average, treatment student post-test scores 5 percentage points higher than the control group ($p < .05$). 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 (i.e., entirely above zero). Using EAP instead of raw scores, we obtained similar results. Since baseline differences between treatment and control group student raw scores were within the range of statistical correction, the similarity between the two *models* (raw score and EAP score models) is important, offering more confidence in the estimates of positive impact.

Implementation Fidelity and Student Learning

Of the 12 treatment instructors, 9 provided sufficient weekly information about the amount of instruction using WATS to determine a level of implementation. Three instructors were coded as high fidelity, 3 as moderate, and 3 as low. We explored whether the level of this category of procedural implementation fidelity in the treatment group correlated with student learning. To estimate *learning* we computed a normalized gain score, calculating *z*-scores for the pre- and post-test EAPs separately, and then subtracting the post-test *z*-scores from the pre-test *z*-scores for every student. These gains represent a difference between a student's relative position on the distribution of pre-test scores to their relative position on the distribution of post-test scores. Thus, a negative gain does not mean that a person (or in the case of Figure 2, a group of people) know less by the end of the course. Rather, it means that students are lower in the standardized distribution at post- than at pre-test. Figure 1 shows the average gain for treatment group students at each of the different fidelity of implementation levels for "Assigned WATS." We report here on this Procedural factor because it had the most notable differences across levels of fidelity.

The results in Figure 1 (next page) suggest that the more regularly instructors assigned WATS lessons, the larger the student gain. However, sample sizes at present are small, so results are not definitive. Nevertheless, these results are consistent with the developer's expectations (and an addition of a second cohort will allow us to examine whether these associations persist).

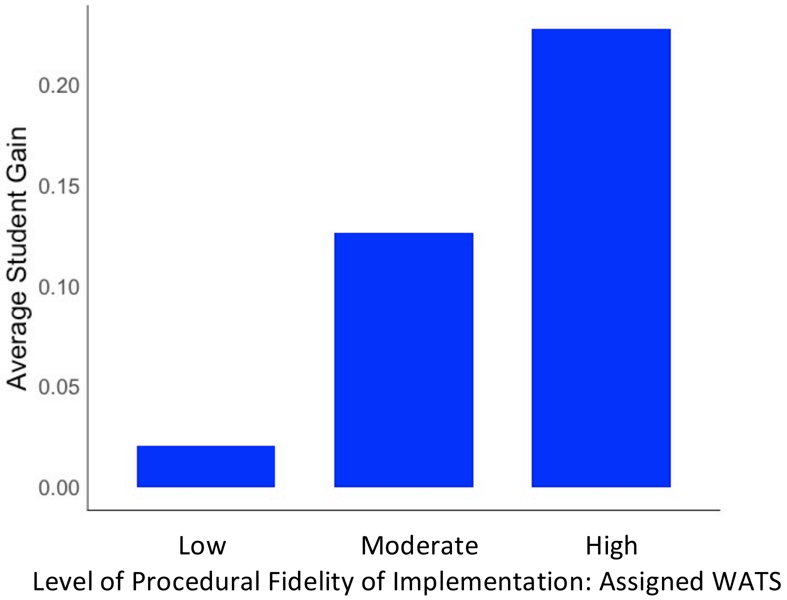

Figure 1. Average student gains in treatment classrooms according to instructor's level of procedural fidelity of implementation in the factor "Assigned WATS."

Control group instructors might use a WATS, but were restricted from using the focal WATS under study. Thus, one question was whether use of *any* WATS, regardless of whether it was the focal one, correlated with student learning. To explore this possibility, we examined average student gains for instructors using the treatment WATS (*n* = 9) compared to those in the control group who used a different WATS (*n* = 8). Figure 2 suggests that use of the study's focal WATS in particular, not just any WATS, had a positive relationship with student learning.

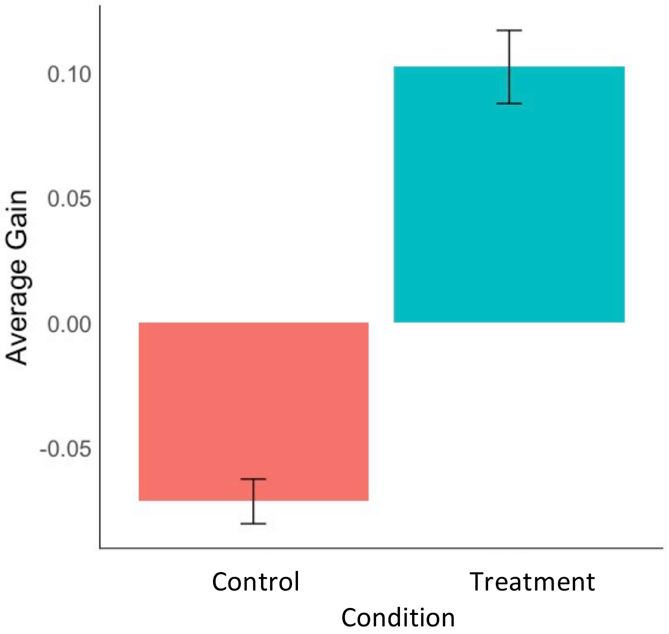

Figure 2. Average student gains for control group classes using a WATS vs. treatment WATS.

Discussion

We continue to explore relationships between fidelity of implementation and student success. While the results to date suggest that the focal WATS had a positive impact on students' elementary algebra achievement, recall there was high instructor attrition. This fact, coupled with moderate to large baseline differences at pre-test, warrant caution in interpreting the results. Still to do is a systematic consideration and testing of alternative explanations for Figures 1 and 2. For example, differential attrition may mean that treatment instructors who stayed in the study were better at incorporating the focal WATS into instruction. Mitigating against this explanation are the exit surveys completed by instructors who left the study in which course reassignment was the primary reason for treatment instructor attrition. Also, while we know the types of WATS used by control group instructors, we do not have a developer-validated fidelity of implementation rubric for each of those other WATS.

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. As we move forward with modeling, implementation indices (or vectors of values, one for each factor) will be used at the instructor level in statistical modeling of the impact of the intervention as part of a "specific fidelity index" (Hulleman & Cordray, 2009).

Conclusion

As indicated above, we have repeated the study with a second cohort of participants in the 2016-17 academic year. The new data, combined with the first study reported here, may provide additional results and insights by the time of the RUME conference.

Implications for practice. For the question: Should faculty use a WATS? The answer is a cautious: It depends. We know that treatment instructors had supports for WATS use in the form of video-based professional development and access to a project consultant who was experienced with the focal WATS in community college algebra. The implementation supports for control group teachers who used another WATS were varied. Taking into account the potentially biased statistical impact results and the exploration of variation in instructor implementation, there is still an open question about what might be the minimal supports needed for an instructor to have high fidelity on procedural components (e.g., Assigns WATS).

Implications for research. There were significant challenges in recruiting and retaining community college mathematics instructors for the study. To build community and assist in future research efforts in two-year colleges, we are sharing the processes and results of this work in materials read by community college faculty and administrators (e.g., *MathAMATYC Educator* – a journal of the American Mathematical Association of Two Year Colleges). It is important for potential faculty participants in research and their chairs/deans to be aware of the enormous contributions faculty can make to research. A second pragmatic implication for research is in how to manage the data generated by such projects (Hubbard, 2017)

Acknowledgement

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140340 to WestEd. The opinions expressed are those of the authors and do not represent views of the Institute or the Department of Education.

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.

Bailey, T., Jeong, D. W., & Cho, S.-W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review, 29*, 255–270.

Bickman, L. (1987). The functions of program theory. *New Directions for Evaluation, 33*, 5-18.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Century, J., & Cassata, A. (2014). Conceptual foundations for measuring the implementation of educational innovations. In L. M. H. Sanetti and T. R. Kratochwill (Eds.), *Treatment Integrity: A Foundation for Evidence-Based Practice in Applied Psychology* (pp. 81-108). Washington DC: American Psychological Association.

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.

Gardenhire, A., Diamond, J., Headlam, C., & Weiss, M. J. (2016). *At Their Own Pace: Interim Findings from an Evaluation of a Computer-Assisted, Modular Approach to Developmental Math*. New York: MDRC. Available in ERIC, http://files.eric.ed.gov/fulltext/ED567012.pdf

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

Hauk, S., & Matlen, B. (2017). Exploration of the factors that support learning: Web-based activity and testing systems in community college algebra, 2. In A. Weinberg, C. Rasmussen, J. Rabin, M. Wawro, and S. Brown (Eds.), *Proceedings of the 20th Conference on Research in Undergraduate Mathematics Education* (pp. 360-372). (ISSN2474-9346). Available online, November 27, 2017 at http://sigmaa.maa.org/rume/RUME20.pdf.

Hauk, S., Salguero, K., Kaser, J. (2016). How “good” is “good enough”? Exploring fidelity of implementation for a web-based activity and testing system in developmental algebra instruction. In T. Fukawa-Connelly, N. Infante, K. Keene, and M. Zandieh (Eds.), *Proceedings of the 19th Conference on Research in Undergraduate Mathematics Education*. ERIC Number: ED567765, http://files.eric.ed.gov/fulltext/ED567765.pdf

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.

Hubbard, A. (2017). Data cleaning in mathematics education research: The overlooked methodological step. In A. Weinberg, C. Rasmussen, J. Rabin, M. Wawro, and S. Brown (Eds.), *Proceedings of the 20th Conference on Research in Undergraduate Mathematics Education* (pp. 129-140). (ISSN2474-9346). Available online, November 27, 2017 at http://sigmaa.maa.org/rume/RUME20.pdf

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.), *The Cambridge Handbook of the Learning Sciences*. Cambridge, MA: Cambridge University Press.

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.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

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.

Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen and H. Wainer (Eds.), *Test Scoring* (pp. 73-140). Mahwah, NJ: Erlbaum.

What Works Clearinghouse (2014). *Procedures and standards handbook* (Version 3.0). Washington DC: Institute of Education Sciences.