(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. 360-372). San Diego, CA.

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

[Conference Long Paper]

Shandy Hauk & Bryan J. Matlen
WestEd

A variety of computerized learning platforms exist. In mathematics, most include sets of problems to complete. 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 such systems also varies – from static dictionaries of problems to 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 mixed-methods study that included a cluster randomized controlled trial design. The study was in community college algebra classes where the intervention was a particular type of web-based activity and testing system.

Key words: Computer-based Learning, 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 online materials appropriate to provide sufficient practice for each students' needs? Did instruction with the intervention equitably prepare students to pass the final exam?

This 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 used in community college elementary algebra classes. 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 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 WATS tool?

Background and Conceptual Framing

There are distinctions among dynamic and static learning environments (see Table 1). Though the focus of this report is a particular dynamic system, we offer information on both to situate what that means. 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 WATS environments based on adaptability and basis in a theory of learning.

		Type of Adaptivity in Design	
		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 (e.g., ALEKS, Khan Academy, ActiveMath)
	Yes	Textbook design and use driven by fidelity to an explicit theory of learning	"Intelligent" tutoring systems (e.g., Cognitive Tutor)

Static learning environments deliver content in a fixed order and contain scaffolds or feedback that are identical for all users. Although often informed by a learning theory, this type of system is distinguished from others in that it is not designed to immediately adapt to individual learning needs of users. An example of this type of environment might be online problem sets from a textbook that give immediate feedback to students such as "correct" or "incorrect." Studies of college algebra student achievement and attitudes when instruction uses these tools in conjunction with face-to-face instruction (e.g., computer-based homework rather than paper-and-pencil homework) is mixed, generally indicating that use will do no harm but is not particularly beneficial (e.g., Bishop, 2010; Buzzetto-More & Ukoha, 2009; Hauk, Powers, & Segalla, 2015).

Dynamic learning environments keep track of some user behaviors (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 estimated mastery of specific skills. An example of an "adaptive" dynamic environment might be a system such as ALEKS or the "mastery challenge" approach now used in the online Khan Academy Mission structure. For example, in working on a particular skill (e.g., the distributive property) in the Algebra Mission, 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. Some studies have found correlations between adaptive-dynamic systems and student learning (e.g., Murphy et al. 2014). However, other than our own, we are unaware of any large-scale experimental studies assessing the efficacy of adaptive-dynamic systems in college mathematics.

Above and beyond 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 instance, 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. For example, more scaffolding through examples can be provided early in learning and scaffolding can fade 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. The efficacy of early computer versions of such an approach has been documented in some large-scale studies in high school and college settings (Koedinger & Sueker, 1996; Koedinger et al., 1997). However, no fully tested cognitively-based web-based activity and testing system currently exists for college students learning algebra.

As mentioned, several adaptive 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 of study. The second year of the study – which repeats the design of the first – is currently underway. Hence, we purposefully under-report some details.

Methods

The study we report here uses a mixed methods approach that combines 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 in their instruction (treatment condition), the other half taught as they usually would, barring the use of the Treatment WATS tool though other WATS might be used (control condition). Faculty participated for two semesters in order to allow instructors to familiarize themselves with implementing the WATS with their local algebra curriculum. Specifically, the first term in Fall was a "practice" semester to field-test the intervention and the second semester of the same academic year was the "efficacy" study from which data were analyzed.

Sampling Strategy

Rather than recruit a sample by convenience, which is likely to result in poor generalizability, we utilized a stratified sampling approach developed by Tipton (2014). This method is a way of recruiting a sample that is compositionally similar to the target population for which the results of the study are meant to generalize. The target population for this study was defined as students at all community colleges in semester-long elementary algebra courses (also known as "developmental" or "beginning" algebra, the equivalent of a first year of algebra), in the U.S. state where the study took place. This population was selected in part because the state is large and diverse, and in part because we sought to decrease variability that may result from differing high school mathematics standards and graduation requirements across multiple states.

To recruit a sample that was compositionally similar to the target population, we first created a database that included information about all eligible community college sites (more than 100 across the state). We included information on college-level characteristics that existing research suggests will correlate with the study outcome (e.g., the average age of students at the college, the proportion of adjunct faculty, the proportion of students enrolled in remedial math courses). We conducted a cluster analysis on these potential covariates with all of the eligible colleges. The analysis resulted in a five-cluster solution that explained 29% of the variance between colleges. Examination of the characteristics that were unique to each cluster yielded the following descriptive observations:

Cluster 1. Represented 25% of colleges. These are colleges with a total student enrollment near the average (across all community colleges in the state) whose students tend to take more credits in the evening relative to colleges in other clusters. Cluster 1 colleges have more Temporary Faculty, and more Hispanic students, African American students, and students over 40 years old.

Cluster 2. Represented 15% of colleges. These colleges serve primarily students aged 25 and above who take fewer credits and more commonly are evening students.

Cluster 3. Represented 22% of colleges. These are colleges with a total student enrollment near the state average where students are more commonly Asian, younger, and enrolled full time during the day.

Cluster 4. Represented 23% of colleges. Cluster 4 represents smaller colleges that have a higher proportion of white students that tend to be younger, mostly full-time, and take fewer evening courses.

Cluster 5. Represented 15% of colleges. These are larger colleges that have more Hispanic and younger students. Students tend to take more daytime courses, with more fulltime loads and many remedial mathematics courses and high remedial math enrollment.

Our recruitment efforts aimed to include a proportionate number of colleges within each of the five clusters. Recruitment for the first cohort of participants yielded a study sample of colleges similar to the overall distribution across clusters that was the target for the sample. Due to attrition (instructors leaving the study), the representation shifted away from the target slightly for Clusters 1 and 4 by the end of the second term (see Figure 1).

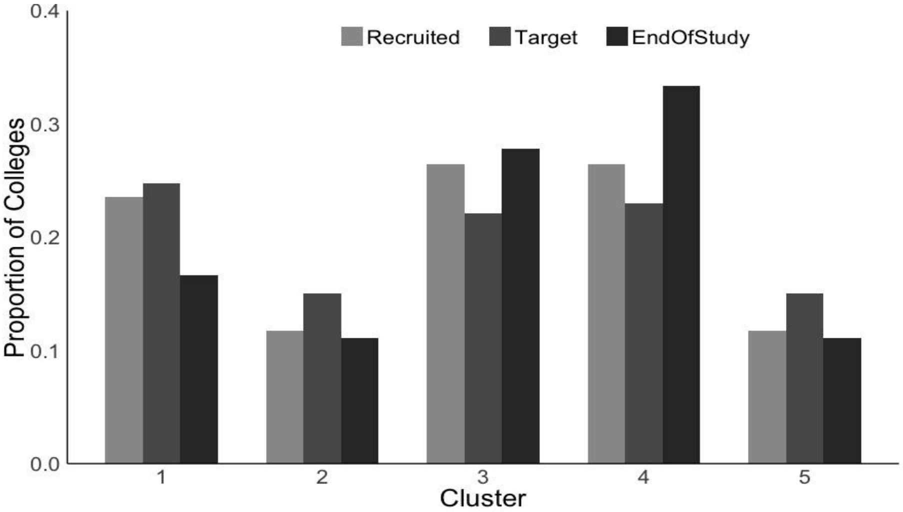

Figure 1. Recruited, target, and end of spring sample proportions across clusters.

Sample for this Report

Initial enrollment in the study included 89 instructors across 38 college sites. Attrition of instructors from initial enrollment to the end of the spring efficacy data semester was significant (68%). For this report, we analyzed the data from 510 students of 29 instructors across 18 colleges. Student and instructor numbers related to the data reported on here are shown in Table 2 and characteristics of the teachers and colleges are presented in Table 3.

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

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

* Note: there were multiple instructors at some colleges.

Table 3. Descriptive statistics for the student and instructor populations across the colleges in the study, by condition.

		Treatment		Control	
		M	*SD*	*M*	*SD*
		Student Characteristics			
	Enrollment	26,520	10,240	25,300	18,200
Proportions	U.S. Citizens	0.88	0.05	0.88	0.12
	Math Basic Retention*	0.80	0.03	0.82	0.06
	African American	0.04	0.03	0.06	0.03
	Asian	0.11	0.07	0.15	0.14
	Hispanic	0.49	0.21	0.41	0.19
	Native American	0.00	0.00	0.01	0.02
	White	0.27	0.17	0.30	0.16
	Below 25	0.61	0.05	0.58	0.07
	25 and Above	0.39	0.02	0.42	0.03
	Day Students	0.76	0.04	0.72	0.11
	Evening Students	0.18	0.04	0.16	0.04
		Instructor Characteristics			
	Part Time Faculty	0.45	NA	0.33	NA
	Years Experience Teaching Math	15.78	8.86	15.54	6.59
	Semesters of Algebra Teaching	18.60	11.99	15.36	13.82

* Proportion retention in remedial mathematics courses

Measures

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 analysis of impact (Research Question 1) and 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 (Research Question 2). Indices of specific and generic fidelity derived from this work also will play a role in HLM generation and interpretation in the coming year. The instruments are summarized below. With the exception of the observation and interview tools, all measures were administered online.

Instructor Instruments

Technology and Teaching Survey. This survey measures teachers' self-reported ability to use technology for teaching.

Perspectives Survey. This survey consists of questions related to teachers' background (e.g., years of experience teaching algebra, demographic information) as well as their attitudes and perspectives about teaching.

Measures of Effective Teaching – Algebra Test. This test was developed, piloted, and validated by the Educational Testing Service as part of the *Measures of Effective Teaching* (MET) project. It assesses instructors' pedagogical content knowledge in developmental algebra.

Weekly Instructor Logs. After extensive pilot testing, weekly logs were developed that ask about course format, topics, and resources used for that week's instruction.

Observation & Interview. The observation protocol captures a variety of information, including frequency of mention of WATS use, work completed in a WATS (treatment or other), teacher in-class use of WATS tools, as well as amount of time spent in whole class, group, and individual work. The interview focus is on the successes and challenges teachers face in using a WATS as part of instruction.

Student Instruments

Mathematics Diagnostic Testing Project (MDTP) Assessment. The MDTP serves as the study's primary student outcome measure. The *Algebra Readiness* form is the pre-test administered at the start of the semester and the *Elementary Algebra* form is the end-of-semester post-test. The MDTP tests have been shown to be valid and reliable measures of students' algebraic understanding (Gerachis & Manaster, 1995).

Student Background Questionnaire. This survey asks students about academic and demographic information such as academic history in mathematics, eligibility for financial aide

Motivated Strategies for Learning Questionnaire (MSLQ). This questionnaire measures students' motivation and attitudes towards mathematics.

Student Evaluation of Teaching Survey. The evaluation survey asks students to assess their experience in the course using Likert-scale questions.

The way performance is calculated is a non-trivial issue in educational measurement. One way to estimate student achievement on the MDTP tests is to calculate the raw percentage correct (i.e., summing the number of correct scores, and dividing by the total possible score). However, such a calculation does not take into consideration other parameters of interest, such as item difficulty, that provide added information that can be used to estimate student ability. To address this issue, we used a multilevel extension of the two-parameter logistic item response theory model 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. Similarly, we created EAP average scores for each classroom (a teacher-level score). We used individual and classroom aggregate student EAP scores in the analytic model described below.

Results

Quantitative Analysis

The study employed Hierarchical Linear Modeling (HLM), controlling for students' pretest MDTP EAP 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 revealed that the HLM choice was justified, as the intra-class correlation in the unconditional model was 0.36, suggesting that the observations were not independent (i.e., student scores varied based on their classroom – statistically, the teacher mattered – so other approaches, such as single-level regression, would be inappropriate). The specific HLM model we used:

$$Y_{ij} = \beta_{00} + \beta_{01}(WATS)_j + \beta_{10}(StuPre)_{ij} + \beta_{02}(InstructorPre)_j + \xi_{0J} + \epsilon_{I0} \quad \text{(Equation A)}$$

In the equation above,

Y_{ij} is the MDTP post-test EAP score for the i-th student of the j-th instructor;
β_{00} is the grand mean of EAP scores across all students;
$(WATS)_j$ is a dichotomous variable indicating instructor assignment to use the particular treatment WATS or not;
$\beta_{10}(StuPre)_{ij}$ is the student MDTP pre-test EAP score;
$\beta_{02}(InstructorPre)_j$ is the MDTP pre-test EAP estimate for all students in the class of instructor j;
ξ_{0J} and ε_{I0} represent a random effect term for instructors and a random error term, respectively.

All covariates were grand-mean centered to achieve the desired model interpretation (i.e., covariates were transformed to be centered on a mean of zero). Importantly, the impact of the treatment WATS use is captured by β_{01}.

Baseline equivalence. The What Works Clearinghouse (2014) considers baseline differences with a Hedges $g \leq .25$ to be within the range of statistical correction. However, differences of Hedges $g > .25$ are considered not amenable to statistical correction. As can be seen in Table 4, both situations occurred. The differences between Instructor mean EAP scores (i.e., classroom average) and student pre-test raw scores were moderate between the two conditions. However, the difference between student pre-test EAP scores was substantive across conditions (g = 0.30). The EAP pretest difference for students is large enough that the analytic sample might be considered non-equivalent at baseline on this variable (below, we discuss details that attempt to address this difference).

Table 4. Baseline equivalence analysis on the analytic sample.

	Effect Size Hedges g	*WATS*		*Control*	
		M	*SD*	*M*	*SD*
Student Pre (Raw Scores)	0.25	30.58	8.27	28.54	7.82
Student Pre (EAP Scores)	0.30	0.45	1.10	0.14	0.99
Instructor Pre (EAP Scores)	0.08	0.22	0.53	0.18	0.39

Intervention impact. The aim of the impact analysis was to address the question: After controlling for student and classroom-level average pre-test scores, what is the impact of the WATS intervention on students' elementary algebra knowledge, as measured by the MDTP? To address this question, we use Equation A to estimate the average impact of going from the control to the treatment condition. Ideally, what we are interested in is this: what would a control students' algebra achievement be if his/her instructor, in an alternative universe, were assigned to the treatment group? Because students cannot participate to both conditions simultaneously, our randomized trial is a proxy for this counterfactual scenario. The results of random and fixed effects in the model are presented in Tables 5 and 6, respectively.

The random effects (Table 5), tell us that the amount of variance that the instructor-level accounts for (i.e., the intraclass correlation) is about 28% (from Table 5 and a quick calculation, we see instructor variance divided by the total variance = **0.16**/(**0.40**+**0.16**)=**0.28**). This means that student level values are not independent. Put another way, students within classrooms were more similar to each other than students between classrooms. The intraclass correlation justifies our hierarchical analytic approach over single level regression. More generally (and in future work), we want to look at what instructors are doing to see how the instructor-level activity is shaping student achievement. The fixed effect model estimates are provided in Table 6. Controlling for students' pretest EAP scores, we found that using this particular WATS platform corresponded to a **0.35** increase in students' post-test EAP scores. This difference is considered a statistically significant positive effect ($p < .05$). The Hedges *g* value for this effect is 0.32, which is judged to be substantively important for educational research studies of this type (WWC, 2014). The 95% confidence interval around the effect estimate was 0.14 - 0.50, which is large, but spans an exclusively positive range.

Table 5. Random effects of the model.

	Variance	*Standard Deviation*
Instructor ξ_{0J}	**0.16**	0.40
Level-1 Error ε_{I0}	**0.40**	0.63

Table 6. Fixed effect results of the model.

	Estimate	*St. Error*	*p-value*
Intercept β_{00}	-0.10	0.10	0.34
WATS β_{01}	**0.35**	0.16	0.04
StudentPre β_{10}	0.73	0.03	< .001
InstructorPre β_{02}	0.30	0.19	0.13

Using raw MDTP scores (instead of EAP estimates) as outcomes and covariates in the model, we obtained similar results. In the raw score model, the impact of WATS was estimated to result, on average, in a 2.57 point increase in student raw score. This was a statistically significant positive effect ($p = 0.04$, *SE* = 1.18, *Hedges g* = 0.32). The control group mean was estimated at 22.04 (out of 50 points), thus, the 2.57 point difference corresponds to nearly 12 percentage points increase in post-test scores relative to the control group (2.57 / 22.04 * 100 = 11.66). 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, providing more confidence in the estimates of positive impact.

The effect size across both analyses was estimated at 0.32. This result can be interpreted as the WATS group of students would have scored an estimated 0.32 standard deviations higher, on average, than the control group of students on the MDTP, had the groups been fully equivalent prior to the intervention. However, to interpret the effect size of 0.32 in a more meaningful way, we converted the effect size using properties of the normal distribution. In a normed sample, a one standard deviation increase from the middle of the distribution corresponds to a 34 percentile point increase in scores. Thus, an effect size of .32 would correspond to an approximate 11 percentile point increase in scores (i.e., .32 * 34 = 10.88). Therefore, if students in the control condition perform at the 50th percentile in a normed sample, the students in the WATS condition would perform at the 61st percentile in the normed sample (50 + 10.88 = 60.88).

While these results suggest that WATS has a positive impact on students' elementary algebra achievement, it is important to note that this study suffered from high instructor attrition. This fact, coupled with moderate to large baseline differences at pretest, warrant caution in interpreting the results. In order to determine whether the results of the present study are robust, we are repeating the study with a second cohort of instructors and their students in the 2016-17 school year. Pooling the results of these two studies will help to determine the extent to which the findings replicate with different samples and will lend more confidence in the study conclusions (Cheung & Slavin, 2015).

Qualitative Analysis

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 (Table 7, next page).

The components in Table 7 are operationalized through a rubric, a 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 identifies measurable attributes for each component (for example, see Table 8 on the next page for some detail on the "educative" component). Data for assessing each row come from the survey, observation, and interview measures described earlier.

Table 7. 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.

Table 8. 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 content (CK), pedagogical (PK), and technological knowledge (TK)	Instructor regularly integrates content, pedagogical, and technological knowledge (TK) 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 philosophy behind the intervention	Instructor understands philosophy of WATS resources (practice items, "mastery mechanics," analytics, and coaching tools),	Instructor is aware of it, but understanding of the philosophy of WATS tool has some gaps.	Instructor is not aware of or does not understand philosophy of WATS resources.
*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 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.
* Note: Disagreeing is okay, this is about instructor knowledge of it.			

Defining and Refining Measures for the Fidelity of Implementation Rubric

The ultimate purpose of a fidelity of implementation rubric is to unpack and articulate the conditions of implementation and the relationship between those conditions and impact on student achievement. In addition to allowing identification of alignment between developer expectations and classroom enactment, an examination of implementation 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 for each teacher across the year of data collection. This can be as simple as the approach shown in Table 8, a 3 for a high level of fidelity, 2 for a moderate level of fidelity, or a 1 for a low level. The general score for a teacher-level index of implementation fidelity will be the total number of points assigned in completing the rubric as a ratio of the total possible. At a more detailed level, once we have completed rubric analysis to create the row by row scores for each instructor, these scores will be used as a vector of values in statistical modeling of the impact of the intervention as part of a "specific fidelity index" (Hulleman & Cordray, 2009).

We are at the beginning of addressing 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 WATS tool? To date, analysis of observations, interviews, and weekly logs has provided the opportunity to discover instructional orientations. Several orientations are emerging from analysis now and include a "denial" orientation in which instructors see the WATS as no different from themselves as a teacher, a "polarized" orientation where an instructor is either indifferent (no/low expectations for success) or enthusiastic (high/excessive expectations for success) about the power of student engagement with the WATS, a "cautious optimism" in which the instructor sees the WATS as one tool in a collection of resources to be used strategically in designing instruction, or an "adaptation" orientation in the sense that the instructor sees the WATS as a resource for which appropriate instructional use is negotiated with and through the students' goals for interaction with the software in the context of the algebra course. In addition to the fidelity scoring of alignment between developer expectations and classroom enactment, these orientations may serve to explain the relationship between implementation and impact, getting at how and for whom WATS are most effective.

Next Steps

As indicated above, we will continue this study with a second cohort of new participants in the 2016-2017 academic year. Our specific objectives in the coming six months are to complete the second cohort's efficacy semester, generate fidelity indices for each instructor in each cohort, and complete separate and collective statistical modeling explorations.

Implications for practice. Though the study is ongoing, the early results might be considered promising. If the question is: Should I use a WATS? The answer is: It depends. Taking into account the potentially biased statistical impact results to date and the exploration of variation in instructor implementation, it appears likely that an orientation of "cautious optimism" or "adaptation" may be required for a dynamic WATS tool like the one in the study to have significant impact on student learning.

Implications for research. A mixed-methods study like the one reported here is large and complex. We note here that there were significant challenges in recruiting and retaining community college mathematics faculty for the project. To build community and assist in future research efforts in two-year colleges (and as part of our dissemination about the work) we have targeted outlets read by community college faculty (e.g., MathAMATYC Educator – a journal of the American Mathematical Association of Two Year Colleges). It is important for practitioners and potential participants in studies on research in undergraduate mathematics education to be aware of research and the enormous contributions they can make to it. Secondly, a major implication for research (for us) was the work in managing all the data generated by the project. The reader is encouraged to review the piece by our colleague Aleata Hubbard that also was presented at the conference, *Data Cleaning in Mathematics Education Research: The Overlooked Methodological Step*.

Acknowledgements

This project is supported by a grant from the U.S. Department of Education, Institute of Education Sciences (IESR305A140340). 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. *Journal of the Learning Sciences, 4*(2), 167-207.

Bishop, A. R. (2010). *The effect of a math emporium course redesign in developmental and introductory mathematics courses on student achievement and students' attitudes toward mathematics at a two-year college* (Doctoral dissertation). Retrieved from http://aquila.usm.edu/dissertations/471

Buzzetto-More, N., & Ukoha, O. (2009). The efficacy of a web-based instruction and remediation program on student learning. *Issues in Informing Science and Information Technology*, 6, 285-298.

Cheung, A., & Slavin, R.E. (2015). *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

Hauk, S., Powers, R. A., & Segalla, A. (2015). A comparison of web-based and paper-and-pencil homework on student performance in college algebra. *PRIMUS*, *25*(1), 61-79.

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 Educational Effectiveness*, 2(1), 88-110.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8,* 30–43.

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.

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. *American Education Finance Association*.

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.

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

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

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

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2014, March). *What Works Clearinghouse: Procedures and Standards Handbook* (Version 3.0). Retrieved from http://whatworks.ed.gov