A Multifaceted Consideration of Motivation and Learning within ASSISTments

by

Korinn S. Ostrow
(ksostrow@wpi.edu)

A Thesis

Submitted to the Faculty

of

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Learning Sciences & Technologies

April 2015

APPROVED:

_____

Neil T. Heffernan, Ph.D., Co-Director, Advisor

_____

Joseph Beck, Ph.D.

_____

Ivon Arroyo, Ed.D.

Abstract

An approach to education gaining popularity in the modern classroom, adaptive tutoring systems offer interactive learning environments in which students can access immediate feedback and rich tutoring while teachers can achieve organized assessment for targeted interventions. Yet despite the benefits that these systems provide, a number of questions remain regarding the optimal inner workings of adaptive platforms. What is the recipe for optimal student performance within these platforms? What elements should be taken into consideration when designing these learning environments? Can facets of these platforms be harnessed to increase students' motivation to learn and to improve both immediate and robust learning gains?

This thesis combines work conducted over the past two years through versatile approaches toward the goal of enhancing student motivation and learning within the ASSISTments platform. Approaches considered include a) enhancing motivation and performance through altered feedback using hypermedia elements, b) instilling motivational messages alongside media enhanced content and feedback, c) allowing students to choose their feedback medium, thereby exerting control over their assignment, d) altering content delivery by interleaving skills to enhance solution strategy development, and e) establishing partial credit assessments to drive motivation and proper system usage while enhancing student modeling.

After a brief introduction regarding the main tenants of this research, each chapter highlights a randomized controlled trial focused around one of these approaches. All studies presented have been conducted or are still running within ASSISTments. Much of this work has already been published at peer reviewed conference venues, some with stringent acceptance rates as low as 25% for full papers. Two of the studies presented here are second iterations of previously published work that are still in progress, and only preliminary analyses are available. A chapter on conclusions and future work is included to discuss the contributions that have been made to the Learning Sciences community thus far, and to briefly discuss potential directions for my continued research.

Acknowledgements

Heartfelt appreciation goes first and foremost to Sam for supporting me through the journey that has fallen into place since first finding the Learning Sciences & Technologies program at WPI and blindly submitting an application two years ago. To the many nights you have allowed me to disturb your sleep with the glow of my laptop. To your endless patience with the turmoil that arises from my stubborn determination. Your patience with my breakdowns, my tears, and the many moments that have challenged my self-worth. And to the pride and strength you continue to provide through both the celebrations and the hardships.

Endless love to a pup named Louie, the best and most patient boy anyone could have. To a sweet love that curls up next to me for hours past bed time, patiently waiting for me to be done. For all the precious moments we have lost to a computer screen, a sacrifice that is sometimes too difficult to face. And for all that you are, maybe without even knowing, that keeps me going.

I would also like to extend gratitude to my family, blood and otherwise. Thank you for walking me through the failures and holding me up until the next success. For accepting sacrifices of our time together all too often, whether I have been a ghost or I have half existed while glued to a screen. Mom and Dad: for pushing me without ever really pushing, but rather trusting that I would succeed. For phone calls that come too sparingly now, and visits that are even more scarce. For constantly letting me know that you believe in what I have already accomplished and what I continue to strive for. For reading my work and for celebrating my successes even when it/they may not make much sense or seem that important. For always reminding me that these are just stepping stones, and that I am still climbing, even when I am too afraid.

To those that have worked alongside me, who have helped me meet my goals and who have allowed me to help them meet theirs. To co-authors and to those who have helped without earning the proper acknowledgments, thank you.

Finally, to one of the craziest and most hardworking people I have ever met, who seeps passion for his work, and somehow stares down each day with a conviction to rule the world. Neil, you gave me the opportunity to run from a place in my life marked by much unhappiness, and to blossom in a world I still sometimes wonder if I have dreamt up. You continue to provide an abundance of support that has allowed me to succeed in ways I never knew I could. And of course, thanks to that fellow's other half, Cristina, a woman who never ceases to support my work and keeps me grounded, generally when it counts the most. To the entire faculty for your patience, your trust, and your tendency to treat me like a colleague rather than the inane student I feel like sometimes. You are helping me build my future, one brick at a time, and for that I am infinitely gracious.

Table of Contents

List of Tables

List of Figures

Chapter 1 – Introduction

Despite the crucial role of education in America, more often than not, curriculum and classroom practices are not held to the scientific rigor of other fields. Attempts to regulate, standardize, and unify best practices throughout the nation have frequently been met by pushback rather than support (i.e., the Common Core State Standards). Coupled with this issue, the role of technology is continuously expanding in the classroom and interactive systems now allow students to learn in new and unique ways. Computer-aided testing platforms and adaptive tutoring systems have led to exponential growth in the availability of educational data and many goals for education have drifted toward individualization rather than standardization. The U.S. Department of Education's National Educational Technology Plan supported the idea that technology will play a key role in delivering personalized educational interventions (U.S. Dept. of Ed., 2010a). Yet there remains a severe lack of research regarding the effectiveness of online learning systems for K-12 education (U.S. Dept. of Ed., 2010b). Adaptive tutoring systems offer interactive learning environments in which students can excel while teachers can maintain organized, data-driven classrooms. Before the development of these adaptive platforms, research within classrooms was costly, difficult, and generally required a longitudinal approach. Studies that could take place within real world classrooms were typically invasive and disruptive to learning, with teachers sacrificing lesson time to pretests, posttests, and formal debriefing practices. As such, much of the evidence that supports educational practice and learning theory is drawn from studies conducted by psychologists in laboratory settings with samples of college undergraduates. Our collective understanding of the intricacies of K-12 learning, including elements of motivation and self-regulation, is built largely on generalizations. However, the technology required to easily conduct non-invasive educational research in real world classrooms now lies at our fingertips in the form of adaptive tutoring systems.

The research studies described in this thesis were conducted, or are still underway, as randomized controlled trials within ASSISTments, an adaptive online tutoring system with a focus on K-12 mathematics content that provides assistance and assessment to over 50,000 students around the world as a free service of Worcester Polytechnic Institute (Heffernan & Heffernan, 2014). ASSISTments provides teachers with the unique ability to build their own content or to access prebuilt certified content and textbook material linking to more than twenty of the top mathematics textbooks in the United States without infringing copyright. The system, commonly used for both classwork and homework, presents students with immediate feedback and a variety of rich tutorial strategies, while offering teachers a toolbox of preferences and settings to craft the optimal the learning experience. ASSISTments is a powerful assessment tool, providing teachers a variety of student and class reports that allow them to pinpoint where students are struggling and enhance classroom techniques through a data driven model. External to the classroom effects, one of the most unique aspects of ASSISTments is that the platform's structure allows educational researchers to design and implement content-based experiments without extensive computer programming knowledge. A recent journal article highlighting the decade-long growth of ASSISTments cited that its capacity for research has led to the publication of over 18 peer-reviewed articles detailing the results of randomized controlled trials (Heffernan & Heffernan, 2014).

The tutorial strategies used within ASSISTments can take a variety of forms. The simplest type of feedback, correctness feedback, merely lets the student know if she answered correctly (signified by a green checkmark) or incorrectly (signified by a red X). ASSISTments can also provide Hints, or small pieces of information pertaining to the problem that are presented upon the student's request for assistance. Problems typically contain at least one hint, known as the Bottom Out Hint, which provides the student with the answer to keep her from getting stuck within the assignment. However, problems often contain multiple hints with compounding specificity and helpfulness (i.e., on a problem with 4 hints, the final being the Bottom Out, the first hint in the series provides little insight, while the third provides much insight). Scaffolding Problems are another form of feedback within ASSISTments. Scaffolding can be used to break a problem down into multiple sub-steps, or to provide the student with a worked example of an isomorphic problem. For questions that contain scaffolding feedback, tutoring is provided automatically if the student answers the question incorrectly or if they request that the problem is broken

down into steps. ASSISTments also features Buggy Messages, or feedback that is carefully crafted based on common wrong answers to provide students a push in the right direction when presented alongside correctness feedback.

Despite a variety of feedback methods, ASSISTments lacks breadth in feedback medium. Until recently, virtually all feedback within the platform was provided using text, typically with font color or typeset as pointers to significant variables within the content. In rare cases, feedback has also taken the form of images or animated gifs, and videos were recently used to promote student engagement and motivation (Kelly, et al., 2013). However, adaptive tutoring systems offer the opportunity to utilize a variety of hypermedia elements, as outlined by Mayer's multimedia principles for the optimal design of e-Learning environments (Clark & Mayer, 2003). These twelve principles, driven by cognitive theory, promote active learning while reducing cognitive load and accounting for the average user's working memory (Mayer, 2005). It is possible that video can be used to provide feedback with more emphasis than text, maintaining identical content but increasing the connected feeling provided by a human tutor. Two of Mayer's principles in particular are significant to the introduction of matched content video feedback to the ASSISTments platform through brief 15-30 second YouTube recordings: the *redundancy principle* and the *modality effect*. The former posits that although hypermedia can be effective, material should only be offered through a single information-processing channel (i.e., a video hint should not be accompanied by text that explains the same content). By reducing the overlap of processing channels, the learner's attention is better appropriated and intake of information is optimized (Clark & Mayer, 2003). The latter principle states that a multimedia approach using audible narration is more effective than content presented using on-screen text (Mayer, 2005). Despite the predominant influence of Mayer's work in the field, his research is largely carried out in laboratory settings, and as such, there is little evidence supporting his impressive findings in actual K-12 classrooms. Yet in the age of educational television programming and iPad applications, students are already familiar with hypermedia learning; as video gains popularity in the classroom, especially in the context of the flipped classroom (Goodwin & Miller, 2013), it is crucial to harness this increasingly accessible medium within adaptive tutoring systems like ASSISTments. Work investigating the characteristics of student users within adaptive tutoring systems has previously suggested that tailoring feedback to individual users can be quite effective (van Seters, Ossevoort, Tramper & Goedhart, 2012), providing the groundwork for offering students additional feedback mediums and control over their experience with feedback.

Further, although ASSISTments is in the process of adopting new feedback mediums, it lacks true student level personalization and typically fails to promote the motivation and affect of users. Considering the current infrastructure, students are not able to make choices or exert control within their assignments despite the fact that choice is an intrinsically motivating force (Patall, Cooper & Robinson, 2008). It is highly possible that the controlling nature of the tutor leaves students feeling powerless, thereby having negative effects on learning and performance. According to the work of Pekrun, choice has the potential to boost subjective control, or a student's perception of their causal influence over their learning outcomes (2006). Feelings of control are balanced by appraisals of subjective value, or a student's perceived importance of her learning outcome. By providing the student with choices at the start of her assignment, it may be possible to enhance expectancies regarding her performance and thereby enhance achievement emotions such as motivation (Pekrun, 2006). It is also possible that supplementing assignments with choice will have a differential effect across genders or other moderating variables; research has shown that similar achievement outcomes can lead to markedly different emotional responses across genders (Frenzel, Pekrun & Goetz, 2007; Patall, Cooper & Robinson, 2008). Considering the control-value theory within the realm of an adaptive tutoring system for mathematics content may help to explain and ameliorate female dropout in STEM fields (Frenzel, Pekrun, & Goetz, 2007). An expansive longitudinal study regarding motivation within mathematics achievement recently revealed that perceptions of control are correlated with active learning, higher rates of persistence, and improved intrinsic engagement (Murayama, et al., 2013). These findings are not shocking, as the modern student is an expert when it comes to control amidst technology. Students with smartphones, iPads, or popular gaming systems make choices on a daily basis that personalize their experience of these interactive systems. Despite the fact that

users can endlessly customize their experiences with commercial products, student preference is not a key element in realm of education. If offering students somewhat trivial choices regarding their learning does not result in negative consequences (i.e., choosing a feedback medium differs drastically from providing students the option of completing an assignment), boosting the perception of control within adaptive tutoring systems seems like an obvious next step.

It is also possible to enhance student motivation and learning outcomes through approaches that optimize content delivery. Interleaving, a simple intervention in which skills have a mixed presentation rather than a blocked presentation, has consistently proven effective in the realm of mathematics education in recent years (Mayfield & Chase, 2002; Rohrer & Taylor, 2007; LeBlanc & Simon, 2008; Taylor & Rohrer, 2010; Li, Cohen, & Koedinger, 2012). These benefits are often credited to the discriminative-contrast hypothesis (Birnbaum, et al., 2013), which purports that the effect is rooted in a student's enhanced ability to pinpoint differences in problem content, allowing student to better practice problem type identification and solution strategy choice (Rohrer, 2012). Although a large collection of work details the benefits of this approach to content delivery, the process is rarely used in practice. Policymakers and educational designers fail to interleave mass-produced skill content in mathematics, claiming that it is detrimental to the student's learning *experience* (Rohrer & Pashler, 2010; Taylor & Rohrer, 2010; Kornell & Bjork, 2008). Essentially, the learning experience itself seems more complex and frustrating, what Bjork terms 'desirable difficulty' (Bjork, 1994). Additionally, many questions still exist when considering the effect of interleaving. Is there an optimal quantity of skills to consider when interleaving content? Perhaps the duration of the intervention is significant; is it possible to observe the effects of interleaving within brief sessions? Do the benefits of interleaving differ when considering student characteristics (i.e., their skill level)? Researching the interleaving of skill content within adaptive tutoring systems can offer a variety of avenues for enhancing student performance.

A final consideration is the use of partial credit assessment to improve student interaction with adaptive tutoring platforms. Much of the focus within educational data mining entails using a binary notion of student's accuracy on each question to make predictions regarding knowledge state, learning, or next problem correctness. State of the art approaches including Knowledge Tracing (Corbett & Anderson, 1995) and Performance Factors Analysis (Pavlik, Cen, & Koedinger, 2009) allow researchers to gain insight into how students learn by considering (typically) binary performance on a sequence of skill opportunities. However, despite expansion in the field that has lead to a number of alternative or supplementary modeling techniques, few have considered more continuous measures of student performance. It can also be difficult to decipher what percentage of variation in predictions hails from the student, versus that which hails from the skill content (Pardos & Heffernan, 2010; Pardos & Heffernan, 2011). It is possible that these models rely on binary correctness due to the complexity of accurately and universally defining an algorithm that validates partial credit scores within adaptive tutors. Still, the primary goal of these platforms is not solely to assess student knowledge, but to simultaneously promote student learning through adaptive feedback, making binary correctness a stale concept. Students often require multiple attempts to solve a problem or request system feedback for guidance, thus assigning value to the concept of partial credit assessment within these platforms. Partial credit could serve to enhance student motivation, prompt appropriate usage of system feedback, and would serve to enhance modeling predictions when attempting to predict next problem correctness.

The following chapters highlight a set of randomized controlled trials conducted within the ASSISTments platform that focus on student motivation through feedback mediums, the provision of student choice, interleaving skill content, and data driven partial credit assessment. Chapters 2-4 describe studies that have considered the learning gains observed through altering feedback mediums to include video and animated characters. Chapters 5-6 detail studies which again altered feedback mediums, but sought to observe how student motivation and learning differed when students were offered choice regarding feedback. Chapter 7 describes a content-based experiment that revealed the potential benefits of interleaving cognitive skills within an adaptive tutor. Finally, Chapters 8-9 describe data mining exercises examining partial credit scoring and the potential benefits that optimizing such assessment could have for students, teachers, and for student modeling and prediction. The whole of these works are considered in

Chapter 10, with conclusions regarding the contributions to the field thus far and directions for future work examined.

Chapter 2 – Testing the Multimedia Principle in the Real World: A Comparison of Video vs. Text Feedback in Authentic Middle School Math Assignments

As technology evolves and design options for web-based homework support systems expand, researchers are left with questions regarding best practices. These platforms often provide students correctness feedback meant to guide learning and offer dynamic tutoring to help students solve difficult problems. Feedback typically consists of bland text and worked examples, but as hypermedia gains prevalence, researchers are turning their focus to the appropriate use of such elements in e-learning environments. The following study assesses the effects of feedback medium within a randomized controlled trial conducted using ASSISTments, an adaptive math tutor. Results suggest that video feedback enhances learning outcomes and is well perceived by student users. These findings are of particular interest to the Learning Sciences, with intent to optimize e-Learning design.

INTRODUCTION
A leader in the field of e-Learning, Richard Mayer has defined various multimedia principles for the optimal design of technology supported learning environments such as web-based homework support systems (Clark & Mayer, 2003). Rooted in cognitive theory, these principles call for the design of learning environments that are driven by an active learning process and that take the restraints of cognitive load and working memory into consideration (Clark & Mayer, 2003; Mayer, 2005). Still, researchers seeking to enhance student engagement, motivation, and persistence, they are left questioning how to optimize the learning environment without overloading learners.

Mayer also posits that learners utilize separate information processing channels to internalize information; under the *redundancy principle*, material offered through one channel (i.e., a narrated passage) should not be simultaneously presented through another (i.e., text accompanying the narration) (Clark & Mayer, 2003). When such circumstances occur, the learner's attention is split across redundant content, depressing intake from both channels and hampering learning. Further, the *modality effect* suggests that learning gains are greater for narrated content than for content presented as text (Mayer, 2005). Based on these principles, the use of video, when presented without redundant textual explanation should appeal to both auditory and visual processing channels without risking overload.

Video is not novel to education, and it is growing increasingly popular due to the concept of the "flipped classroom," which often parallels the use of web-based homework support systems. While the quality of evidence for the flipped classroom has not yet proven impressive (Goodwin & Miller, 2013), the trend speaks to the growing accessibility of technological resources in education. Self-recorded video

lectures and feedback offer teachers the opportunity to be deeply involved in student learning while simultaneously enhancing ownership of the technology (Kelly, et al., 2013).

Contrary research has suggested that video is not universally successful in promoting learning gains. In his early work on the effect of educational movies, Pane (1994) noted mixed results as a function of content, offering evidence that the use of video may improve the speed of immediate recall, yet potentially harm long-term learning. Negative effects of video may include prolonged time-on-task that potentially leads to boredom or frustration, the inability to appropriately convey abstract content material, and the likelihood of technological difficulties that prevent students from adequately accessing materials.

In the present study, the ASSISTments platform is used to compare the delivery methods of feedback messages within a mathematics e-Learning environment. Prior research has found that dynamic graphics are more effective than static graphics in mathematics realms (Mayer, 2005), and thus, we hypothesize that video will have a positive effect on learning gains within this system. Since its inception, ASSISTments has delivered significant results surrounding the use of textual feedback in the form of scaffolding and hints (Razzaq & Heffernan, 2006; Singh, et al., 2011; Wang & Heffernan, 2011); the present study serves as a preliminary exploration into replacing textual feedback with video. Thus, we pose the following research questions:

1. Are learning outcomes enhanced when scaffold feedback is delivered using video rather than text?
2. Can we determine if students disproportionately internalize feedback based on the medium, given next question performance and response time?
3. Based on self-report measures, do students respond positively to the addition of video to their assignment?

METHODS

Participants

A set of six questions requiring students to use the Pythagorean theorem was assigned to 139 8th grade students using ASSISTments. This student population was comprised of four classes of differing skill levels that spanned four suburban middle schools in Massachusetts and Ohio. All students were familiar with ASSISTments and used the system on a regular basis as part of classwork and homework assignments.

Design

The Pythagorean theorem problem set was derived from pre-existing ASSISTments certified material, based on Common Core State Standards and chosen in an attempt to match 8th grade fall curriculum. The structure of the problem set relied on three questions with text feedback (A, B, C) and three isomorphic questions with video feedback (A*, B*, C*). Each question and its morph were of similar difficulty and were therefore considered interchangeable (i.e., A and A*). The questions are available at (Ostrow, 2013b) for further comparison.

The fixed question patterns depicted in Table 2-1 were rooted in the intention to allow all students an equivalent opportunity to experience both feedback styles. Thus, the four groups were designed to house fixed question patterns from which we could assess the impact of video versus text at various points throughout the problem set. Random assignment was attained by allowing ASSISTments to allocate students into one of the four groups at the start of the assignment. As depicted in Table 2-1, students assigned to Group 1 received video feedback if they answered question 1 incorrectly, text feedback if they answered question 2 incorrectly, and so on.

Video content was designed to mirror textual feedback in an attempt to provide identical assistance through both mediums. Each video simply featured the lead researcher reading a feedback message while referring to the question content on a whiteboard. Figure 2-1 depicts question C* with video feedback, while Figure 2-2 depicts the question morph (C) with text feedback. All video material can be accessed at (Ostrow, 2013b).

Table 2-1. Group Design

Linear Order	1	2	3	4	5	6
Group 1	A*	B	C*	A	B*	C
Group 2	A	B*	C	A*	B	C*
Group 3	A*	B*	C	A	B	C*
Group 4	A	B	C*	A*	B*	C

*Depicts question morph with video feedback

Both types of feedback were set to load incrementally with incorrect responses or if the student requested to break the question down into steps. Videos were set to play automatically, allowing students to gain information with equal efficiency regardless of feedback medium, and perhaps making it harder or more inconvenient to "game the system," or click through the scaffold steps in rapid succession.

For each group, four post-test survey questions asked students to judge, using a simple three-measure Likert scale (i.e., not at all, somewhat, a lot), if they felt video feedback was helpful, if it was enjoyable, if they would prefer similar videos in future assignments, and what effect video feedback had on their focus. For the entire student experience, see (Ostrow, 2013b).

Figure 2-1. Video Feedback for Question C*
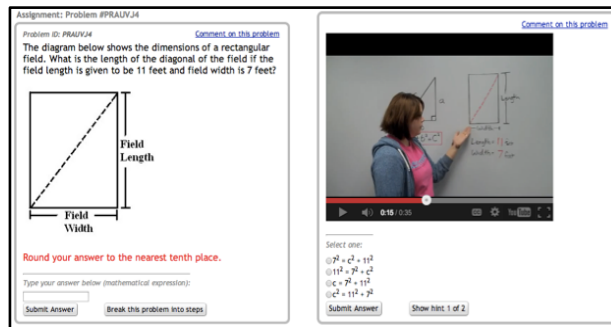
Figure 2-2. Text Feedback for Question C
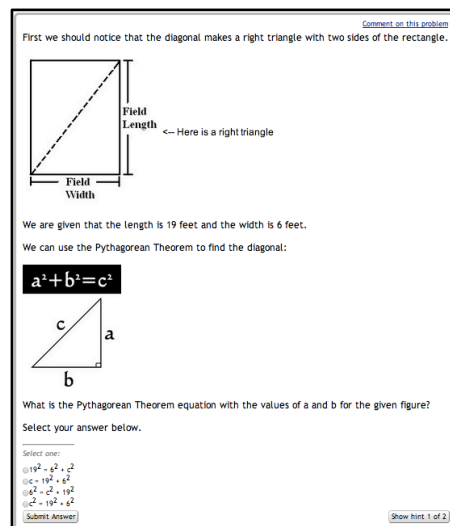
Procedure

The problem set was assigned to students in the manner consistent with their teacher's usual use of ASSISTments (i.e., as either classwork or homework). Students were free to work at their own pace and were not required to complete the assignment in one sitting. Log data was compiled for each student's

performance, including elements such as first action, correctness, response time, attempts, and hints requested. Delegating random assignment to the tutor produced results that were less than optimal, as significantly fewer students were assigned to Group 2 and Group 4. However, assessment of the code controlling ASSISTments' random assignment function concluded that this anomaly was not influenced by any student attribute or system characteristic.

Table 2-2 explains initial group assignment and the process for excluding students from analysis. A total of 139 students were originally assigned (OA) the problem set. Six students failed to log enough progress to initiate a group assignment and were therefore excluded. Of the remaining 133 students, 13 students did not complete the problem set (I), and 31 students tested out (TO) (these students answered each question correctly and failed to receive feedback of either style). A disproportionate number of students tested out of Group 3, likely as a function of random assignment and small sample size.

Table 2-2. Students Excluded From Analysis

	OA	I	TO	G	Remaining
Group 1	35	4	7	0	24
Group 2	29	3	6	4	16
Group 3	43	4	11	2	26
Group 4	26	2	7	4	13
Total	133*	13	31	10	79

*Six students failed to initiate a condition.

In prior research, "gaming the system" within ASSISTments has been defined as consistent answer seeking behavior displayed in rapid succession (i.e., clicking through all hints or scaffolds for completion) (Baker, et al., 2008). As such, "gamers" were operationalized as any student who clicked through question A (or A*) and its four scaffolds, regardless of feedback medium, at a rate faster than five seconds per response. By this loose definition, a total of 10 students qualified as "gamers" (G) and were removed prior to analysis as shown in Table 2-2.

Our primary analysis assessed student performance on the second question as a function of the feedback medium they experienced after incorrectly answering the first question. For question 1, Groups 1 and 3 were presented video feedback (A*), while Groups 2 and 4 were presented text feedback (A). We were therefore able to collapse these groups when analyzing second question performance. Based on Table 2-2, the removal of gamers significantly differs when the Groups are collapsed: for Groups 1 and 3, only 7.1% of students are removed from the remaining sample, while Groups 2 and 4 lose 43.5% of the remaining students. Considering our operational definition of gamers, and noting that Groups 2 and 4 received text feedback upon incorrectly answering question 1, the discrepancy found here suggests that video feedback may deter gaming. To better understand this bias, the proceeding analysis is carried out both with and without gamers for comparison.

RESULTS

Second Question Analysis

After considering the aforementioned exclusion methods, 79 students were remaining for analysis (89 when gamers were included). To address our initial research question, we assessed second question performance in students who had received feedback on question 1, as summarized in Table 2-3. Learning outcomes were enhanced for students who received video feedback (M = 0.77, SD = 0.43) rather than text feedback (M = 0.63, SD = 0.50), approaching significance at p = 0.143, with an effect size (CEM, 2013) of 0.32, 95% CI [-0.28, 0.91]. When gamers were included to analyze the effect of the selection bias, the improvement for students who had received video (M = 0.76, SD = 0.44) versus text (M = 0.52, SD = 0.51) became statistically significant, p < .05, with an effect size of 0.50, 95% CI [-0.03, 1.03].

Table 2-3. Summary of Second Question Analysis

	n	*Video*	*n*	*Text*	*t*	*p**	*Hedge's g*	*95% CI of g*
Performance								
w/o Gamers	35	0.77 (0.43)	16	0.63 (0.50)	1.08	0.143	0.32	[-0.28, 0.91]
w/ Gamers	37	0.76 (0.44)	23	0.52 (0.51)	1.90	0.031	0.50	[-0.03, 1.03]
Response Time								
w/o Gamers	35	134.86 (118.76)	16	421.77 (1122.27)	-1.51	0.068	-0.45	[-1.05, 0.15]
w/ Gamers	37	129.72 (117.46)	23	307.33 (943.50)	-1.14	0.130	-0.30	[-0.82, 0.23]

Note. Time is depicted in seconds. Mean (SD). *Significance presented as a one-tailed t-test.

Further analysis of second question performance suggested that response time was faster for students who had received video (M = 134.86, SD = 118.76) rather than text (M = 421.77, SD = 1122.27), approaching significance at p = 0.068, with an effect size of -0.45, 95% CI [-1.05, 0.15]. When gamers were included for comparison, students who had incorrectly answered the first question and received video feedback performed faster (M = 129.72, SD = 117.46) than those receiving a text scaffold (M = 307.33, SD = 943.50), but results were not significant and the effect size dropped to -0.30, 95% CI [-0.82, 0.23]. As gaming was defined as rapidly clicking through questions and feedback, it is not surprising that time measures would drop in this manner. While these results portray consistent trends approaching significance, they should be taken with caution, as the number of students who received text feedback was disproportionately smaller than the number of students who received video feedback.

Response Time Within Feedback

To address our second research question, we examined students' overall experience within each type of feedback. Students saw a total of 186 scaffold levels of video feedback, and 171 scaffold levels of text feedback while completing their assignment. On average, response time during video feedback (M = 202.51, SD = 337.99) was longer than response time during text feedback (M = 35.18, SD = 28.74) approaching significance at p = 0.085, with an effect size of 0.68, 95% CI [0.47, 0.90]. When gamers were included for comparison, students saw a total of 241 levels of video feedback, and 231 levels of text feedback. Average response times dropped within both feedback styles, yet response time during video feedback remained longer (M = 169.28, SD = 268.44) than response time during text feedback (M = 28.38, SD = 21.67), approaching significance at p = 0.076, with an effect size of 0.73, 95% CI [0.54, 0.92].

These results suggest that there was no significant difference in the overall number of feedback levels experienced by students as a function of feedback medium. On average, students spent 3 minutes and 23 seconds within video feedback and only 35 seconds within text feedback. When gamers were considered, less time on average was spent within each feedback style, with students spending 2 minutes 49 seconds within video feedback and only 28 seconds within text feedback. This difference is presented in Table 2-4. Students consistently spent more time within video feedback, suggesting that they actually took the time to watch the videos and internalize the content whereas they seemed to gloss over text feedback.

Table 2-4. Summary of Response Time Within Feedback

	n	*Video*	*n*	*Text*	*t*	*p**	*Hedge's g*	*95% CI of g*
w/o Gamers	186	202.51 (337.99)	171	35.18 (28.74)	1.48	0.079	0.68	[0.47, 0.90]
w/ Gamers	241	169.28 (268.44)	231	28.38 (21.67)	1.57	0.068	0.73	[0.54, 0.92]

Note. N is equivalent to levels of feedback seen by students over the course of the entire assignment. Each student may have seen multiple scaffolds. Mean (SD) in seconds. *Significance denoted by a one-tailed t-test.

Survey Response Analysis

Of all students available for analysis, 53 answered the four post-test survey questions. Student responses are proportioned in Table 2-5. Taken together, we consider the survey results to suggest that video feedback is well perceived by students. Essentially, 83% of students reported that they would at least somewhat prefer ASSISTments to use video more often. Coupled with the student performance findings discussed above, we feel that video may be a beneficial tool for ASSISTments and that further exploration regarding the long-term effect on learning is required.

Table 2-5. Student Responses to Posttest Survey Questions

	Not at all	Somewhat	A lot
How helpful did you find the videos as you completed your assignment?	14%	43%	43%
How much did you enjoy the videos?	17%	57%	26%
Would you like it if more of your ASSISTments assignments used videos?	17%	43%	40%
Did you feel more focused on your assignment when the question had videos?	30%	38%	32%

DISCUSSION

Although Mayer's work has been a predominant influence on the field of multimedia infused learning, much of his research has assessed college undergraduates in psychology labs. Thus, his results suggest a massive and seemingly unrealistic effect when compared to most real-world educational interventions. According to recent research detailing average effect sizes in educational settings, Lipsey, et al. (2012) note that at the middle school level, researcher developed studies with specialized topics tend to show strength with effect sizes of approximately 0.43. The present study is on par with this trend, with effect sizes for second question analysis ranging from 0.32 to 0.51. We argue that these results provide a contribution to the Learning Sciences and help establish a basis for future research.

Based on our findings, we feel that video feedback may be a significantly beneficial tool for e-Learning. Immediate learning gains, represented by second question performance after receiving feedback on question 1, were significantly greater in students who experienced video feedback. Our results suggest that the use of video forces the learner to slow down and internalize the concept that is being taught, as depicted by consistent trends for response times within the feedback experience. Although text feedback consistently provides a faster alternative for skilled readers, perhaps adaptively slowing the pace more closely mimics the actions of a human tutor.

It should be noted that video feedback appears to have deterred gaming behavior. This may have been due in part to novelty, but was likely a function of the automatic nature of video playback. When a student tried to game through a question, each scaffold level would present another video until they were all playing simultaneously. A slightly more qualitative inspection of gaming behavior within this problem set suggested that at least three of the students labeled as gamers corrected their behavior after being exposed to video feedback. Future research is required to determine if video feedback provides a beneficial intervention for this population in general.

Regardless of the cause, let us assume for a moment that these effects are valid and reliable, and that video feedback significantly enhances student performance. With the growing popularity of web-based homework support systems and the ubiquitous nature of video servers such as YouTube and SchoolTube, teachers and instructional designers may be overlooking a valuable tool. The videos used in this study were of low production quality, shot in a single take, and featured a non-professional actress reading from a script. Teachers with years of expertise in providing feedback could arguably record a short video on their smartphone or tablet that would outperform the content used in this study. The use of video within e-Learning environments has the potential to streamline the process of repetitive one-to-one tutoring and boost the teacher's efficiency in the classroom. While pedagogical agents have become a popular tool for

feedback delivery within e-Learning environments, the same messages may carry significantly more power when delivered by the student's teacher. A multitude of brief interactions offering personalized and appropriately timed feedback, guidance, and motivation, could become an important step toward truly adaptive tutoring.

Future implementations of this study should utilize a more powerful pre/post-test design with additional far transfer items and the use of open-ended survey response options to gauge student feedback. We also suggest that future endeavors compare a purely video condition to a control containing only textual feedback, perhaps using an AB design with multiple content topics to maintain fair treatment. Future work should also attempt to pinpoint critical elements driving the effects of video, such as motivation, novelty, personalization, and engagement.

Although further exploration regarding the effect of video feedback on robust learning is required, the findings of this work provided a basis for a multitude of refinements within the ASSISTments platform. Multiple grant proposals have been written to reflect the inclusion of short video feedback strategies recorded by teachers and students, and an initiative was established with partnering teachers in the state of Maine to develop brief video tutoring strategies for prebuilt certified content and textbook material. This influx of video feedback will allow for future studies analyzing the subtleties of content delivery. A multitude of brief recordings offering personalized and appropriately timed feedback, guidance, and motivation, could become an important step toward truly adaptive tutoring.

Chapter 3 – Scaling-Up the Multimedia Principle within ASSISTments: Further Examination of Video vs. Text Feedback

A second version of the experiment presented in Chapter 2 was established as a Skill Builder problem set to modify and scale-up the original research design within ASSISTments. Skill Builder problem sets are readily available to all users when placed in the ASSISTments Certified folder. Thus, rather than conducting an orchestrated study within just a few classrooms, the research sample can be expanded to include a vast number of student users. While this likely adds noise to the analysis, it allows for an assessment of how the effects of video feedback generalize to new users. The work highlighted here is still in progress and thus has not been submitted for publication; preliminary analyses are presented.

The ASSISTments platform was used to compare matched content video and text feedback within the context of the Pythagorean theorem. Given the results of the study in Chapter 2, it was hypothesized that video snippets would continue to outperform customary text feedback. When scaled up, the present study sought to reinforce the first two research questions examined in Chapter 2; this design did not offer an opportunity for self-report and thus could not reinforce results observed for the third research question.

METHODS
This study utilizes a more powerful design than that discussed in Chapter 2 to test immediate learning gains and to control the experience of feedback mediums. The content originally created for Experiment 1 was reused for this study in a modified delivery design, as shown in Table 3-1. The six Pythagorean theorem questions, three with text feedback (A, B, C) and three with video feedback (A*, B*, C*) served as the first question for one of six possible experimental conditions. The second question, "D," was identical for all groups. This question is depicted in Figure 3-1. It featured a right triangle drawn on a unit grid, and asked students to determine the length of the hypotenuse. This question was considered a far transfer item, as it was dissimilar to questions A, B, and C in that it required students to derive side lengths of the right triangle from the unit grid rather than simply presenting lengths in the question text. As shown in Table 3-1, each group also received an introductory 'problem' explaining that students would be presented with video and should turn on their speakers, as well as a 'video check,' or a problem used to test whether students were able to access YouTube videos without technical difficulty.

Table 3-1. Group Design

Linear Order	Introduction	Q1	Q2	Q3+
Group A		A	D	
Group A*	All Groups:	A*	D	All Groups:
Group B	Introduction and	B	D	Random Questions for
Group B*	Video Check	B*	D	Remainder of Assignment
Group C		C	D	
Group C*		C*	D	

*Depicts question morph with video feedback.
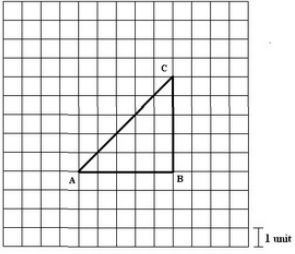
Figure 3-1. Second Question ("D") for All Students

Students were randomly assigned to one of the six groups, and after completing the brief experimental portion, the remainder of the assignment functioned as a normal Skill Builder. This design did not allow for students to receive both text and video feedback, but more appropriately controlled for the immediate effects of feedback on the first question. The content and settings of feedback remained identical to that originally explained in Chapter 2. That is, feedback was still presented using scaffolds and set to load incrementally with incorrect responses or upon the student's request for assistance. This method was maintained rather than using the hint strategy within ASSISTments, as it was observed that the majority of students fail to request hints when working through problems, even when struggling, and therefore the sample of treated students would be small. Hint usage may be low due to the nature of the hinting feature: when students request a hint the system considers their overall answer for that problem as incorrect.

PROCEDURE

Although this study is still in progress, preliminary analysis has been conducted for presentation here. Log files were collected and analyzed for the sample of students who have thus far completed the problem set. Teachers from 27 schools have assigned the problem set to a total of 1021 students. Through a post-hoc estimate of gender, based on the U.S. Social Security's list of the top 1000 names for both genders for children born in 2000 (roughly middle school aged children), approximately 70% of this sample was assigned gender as a potential covariate. It was estimated that this sample included at least 346 girls and 366 boys, with 309 students unable to be classified by name alone. The majority of students in the sample were estimated to be in 8th grade.

A large proportion of students had to be removed prior to analysis. A total of 61 students did not get past the introduction and video check, and therefore were never randomly assigned into one of the six

conditions. Of the remaining 960 students, 267 reported have technical difficulties in response to the video check, thus leaving only 693 students who were assigned to a condition. Additionally, 13 students dropped out before answering the first question, and therefore could not be included in analysis. These students were distributed across five of the six conditions. Of the remaining sample, the distribution is as shown in Table 3-2, along with means and standard deviations for question 1 and question 'D' for the full sample.

Table 3-2. Distribution of Full Sample, with Performance Analysis

Problem ID	Condition	N (full sample)	Q1 M (SD)	'D' M (SD)
PRAUVJ2	P1video	119	0.59 (0.49)	0.58 (0.50)
PRAUVJ3	P2video	120	0.64 (0.48)	0.46 (0.50)
PRAUVJ4	P3video	111	0.65 (0.48)	0.51 (0.50)
PRAUVJX	P1text	116	0.65 (0.48)	0.57 (0.50)
PRAUVJY	P2text	107	0.56 (0.50)	0.54 (0.50)
PRAUVJZ	P3text	106	0.63 (0.48)	0.57 (0.50)

However, many of these students effectively 'tested out' or got question 1 correct and therefore did not receive feedback of either type prior to question 'D.' These students offer no information regarding the differential effectiveness of feedback medium, and therefore were excluded from analysis. Of these students, 202 had been assigned into text feedback conditions and 219 had been assigned into video feedback conditions. The treated sample consisted of 258 students, with 127 receiving text feedback and 131 receiving video feedback, distributed as shown in Table 3-3.

Table 3-3. Distribution of Treated Sample, with Performance Analysis

Condition	N (treated)	Q1 M (SD)	'D' M (SD)
P1video	49	0.59 (0.49)	0.58 (0.50)
P2video	43	0.64 (0.48)	0.46 (0.50)
P3video	39	0.65 (0.48)	0.51 (0.50)
P1text	41	0.65 (0.48)	0.57 (0.50)
P2text	47	0.56 (0.50)	0.54 (0.50)
P3text	39	0.63 (0.48)	0.57 (0.50)

For analysis of response times, a trimming method was used to amend outliers. Response times were bound to a ten-minute maximum, thereby reducing noise that occurs naturally when students use the ASSISTments tutor for homework (i.e., students may walk away from their assignment for a prolonged period of time). The trimming method used in Chapter 2 was not as strict, as the majority of students had completed the assignment as classwork and only multi-day assignment logs required correction.

RESULTS
Within treated students, the distribution of the sample was far sounder than in Chapter 2. Findings are summarized in Table 3-4 for performance, response time, and attempts exhibited on question 'D' after receiving feedback on question 1. The sample of students who actually answered question 'D' only included 231 students, suggesting that 27 students dropped out after answering question 1 but prior to answering question 'D.' Distribution of this dropout was not significantly skewed, as a total of 15 students dropped out of text conditions, and a total of 12 students dropped out of video conditions.

Students performed mildly better after receiving video feedback, but the difference was not significantly reliable. Trimmed response times revealed that students took about 24 seconds longer on average to respond to the question 'D' after receiving text, although this difference was not reliable.

Treated students experienced 225 levels of text feedback and 233 levels of video feedback, suggesting that students assigned to video feedback conditions requested additional scaffolds or made additional incorrect attempts, thereby prompting further videos. This discrepancy in the amount of feedback experienced was not significantly different across groups, with students averaging 1.77 levels of text feedback (SD = 1.46) and 1.78 levels of video feedback (SD = 1.47), p > .05. Students also spent approximately 38 seconds longer on average within video feedback, a finding in line with that observed in Chapter 2.

Table 3-4. Means, Standard Deviations, and T-test Results for Three Measures of Performance

	Video (n = 119)	Text (n = 112)	t	p*	Hedge's g	95% CI of g
Performance 'D'	0.33 (0.47)	0.31 (0.47)	-0.247	0.805	0.04	[-0.22, 0.30]
Trimmed Response Time 'D'	121.67 (169.71)	145.17 (188.56)	0.997	0.320	-0.13	[-0.39, 0.13]
Trimmed Time in Feedback Q1	201.83 (196.79)	163.69 (184.39)	-1.605	0.110	0.20	[-0.06, 0.46]

Note. Time is depicted in seconds. *Significance presented as a 2-tailed t-test.

Secondary analyses were conducted to take into account the type of problem presented on question 1 (A, B, or C). An ANCOVA was conducted examining correctness on problem 'D' as a function of feedback medium after controlling for problem type on question 1. Results still suggested that feedback medium did not link to significant differences in correctness, $F(1, 228) = 0.074$, p > .05. Further, problem type on question 1 was not a significant covariate, $F(1, 228) = 0.211$, p > .05. A similar analysis conducted on the number of scaffolds used following question 1 revealed that the main effect of feedback medium trended toward significance, $F(1, 255) = 3.038$, $p = 0.083$, $\eta^2 = 0.012$. An investigation into each problem type revealed that for problems A and A*, students used significantly more scaffolds when receiving text feedback (M = 3.44, SD = 0.90) than when receiving video feedback (M = 2.73, SD = 1.38), t = 2.807, p < .01. However, this trend did not hold for either of the other two problem types (B, B* and C, C*).

DISCUSSION
The findings of this preliminary analysis are somewhat in line with the findings observed in Chapter 2, and suggest that while subtle differences exist between feedback mediums, more work will be required to detect reliable differences. Students appear to perform slightly better on an identical problem after receiving video feedback rather than text feedback, answering the next problem more efficiently after spending longer absorbing the feedback they were provided. The time difference here effectively cancels out for most students, with overall assignment time working out to be similar. Limitations of this design include the continued emphasis on short term learning effects, and the fact that the experimental treatment is limited to just the first question. Some students may answer this first question correctly and thus effectively test out, but then require feedback following later questions. Thus, this design is not receptive to the student's needs later in the problem set and merely offers a snapshot of initial knowledge and immediate response to feedback styles. Still, this experimental design provides a contribution to the understanding of how feedback mediums affect immediate learning gains. The prudent next step will be to measure more robust facets of learning after considering prolonged provision of a feedback medium, possibly throughout multiple assignments. While immediate effects of video are apparent but lack reliability, it is crucial to determine how video effects learning through more robust measures. Ultimately, this study will continue to run until and analyses will be recomputed to account for larger samples.

Chapter 4 – Promoting Growth Mindset Within Intelligent Tutoring Systems

When designing adaptive tutoring systems, a myriad of psychological theories must be taken into account. Popular notion follows cognitive theory in supporting multi-channel processing, while working under assumptions that pedagogical agents and affect detection are of the utmost significance. However, motivation and affect are complex human characteristics that can muddle human-computer interactions. The following study considers the promotion of the growth mindset, as defined by Carol Dweck, within middle school students using an intelligent tutoring system. A randomized controlled trial comprised of six conditions is used to assess various delivery mediums of growth mindset oriented motivational messages. Student persistence and mastery speed are examined across multiple math domains, and self-response items are used to gauge student mindset, enjoyment, and perception of system helpfulness upon completion of the assignment. Findings, design limitations, and suggestions for future analysis are discussed. In a sense, this work extends beyond that presented in Chapters 2 & 3 to examine alternative feedback mediums when presenting motivational materials.

INTRODUCTION
The optimal design of adaptive tutoring systems is a continuous debate for researchers in the Learning Sciences. Decisions when authoring content can be immense, including not only the user interface and tutor material, but also the presence of adaptive feedback strategies such as hints or scaffolding, the use of affect detectors, and in growing popularity, the use of pedagogical agents. While many adaptive tutors share designs rooted in cognitive theory, creators should also incorporate elements that improve student motivation, engagement, persistence, metacognition, and self-regulation skills. These elements aid in the promotion of active learning, an experience that has been shown to heighten the creation of mental connections (Mayer, 2014). However, successful adaptive tutoring systems are not just a random conglomeration of these learning goals. All too often, adaptive tutors are designed under the assumption that students are ideal learners, driven and motivated, ready to employ a full range of self-regulation skills coupled with technological prowess (Arroyo, et al., 2013). Thus, researchers have recently undertaken a more thorough examination of how to universally encourage and motivate students while still promoting self-regulated learning skills and optimizing system design (Bernacki, Nokes-Malach, & Aleven, 2013; Graesser, et al., 2007).

Human motivation has historically been explained and argued by an array of theories, as intrinsic or as extrinsic, as static or as the constant flow of needs, emotions, and cognitions (Reeve, 2009). In a somewhat similar sense, recent research promoting affect detection within educational technology suggests that affect plays a primary role in learning success (Baker, et al., 2008). How can researchers incorporate deeply rooted human characteristics like motivation and affect into the design of an adaptive tutoring system? A renowned leader in the field of psychology, Carol Dweck has helped to establish theories of intelligence that marry these complex constructs within the confines of learning studies (Dweck, 2002). Her research has shown that students approach learning tasks largely with one of two 'mindsets.' The *fixed mindset* is characterized by the notion that intelligence is somehow innate or immutable. Students who live within this fixed realm generally emit lower learning and performance outcomes as well as higher attrition rates based in the notion that effort will not lead to intellectual advancement (Dweck, 2006). Much of American society is rooted in this view; strong emphasis is placed on standardized testing and zero sum competition, with the goal of comparing student intelligence rather than promoting learning. Alternatively, students with a *growth mindset* believe that intelligence is

malleable and that effort and persistence can lead to success. While Dweck (2013) argues that neither mindset is necessarily 'correct,' she promotes the notion that mindset can be altered, and explains the growth mindset as offering a healthier mental lifestyle. Altering mindset is best achieved by varying the type of praise students receive and by realigning their definition of successful learning. By highlighting the learning process rather than the student's intelligence or performance, 'process praise' and the promotion of malleable intelligence has led to positive, long-term learning gains (Dweck, 2002). Students trained in the growth mindset show increased enjoyment in difficult learning tasks as well as higher overall achievement and performance (Dweck, 2006).

An expert in his own right, Richard Mayer has devoted much of his career to promoting a series of multi-media learning principles that enhance e-learning design. These principles call for learning environments to be driven by active learning processes while considering the cognitive load and working memory of users (Clark & Mayer, 2003). As such, those authoring adaptive tutors should utilize audio, animation, graphics, video, and other hypermedia elements to appease multiple sensory channels and thereby reduce the user's overall cognitive load. It is important to note that powerful design requires a fine balance of these resources, as exorbitance may serve to distract or disrupt learners. The evolution of pedagogical agents and learning companions within adaptive tutoring systems has served as a primary way to incorporate both multi-media elements and non-cognitive support. As guidelines for the design of human-computer interaction have followed those set forth by human-human interaction, the art of appropriating the cognitive and affective responses of pedagogical agents has been of major concern (Kapoor, Burleson & Picard, 2007). Agents are typically designed with the premise that they should respond happily to student successes and with a shared disappointment upon failures (Kapoor, Burleson & Picard, 2007).

Considering the optimal design of adaptive tutoring systems and the incorporation of hypermedia and pedagogical agents to engage students in active learning, the current study seeks to analyze the promotion of Dweck's growth mindset theory within ASSISTments, an adaptive mathematics tutor. The following research questions were derived from themes relevant to Dweck's (2006) work, in combination with adaptive tutoring structures unique to ASSISTments:

1. Does the addition of motivational messaging within the tutoring system affect the likelihood of student persistence or attrition?
2. Does the presence of motivational messaging within the tutoring system affect mastery speed as defined by how many items, on average, it takes for students to complete the problem set?
3. Can specific elements within message delivery be pinpointed as significantly powerful? That is, can researchers isolate an element (e.g., the presence of a pedagogical agent, the audio component, static images, or a combination of these elements) that is responsible for the majority of variance in persistence and learning efficiency?

It is hypothesized that students randomly assigned to a messaging condition will be more likely to show continued, persistent effort than those in the control condition. Similarly, regardless of the delivery medium, researchers expect students who receive mindset messages to show improved mastery speed, with fewer items, on average, required to complete a problem set. In the assessment of message delivery, it is hypothesized that motivational messages delivered using an animated version of Jane, a learning companion that originates from partnering tutor Wayang Outpost, will have a stronger effect on student persistence and learning efficiency than alternative message mediums.

METHODS

To determine appropriate math content for this study, the tutor's database was queried to compile a historical record of usage data for a variety of problem sets that fit within Common Core State Standards across various grade levels. All observed problem sets were of a style unique to the ASSISTments tutor, requiring students to answer three consecutive questions correctly in the same day in order to complete the assignment. If the student were to reach a preset 'daily limit' (i.e., ten problems) while attempting to solve three consecutive questions, they are prompted to consult with their teacher and try again tomorrow.

Five problem sets were chosen based on high usage, with math content spanning grades four through seven. The skill topics assessed by these problem sets included finding missing values using percent on a circle graph, equivalent fractions, multiplying decimals, rounding, and order of operations. The goal in designing multiple problem sets was three-fold: to increase data collection, to determine any significant effect for student skill level, and to determine if content was linked to student motivation, perhaps due to difficulty level. Six conditions were then established for each problem set, as defined in Table 4-1. These conditions were designed following the principles set forth by Clark & Mayer (2003), to test matched content messages across a variety of processing channels. The student experience for each problem set was formatted in the same manner. An introductory 'question' explained the format of the problem set and alerted the student to turn on their computer volume and to use headphones if necessary. The second 'question' tested whether or not the student was able to see and hear the pedagogical agent Jane as she introduced herself as a problem-solving partner. This question was included to test the compatibility of the HTML files that supported the pedagogical agent's animation and sound conditions, thus serving as confirmation of fair random assignment. Researchers then relied on a randomization feature unique to ASSISTments that randomly assigned students to one of the six conditions depicted in Table 4-1. Math content was isomorphic across conditions, and was thus considered comparable in difficulty. A test drive of the student experience for each problem set can be found at (Ostrow, 2013a).

Table 4-1. Motivational Messaging Conditions

Control	ASSISTments as usual; no messages added
Animation	Jane, a female pedagogical agent, delivers messages with motion and sound
Static Image with Text	The agent is presented as a static image, with a speech bubble to deliver motivational text messages
Static Image with Audio	The agent is presented as a static image, supplemented by audio files to deliver motivational messages
Word Art	A speech cloud shows motivational text messages, with no agent involvement
Audio	The agent's voice delivers motivational messages with no graphical changes to tutor content

Motivational message content, as depicted in Table 4-2, was matched across conditions to reduce confounding. These messages were validated in and derived from (Arroyo, et al., 2013). Each problem set was designed to randomly select questions from a pool of approximately 100 problems, containing two types of motivational message delivery: *general attributions*, in which the motivational message was presented with the primary question, and *incorrect attributions*, in which the motivational message was presented alongside content feedback if the student responded incorrectly or employed a tutoring strategy. Following this design structure, students saw general attributions on approximately half of the questions, with the remaining half displaying incorrect attributions only to students who answered a problem correctly. Therefore, each student's experience of motivational messaging may have differed slightly, even within each condition. This design was established to reduce persistent message delivery and to avoid inundating students with messages on each question, with the goal of optimizing the effects of motivational messages while retaining a primary focus on math content. All visual motivational messages appeared within the tutor and remained until the student completed the problem; audio messages were played once upon loading the problem or tutoring strategy.

At the end of each problem set, students were asked to partake in a series of four survey questions developed based on previously validated content from (Mueller & Dweck, 1998), to assess student mindset, goal orientation, and perceptions of enjoyment and system helpfulness. All students received these questions regardless of condition. All survey content can be accessed at (Ostrow, 2013a).

Table 4-2. Motivational Message Item Content

General Attributions
• Did you know that when we learn something new our brain actually changes? It forms new connections inside that help us solve problems in the future. Pretty amazing, huh?
• Did you know that when we practice to learn new math skills our brain grows and gets stronger? That is so cool!
• Hey, I found out that people have myths about math… like that only some people are "good" at math. The truth is we can all be successful in math if we give it a try.
• I think the most important thing is to have an open mind and believe that one can actually do math!
• I think that more important than getting the problem right is putting in the effort and keeping in mind the fact that we can all be good at math if we try.

Incorrect Attributions
• Making a mistake is not a bad thing. It's what learning is all about!
• When we realize we don't know why that was not the right answer, it helps us understand better what we need to practice.
• We may need to practice a lot, but our brains will develop with what we learn.

PROCEDURE

Teachers in the state of Massachusetts who frequently use ASSISTments with their students were approached with a brief presentation explaining the study and providing examples of the conditions, motivational messages, and math content. Teachers assigned one or more of the problem sets to their students in accordance with the teachers' usual use of the tutoring system (i.e., as either classwork or homework). Material was assigned as current content and/or review, for a total of 765 student assignments. Log data was compiled for each student's performance. Prior to analysis of persistence and mastery speed, students were removed if they had noted experiencing technical difficulties or if they failed to log enough progress to enter one of the six conditions. Additional students were removed prior to survey analysis due to incompletion. Students remaining after each step are examined across problem sets in Table 4-3.

An ex post facto judgment of student gender was determined for 570 students within the sample remaining for math content analysis. Due to incompletion rates within this subset of students, gender was determined for 554 students within the sample remaining for survey content analysis.

Table 4-3. Explanation of Students Remaining After Removals

Problem Set	A[1]	MA*	SA**
Percent on a Circle Graph	87	69	62
Equivalent Fractions	255	208	205
Multiplying Decimals	62	48	47
Rounding	253	208	205
Order of Operations	108	88	86
REMAINING	765	621	605

A[1] = Assigned. MA = Math Analysis. SA = Survey Analysis.
*Students removed prior to math analysis due to technical difficulties or failure to initiate a condition.
**Additional students were removed prior to survey analysis due to incompletion.

RESULTS

Analyses of student persistence and mastery speed were performed at the condition level for each problem set, as well as for an aggregate of the five sets to serve as a composite analysis of the conditions across math content. To determine if an effect existed within a particular processing channel, similar conditions were compiled based on delivery elements. For example, all conditions utilizing audio were compiled to assess the effect of audio (i.e., audio, animation, static image with audio). Similar analyses

were performed to determine the effect of textual messages and the effect of the pedagogical agent's presence. Researchers also compared a compilation of all conditions containing motivational messages to the control condition in order to determine the effectiveness of motivational messages in general. Initial findings suggested that in general, the sample was too advanced for the math content as students were found to be at ceiling across many of the problem sets. Thus, secondary analyses examined gender differences and assessed the aforementioned variables for a subset of students operationally defined as "strugglers," or those requiring more than three questions to complete their assignment.

When considering student persistence, as defined by continuing until reaching completion, ANOVA results suggested null results ($p > .05$) across all problem sets except for multiplying decimals $F(5, 42) = 2.57$, $p < .05$, $\eta^2 = 0.23$. No significant results were observed when the problem sets were compiled or when specific delivery elements were isolated, and there was no significant difference between messaging conditions and the control. For the full sample, gender was found to differ significantly on persistence, $F(1, 568) = 3.84$, $p = 0.051$, $\eta^2 = 0.01$, with girls showing significantly more persistence ($M = 0.99$, $SD = 0.12$) across conditions than boys ($M = 0.96$, $SD = 0.20$). While girls were found to be approaching completion in all conditions ($p < .05$), boys showed lower completion overall, with the lowest performance apparent in the control condition.

When considering mastery speed, as defined by the number of questions required for problem set completion, ANOVA results suggested null results ($p > .05$) across all problem sets analyzed individually. Further, no significant results were observed when problem sets were compiled or when specific delivery elements were isolated, and there was no significant difference between messaging conditions and control. Although there was no significant difference in mastery speed across genders, trends suggested that girls had faster mastery speed in general, requiring consistently fewer questions to complete problem sets regardless of condition ($M = 4.25$, $SD = 2.65$) than boys ($M = 4.43$, $SD = 2.86$). Means and standard deviations for the full sample are presented in Table 4-4.

Table 4-4. Means and Standard Deviations for Persistence, Mastery Speed, and Survey Measures Across Control and Messaging Conditions for All Students

	Control (104[a], 99[b])		All Messaging (517[a], 506[b])		Animation (106[a], 103[b])		Static Image with Text (116[a], 113[b])		Static Image with Audio (117[a], 115[b])		Word Art (90[ab])		Audio (88[a], 85[b])	
Analysis	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Persistence	0.95	0.21	0.98	0.14	0.97	0.17	0.97	0.16	0.98	0.13	1.00	0.00	0.97	0.18
Mastery	4.74	3.35	4.32	2.67	4.24	2.69	4.62	2.83	4.32	2.42	4.28	3.33	4.09	1.91
Mindset	1.06	0.81	0.96	0.78	1.01	0.80	0.96	0.77	1.02	0.77	1.00	0.79	0.78	0.75
Enjoyment	1.83	0.80	1.67	0.89	1.74	0.87	1.66	0.90	1.77	0.82	1.49	0.91	1.67	0.96
Helpfulness	1.99	0.85	1.94	0.86	1.86	0.89	2.01	0.89	2.01	0.77	1.82	0.79	1.95	0.95

[a]Sample size for Persistence and Mastery Speed.
[b]Sample size for Mindset, Enjoyment, and Helpfulness.
Note. "Mindset" is measured by two questions (0 = Fixed Mindset, 1 = Growth Mindset) and scores are compiled. "Enjoyment" is measured by one question (Likert Scale, 0-3). "Helpfulness" is measured by one question (Likert Scale, 0-3).

ANOVA comparisons of the survey measures of mindset, enjoyment, and system helpfulness similarly conveyed null results within the full sample. The "mindset" variable was established from an average of two binary survey questions, with a composite score scaled from 0-2 representing the spectrum from fixed mindset (0) to growth mindset (2). The "enjoyment" variable was based on one question with Likert scale scores from 0-3, representing how much the student enjoyed their assignment. The "helpfulness" variable is represented in the same manner, based on the student's perception of how helpful the tutoring system was in completing their assignment. Null results were found for all three measures across problem sets when analyzed individually, and no significant differences were observed between conditions when problem sets were compiled or when specific delivery elements were isolated. Further, there was no significant difference between all messaging conditions and the control group. Gender was found to have a significant effect on enjoyment, regardless of condition $F(1, 552) = 19.50$, $p < .001$, $\eta^2 = 0.03$, with girls measuring more enjoyment on average ($M = 1.84$, $SD = 0.81$) than

boys (M = 1.52, SD = 0.90). As shown by Table 4-4, the Control was found to be the most enjoyable condition, while WordArt was enjoyed significantly less (p < .10). Gender was also approaching significance on the mindset measure, $F (1, 552) = 3.31$, $p = 0.069$, $\eta^2 = 0.01$, with boys exhibiting a lower mindset in general (M = 0.93, SD = 0.78) than girls (M = 1.05, SD = 0.77). Gender was not found to have a significant effect on student's perception of tutor helpfulness.

In an attempt to answer our third research question, elements within message delivery were collapsed based on similarity to better understand if a certain processing channel (i.e., audio) was providing the main effect for messaging results. As noted briefly in results for persistence, mastery speed, and survey measures, researchers were not able to isolate any significant differences among delivery elements (p > .05).

While few significant findings were observed in the full sample, it became clear that many students were at ceiling in the math content and therefore showing high persistence (completion) in minimum mastery speed (three consecutive correct questions). When we reassessed the sample for students operationally defined as 'struggling,' or those who required more than three questions to complete their assignments, our analysis became a bit more informative. Among 253 student assignments, no significant differences were found among conditions in persistence or mastery speed (p > .05). However, findings suggested that it took struggling students less questions on average to reach mastery when in the audio condition (M = 5.59, SD = 2.00) compared to all other conditions, as shown in Table 4-5.

When considering gender, struggling boys exhibited lower mastery in conditions including audio (p < .05) yet were found to persevere more when an image of Jane was present, while girls persevered less with the female presence (p < .05). Survey results for struggling students suggested that boys exhibited the lowest mindset measures after experiencing the control condition (p < .05), and trends suggested that regardless of condition, girls exhibited the growth mindset more consistently (M = 1.00, SD = 0.79) than boys (M = 0.91, SD = 0.75). As with the primary analysis, trends suggested that boys exhibited the growth mindset after experiencing the animation condition (p < .10). It was also found that regardless of condition, girls enjoyed their assignments (M = 1.72, SD = 0.87) significantly more than boys (M = 1.42, SD = 0.92), p < .05, and that girls consistently found the tutoring system more helpful in completing their assignment (M = 2.10, SD = 0.83) than did boys (M = 1.92, SD = 0.90).

Table 4-5. Means and Standard Deviations for Persistence, Mastery Speed, and Survey Measures Across Control and Messaging Condition for Struggling Students

	Control (46^a, 45^b)		All Messaging (207^a, 204^b)		Animation (42^a, 41^b)		Static Image with Text (49^a, 47^b)		Static Image with Audio ($49^{a,b}$)		Word Art ($28^{a,b}$)		Audio ($39^{a,b}$)	
Analysis	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Persistence	0.98	0.15	0.99	0.12	0.98	0.15	0.96	0.20	**1.00**	**0.00**	**1.00**	**0.00**	**1.00**	**0.00**
Mastery	7.07	3.95	6.34	3.32	6.17	3.48	6.84	3.24	6.14	2.88	**7.11**	**4.95**	5.59	2.00
Mindset	0.93	0.75	0.95	0.78	1.00	0.81	0.89	0.73	**1.04**	**0.82**	0.82	0.86	0.92	0.70
Enjoyment	1.60	0.86	1.58	0.94	**1.76**	**0.92**	1.45	1.00	1.71	0.79	1.43	1.07	1.51	0.97
Helpfulness	1.98	0.92	2.01	0.87	1.98	0.94	1.98	0.82	2.04	0.87	2.00	0.82	**2.05**	**0.94**

[a]Sample size for Persistence and Mastery Speed.
[b]Sample size for Mindset, Enjoyment, and Helpfulness.
Note. "Mindset" is measured by two questions (0 = Fixed Mindset, 1 = Growth Mindset) and scores are compiled. "Enjoyment" is measured by one question (Likert Scale, 0-3). "Helpfulness" is measured by one question (Likert Scale, 0-3).

Approximately 60% of students in the full sample exhibited the growth mindset in their survey responses, regardless of condition. Noting Table 4-5, students in the control condition actually reported the highest levels of growth mindset (M = 1.06, SD = 0.81), with those in the audio condition reporting the lowest levels (M = 0.78, SD = 0.75). Among struggling students, the highest levels of growth mindset were reported by students in the static image with audio condition (M = 1.04, SD = 0.82), while those in the word art condition reported the lowest levels (M = 0.82, SD = 0.86). Responses to measures of enjoyment and helpfulness followed normal distributions, with approximately 60% finding the

assignments at least "somewhat" enjoyable, and approximately 78% finding the tutoring system at least "somewhat" helpful.

DISCUSSION

Within the current study, the addition of motivational messaging to the ASSISTments tutor did not significantly affect the likelihood of student persistence or mastery speed. Further, there was little evidence that the motivational messages had the intended effect on mindset within the full sample. Trends suggested that those in messaging conditions experienced a slight increase in persistence and a decrease in mastery speed in comparison to those in the control condition. However, students in the messaging conditions also exhibited consistently lower levels for measures of mindset, enjoyment of the assignment, and perception of system helpfulness. A larger student population would be required to discern a truly significant effect within these trends.

Interestingly, struggling students appeared to benefit from the presence of messages, showing an increase in persistence, a decrease in mastery speed, and slightly increased measures of the growth mindset. It can be argued that struggling students, or those facing a challenge, are most in need of motivational interventions, and that they are more likely to respond to messaging, regardless of condition. Motivational messages produced distinctly higher adoption of the growth mindset in struggling students who experienced the static image with audio condition. Thus when designing motivational content for struggling students, current findings promote the addition of audio as an alternative processing channel to assist students. Researchers were not able to pinpoint an optimal processing channel for the delivery of growth mindset messages when targeting the general population.

One participating teacher requested that her students use a feature within the tutoring system to comment on their experience while completing their assignment. Feedback was predominantly negative, with students citing the messages as distracting or confusing. One student specifically questioned why the animated learning companion simply repeated messages rather than helping to solve the problems. This suggests that students are familiar with systems that utilize pedagogical agents, and that they have developed expectations for characters that are associated with learning. This echoes the argument set forth by Kapoor, et al. (2007) regarding the necessity for tutors to provide appropriate cognitive and affective responses, and aids in the design of tutoring systems hoping to incorporate learning companions.

This study had a variety of limitations. The ASSISTments math content chosen due to popular usage lead to a high percentage of ceiling effects within the sample. Teachers assigned multiple problem sets to their students, often as review. Thus, many students easily mastered the content intended for lower grades and thereby skewed rates of persistence and mastery speed. Further, the null effects found in the full sample raise important questions regarding the generalizability of mindset interventions outside of struggling student populations. Within the context of an adaptive mathematics tutor, students who appear to be at ceiling in math content may not require motivational messaging, and it may become detrimental to the learning process.

We also note that approximately 18.8% of students reported having technical difficulties and were removed prior to analysis. The incompatibility of simple HTML files serves as a reminder that many classrooms struggle to maintain up-to-date technological resources. Students are often required to share computers or iPads that come equipped with outdated software and generally slow internet connections. Future research should incorporate allowance for these issues within the experimental design, as incompatibilities may lead to selection bias.

It is also difficult to justify whether or not students consistently attended to the motivational messages. As students were simply presented the messages and were not asked to respond in any manner, the levels of message internalization may be broad. We also note that the duration of the intervention may have been too short to observe reliable differences among messaging conditions. In much of her work, Dweck has provided longer interventions upfront, coupled with 'reminders' such as the messages used in the current study (Dweck, 2013). Further, her studies often run longitudinally across the course of a school year or more. Still, regardless of condition, the majority of students in our sample exhibited the

growth mindset. Future research should include a pretest mindset survey to determine if these results can be credited solely to the motivational messages provided throughout the learning experience.

Finally, it should be noted that researchers relied on the tutoring system to perform random assignment. While prior research has suggested that this practice is sound, assignment for this study appears to have favored the static image with audio condition. Future research using ASSISTments should take this bias into consideration.

Future iterations of this study should focus on struggling students, or those undertaking challenging academic tasks. Future research should also seek to assess these conditions in an even more adaptive environment. It seems as though students were not reaping the benefits of the "persona effect" found in prior research (Arroyo, et al., 2013), due to a lack of bonding with the agent. A truly adaptive agent, one consistently present and building rapport, may be more effective in message delivery. Rather than repeating the same select set of general and incorrect attributions, struggling students may require motivational messages linked with the tutor content and their progress. Perhaps just as a pedagogical agent, these messages must be fine-tuned to a student's cognitive and affective states. Alternative message delivery methods, including video feedback with human tutors used as hints, scaffolding, and misconception messages, should also be considered in future research.

Chapter 5 – The Role of Student Choice in Feedback Mediums Within an Adaptive Tutoring System

While adaptive tutoring systems have improved classroom education through individualization, few platforms offer students preference in regard to their education. In the present study, a randomized controlled trial is used to investigate the effects of student choice within ASSISTments. A problem set featuring either text feedback or matched content video feedback was assigned to a sample of 82 middle school students. Those who were able to choose their feedback medium at the start of the assignment outperformed those who were randomly assigned a medium. Results suggest that even if feedback is not ultimately observed, students average significantly higher assignment scores after voicing a choice. Findings offer evidence for enhancing intrinsic motivation through the provision of choice within adaptive tutoring systems. This chapter serves as a pilot study for questioning how motivation and learning are altered when students are able to choose their own feedback medium or to voice control over their assignment within ASSISTments.

INTRODUCTION
Although the perception of autonomy has been proven as an intrinsically motivating factor for learning (Pekrun, 2006; Frenzel, Pekrun, & Goetz, 2007; Patall, Cooper, & Robinson, 2008; Murayama, Pekrun, Lichtenfeld, & vom Hofe, 2013), student preference is rarely employed in education. Perhaps traditional classroom practices have failed to capitalize on student choice due to limitations in materials or resources. However, adaptive tutoring systems offer unique opportunities for students to invest in their learning experience. These platforms are becoming a staple for the modern classroom, serving to individualize the learning experience while providing students with more powerful feedback and teachers with more powerful assessment. One of these systems, ASSISTments, is fast growing platform used for homework and classwork by over 50,000 students around the world.

The present study was influenced by Cordova & Lepper's (1996) landmark study that unveiled the beneficial effects of choice within educational computer activities. Coupled with findings from previous work surrounding feedback mediums within ASSISTments (Ostrow & Heffernan, 2014), the present study examines 1) how learning outcomes are affected if students are able to choose the feedback medium they will experience within a mathematics assignment, 2) whether a particular feedback medium is more popular or more effective, and 3) if an interaction exists between choice and feedback medium as measured by a variety of performance outcomes.

METHODS
A randomized controlled trial was designed using problem content aligned to the fifth grade Common Core State Standard of Multiplying Simple Fractions. Two isomorphic problem sets were created within ASSISTments: a set of 40 problems, each containing three hints presented as text feedback, and an isomorphic set of 40 problems, each containing three hints presented as short (15-30 second) video snippets. For each problem, regardless of feedback medium, the first two hints served as a static worked example and its solution. The third and final hint for each problem walked students through the solution to the original problem. All problem content and feedback is available at (Ostrow, 2015a) for further reference. These problem sets were then embedded in a complex experimental design within ASSISTments, establishing a solitary assignment with multiple conditions. At the beginning of the assignment each student was randomly assigned to either the Choice (experimental) or No Choice (control) conditions. Those assigned to the control were immediately reassigned to either video or text feedback. Students who were assigned to the experimental condition were asked to choose the type of feedback they wished to receive while working on their assignment. The student experience is available at (Ostrow, 2015a) for reference.

PROCEDURE
The study problem set was made openly accessible to all teachers for assignment to their students, allowing for natural and unbiased data collection. Log files were accumulated approximately one month after the release of the experiment. A total of 82 students from 4 classes spanning 2 middle schools in suburban Massachusetts had been assigned the problem set. All students within the sample were familiar with the ASSISTments platform. Of the 82 students originally assigned this problem set, 78 completed the assignment, following the distribution depicted in Figure 5-1. As shown, regardless of condition, the majority of students did not actually request hint feedback during the assignment. Thus, the results presented herein are primarily intended to guide future work.

Figure 5-1. Distribution of Students Experiencing Feedback Within Full Sample
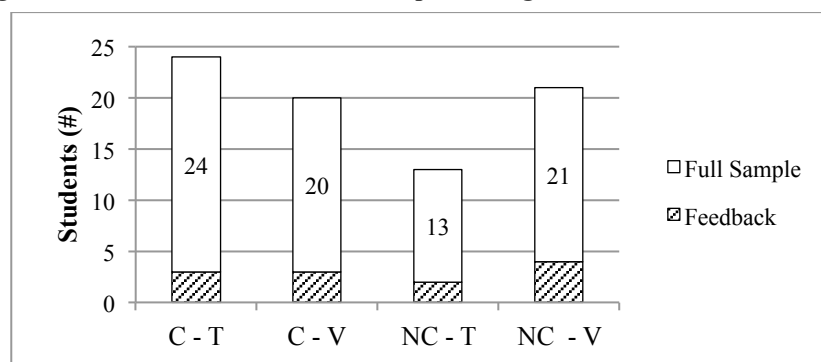
Note. Condition labeled as C (Choice, n = 44) and NC (No Choice, n = 34), Feedback Medium labeled as T (Text, n = 37) and V (Video, n = 41).

RESULTS
It was hypothesized that students would excel when provided choice, and that those receiving video feedback would outperform those receiving text feedback. A MANOVA was conducted to examine the interaction between condition and feedback medium across a number of dependent variables measuring student performance within the assignment. Within the 78 students who completed the assignment, there was no significant interaction effect, Pillai's Trace = 0.110, $F(6, 69) = 1.416$, $p = 0.221$. Further, although there was no significant main effect of condition, Pillai's Trace = 0.077, $F(6, 69) = 0.962$, $p = 0.457$, Table 5-1 reveals that students who made a preference about their feedback medium had significantly higher correctness on average than those in the control condition, $p < .05$, $\eta^2 = 0.05$. Further, students who were given choice were more likely to master their assignment than those in the control condition, trending toward significance $p < .10$, $\eta^2 = 0.04$, they used fewer hints and attempts, and spent longer working on each problem. While these findings were not significantly reliable, they support further investigation of choice within adaptive tutoring contexts. Feedback medium was less relevant to performance than hypothesized; no significant differences were observed within any of the dependent variables.

Table 5-1. Means, SDs, & Univariate Results for Main Effect of Condition (Intent-To-Treat)

Variable	n	Choice	n	No Choice	F (1,74)	p	η^2	R^2
Ave. Correctness	44	0.95 (0.10)	34	0.87 (0.25)	4.03	.048	0.05	0.05
Ave. Hints	45	0.23 (0.68)	36	0.35 (1.15)	0.61	.436	0.01	0.02
Ave. Attempts	45	3.48 (1.19)	36	3.76 (1.74)	0.89	.348	0.01	0.02
Mastery	45	1.00 (0.00)	36	0.94 (0.24)	2.83	.097	0.04	0.04
Ave. Time (sec)	44	44.94 (45.76)	34	40.29 (34.52)	0.55	.461	0.01	0.04
Med. Time (sec)	44	36.45 (42.24)	34	27.00 (16.33)	1.90	.172	0.02	0.09

Note. Averages represent average student performance across all problems experienced in the assignment.

Across the full sample, only 12 students actually requested hint feedback (14.6%). A MANOVA of treated students lacked enough power to suggest a significant interaction effect, Pillai's Trace = 0.724, $F(6, 3) = 1.31$, $p = 0.445$. The main effect of feedback medium trended toward significance, Pillai's Trace = 0.889, $F(6, 3) = 4.02$, $p = 0.141$, with students requesting more hints (M = 2.80, SD = 2.05) and using more attempts (M = 6.20, SD = 2.17) when receiving text than when receiving video (M = 1.14, SD = 0.90; M = 4.86, SD = 2.91). Further, although there was no main effect for condition, Pillai's Trace = 0.641, $F(6, 3) = 0.89$, $p = 0.588$, the means and univariate results presented in Table 5-2 suggest that students showed consistently better performance when they were able to choose their feedback medium.

Table 5-2. Means, SDs, & Univariate Results for Main Effect of Condition (Treated)

Variable	Choice, n=6	No Choice, n=6	F (1, 8)	p	η^2	R^2
Ave. Correctness	0.74 (0.02)	0.66 (0.35)	0.23	.647	0.03	0.04
Ave. Hints	1.67 (1.03)	2.00 (2.19)	0.57	.472	0.05	0.33
Ave. Attempts	5.83 (1.72)	5.00 (3.41)	0.02	.895	0.00	0.30
Mastery	1.00 (0.00)	0.83 (0.41)	0.47	.512	0.05	0.18
Ave. Time (sec)	24.72 (10.14)	59.26 (37.92)	3.99	.081	0.33	0.34
Med. Time (sec)	14.52 (5.93)	35.30 (23.86)	3.49	.099	0.30	0.31

DISCUSSION
This study served as an initial foray into implementing student choice within ASSISTments, an adaptive tutoring platform that was previously unable to individualize learning via student preference. Results suggested that students who were able to invest in their learning experience outperformed those who were not asked their preference. Those provided choice averaged higher correctness on the assignment while using fewer hints and attempts. Further, choice significantly impacted performance, even when the outcome of choosing was not ultimately experienced. Aside from small sample size, this study was also

somewhat limited in that the experimental design utilized feedback that was only provided upon the student's request. As such, proper analysis of main effects would require a much larger treated sample. The results of this study inspired infrastructure changes within the ASSISTments platform that will allow for future research in this area. Similar hypotheses can now be examined using ASSISTments on larger samples and within additional content domains. Findings offer evidence in support of allowing student autonomy within adaptive education.

## Chapter 6 – Understanding the Effects of Student Choice at Scale

Shortly after publication of the work in Chapter 5, production updates effectively broke the experimental design and kept students from being properly assigned to conditions. Given the results that were observed through pilot analysis, a second version of the experiment presented in Chapter 5 was established using the If-Then-Else navigator recently developed by Christopher Donnelly and David Magid. The work in Chapter 5 presented the need for this type of path navigation within ASSISTments, that would allow for expansion of these and similar hypotheses at scale and across domains. This navigator was developed under the SI2 NSF grant (SI2-SSE&SSI: 1440753) helping to build out the platform as a shared scientific instrument for a collaborative of external researchers. This navigator will provide a variety of opportunities to research student choice in a more valid manner than the work already presented, and routing paths will allow the system to offer truly adaptive learning paths based on student performance or logged characteristics. Detailed specifications of this navigator, including its current state and plans for future design implications, can be found in Mr. Donnelly's thesis proposal. However, to better understand how the choice study presented in Chapter 5 was reestablished using this infrastructure, a summary is provided in the following section. Thus, the present study seeks to confirm and expand upon the findings presented in Chapter 5, reiterating the research questions:
1. What feedback medium do students choose most often, and is it producing optimal learning gains? Are students able to effectively judge the type of feedback that would benefit them the most?
2. Does student choice impact learning gains and is the effect different across feedback mediums?

INFRASTRUCTURE
A problem set using the navigator is shown in Figure 6-1 below. Essentially, this section type requires three elements: a conditional statement, a true path, and a false path. The conditional statement can be fulfilled by a problem, problem set, or separate section, with a setting that can be manipulated to guide path routing based on various percentages of completion or correctness. If the threshold is met, students are routed into the true path, or the second element in Figure 6-1, labeled "Video Chosen." If the threshold is not met, students are routed into the false path, or the third element in Figure 6-1, labeled "Text Chosen." In this example, if the conditional statement is a preference question, in line with providing students a choice in their feedback medium, one path is set to be the 'correct' answer while the alternate path is set to be the 'incorrect' answer within the conditional problem. The problem is presented in test mode (i.e., without correctness feedback) and therefore the student has no knowledge of the inner workings of the routing system and does not feel penalized for their choice. If video feedback is ultimately the 'true' path, students- choosing video will be marked 'correct' and routed to the proper section of the problem set.

Figure 6-1. The If-Then-Else Navigator Used in a New Iteration of the Work Presented in Chapter 5
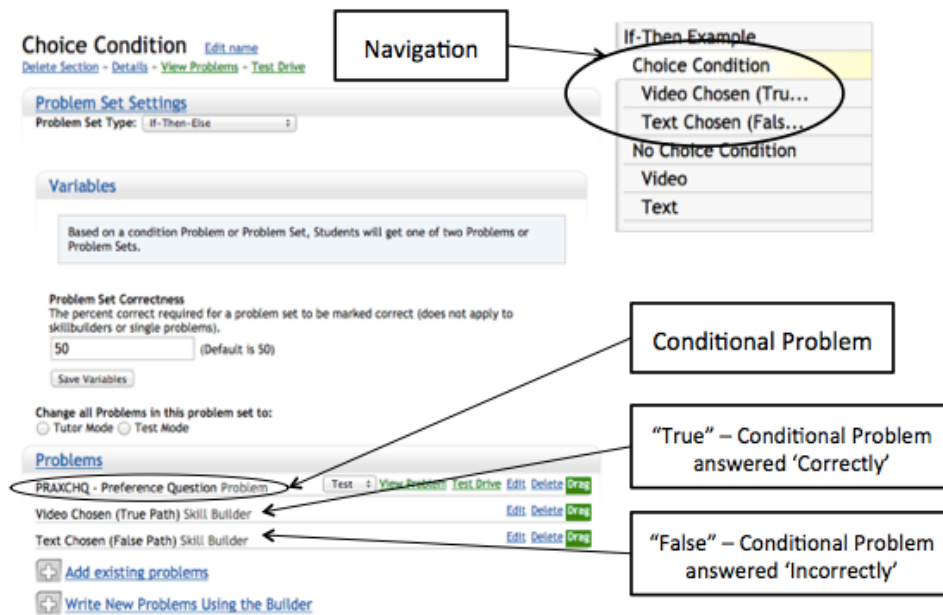

Early testing has shown that the navigator will be useful in a number of complex designs. It is possible to embed multiple If-Then-Else navigators to create a series of paths that solve issues that were previously complicated matters within ASSISTments, including the routine provision of a video check that isolates students who have technical difficulties. In the studies explained in previous chapters, students who reported technical difficulties simply had to wade their way through a disjointed assignment if the tutor randomly assigned them to a condition they were not able to access. In in some cases, video checks were impossible due to the complex structure of the experimental design. However, with the If-Then-Else navigator it is possible to provide a video check as the conditional element, with students who report technical difficulties being automatically routed out of the experiment yet allowing them to receive a proper 'default' assignment. Students who are able to view the video content can be routed into the experiment, and in the case of the present choice study (a stronger implementation of the work presented in Chapter 5) the student would then be randomly assigned into either the control (no choice) or experimental (choice) conditions. Additionally, it is now possible using the If-Then-Else navigator to easily embed complex experiments within a more structured linear design between pretest and posttest sections. This was previously a somewhat limited feature, available when using a simple design and the Choose Condition navigator, a section type that essentially conducts random assignment into any number of subsections. Thus, the If-Then-Else navigator works in many ways to allow researchers to conduct studies more seamlessly within the classroom, without hampering the learning of a single child.

METHODS
Building on the design presented in Chapter 5, this second iteration of the study employed the If-Then-Else navigator to strengthen the validity of this work and scale up the design to a larger sample population. Content was kept identical to that of the pilot study, refashioning ASSISTments Certified material on the fifth grade Common Core State Standard of Multiplying Simple Fractions. The design implemented using the If-Then-Else navigator is presented in Figure 6-1 below. As in Chapter 5, for each condition (regardless of medium), feedback was presented using hints, with the first two hints providing a worked example of a similar problem, and the final hint working through the current problem and providing the answer.

Figure 6-2. Experimental Design Established Using If-Then-Else Navigator
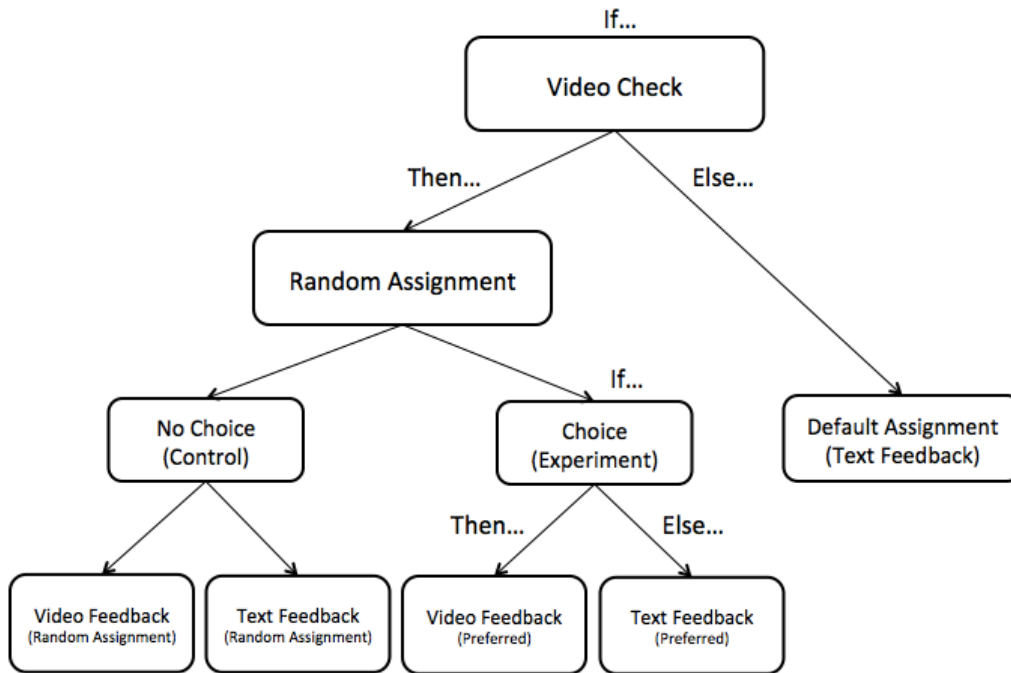
As shown in Figure 6-2, when students began their assignment they were subjected to a standard video check. Those who failed the video check were defaulted to an alternative assignment and excluded from the study. Those who passed the video check by confirming their ability to see video content were randomly assigned to either the choice (experiment) or no choice (control) conditions. This design boils down to a simple 2x2 factorial design with conditions that are identical short of the student preference element. Those who were randomly assigned to the control condition were not offered choice, but rather, were immediately subjected to another random assignment. These students were routed into either a purely video feedback or purely text feedback version of the assignment and were not polled on their preference of feedback style. Students originally assigned to the experimental condition were asked what type of feedback they would like to receive while working on their assignment, using the conditional question of another If-Then structure presented in Figure 6-3. Those who chose video were appropriately routed into the version of the Skill Builder containing simple fraction multiplication problems with video feedback, while those who choose text feedback were routed into the set containing text feedback.

Figure 6-3. Conditional Problem Controlling Path Routing for Students with Choice

The first two steps of feedback for each question, regardless of medium, were a static worked example identical across each problem, with the final hint providing the steps to work through the current problem. Video problems were recorded using Educreations, an iPad app that allows for the creation, storage, and distribution of educational video content (Educreations, 2014). A total of 42 videos were recorded and embedded as hints. Educreations was used in this study rather than YouTube, as it has been observed that many schools block YouTube content causing greater potential for technical difficulties. Cues and colors were kept as similar as possible in the transition across mediums, as shown in Figure 6-4. However, it should be noted that students using iPads were prompted to view the videos in the application (if already downloaded) or to download the application (often a blocked feature) and thus allowed a new type of technical difficulty.

Figure 6-4. Example of Question and Feedback for Isomorphic Problems of Each Medium

RESULTS

As this study has only been running using the If-Then-Else navigator for approximately one month, only 58 students have completed the assignment. Five of these students reported not having access to video, and were therefore excluded from analysis, leaving a sample of only 53 students thus far. These students come from two middle school classes at two different schools in New England. Primary analysis explores various metrics of student performance, similar to those considered in Chapter 5. The main effect of choice is examined across variables for the full sample (those 'treated' or placed into either 'choice' (n = 31) or 'no choice' (n = 22) conditions). Findings are summarized in Table 6-1, and discussed further in the following subsections. To examine the main effect of feedback medium, only students who actually used feedback by requesting hints are considered (those 'treated' with either video feedback (n = 5) or

Figure 6-5. Distribution of Sample (Treated/Distributed)
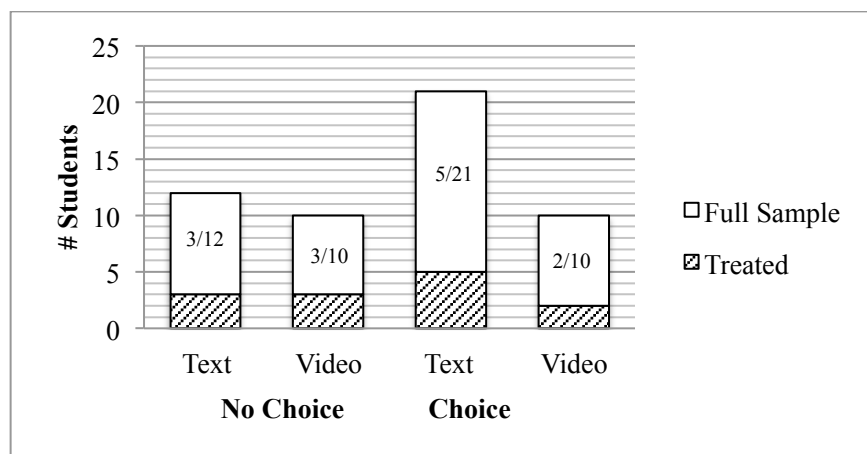
Table 6-1. Means, SDs, and ANOVA Results for Main Effects of Choice in Full Sample

	Choice *(n = 31)*	No Choice *(n = 22)*			
Dependent Variable	*M (SD)*	*M (SD)*	*F*	*p*	η^2
Problems Seen	4.16 (1.90)	3.91 (1.34)	0.286	0.595	0.006
Overall Score	0.84 (0.22)	0.89 (0.15)	0.965	0.330	0.019
Partial Credit Score	0.94 (0.12)	0.97 (0.05)	0.823	0.369	0.016
Total Hints	0.26 (0.51)	0.41 (0.80)	0.704	0.405	0.014
Average Hint Usage	0.05 (0.09)	0.08 (0.14)	1.069	0.306	0.021
Total Attempts	5.52 (3.58)	5.14 (3.11)	0.161	0.690	0.003
Average Attempts	1.35 (0.94)	1.24 (0.41)	0.262	0.611	0.005

Table 6-2. Means, SDs, and ANOVA Results for Main Effects of Feedback Medium in Treated Sample

	Video *(n = 5)*	Text *(n = 8)*			
Dependent Variable	*M (SD)*	*M (SD)*	*F*	*p*	η^2
Problems Seen	4.80 (0.84)	6.00 (2.00)	1.582	0.234	0.126
Overall Score	0.79 (0.04)	0.68 (0.09)	6.211	0.030	0.365
Partial Credit Score	0.91 (0.06)	0.92 (0.03)	0.268	0.615	0.024
Total Hints	1.20 (0.45)	1.38 (0.74)	0.222	0.647	0.020
Average Hint Usage	0.25 (0.09)	0.24 (0.10)	0.115	0.741	0.011
Total Attempts	7.80 (2.77)	9.00 (3.21)	0.474	0.505	0.041
Average Attempts	1.64 (0.54)	1.53 (0.36)	0.200	0.663	0.018

text feedback (n = 8)). These findings are summarized in Table 6-2, and discussed further in the following subsections. Interactions between choice and feedback medium are considered within the treated sample and discussed for particular variables, but should be taken with caution considering the low number of students in each condition within the 2x2 factorial design, as shown in Figure 6-5.

Problems Seen
When examining the number of problems seen, the main effect of choice was not significant within the full sample, $F(1, 51) = 0.286$, $p > .05$. Students with choice saw more problems ($M = 4.16$, $SD = 1.90$) than students without choice ($M = 3.91$, $SD = 1.34$) but these differences were not reliable, $t = -.535$, $p > .05$. In treated students, the main effect of feedback medium was not significant, $F(1, 10) = 1.582$, $p > .05$. Students in the text condition saw more problems ($M = 6.00$, $SD = 2.00$) than those in the video condition ($M = 4.80$, $SD = 0.84$), although this difference was not reliable, $t = 1.258$, $p > .05$. Examining the interaction between feedback medium and choice condition within treated students, results were not significant, $F(1, 9) = 0.124$, $p > .05$. Although differences were not significantly reliable, the mean differences presented in Figure 6-6 suggest that students saw more problems when given choice, and fewer problems when given video.

Figure 6-6. Problems Seen as a Function of Feedback Medium and Choice within Treated Sample
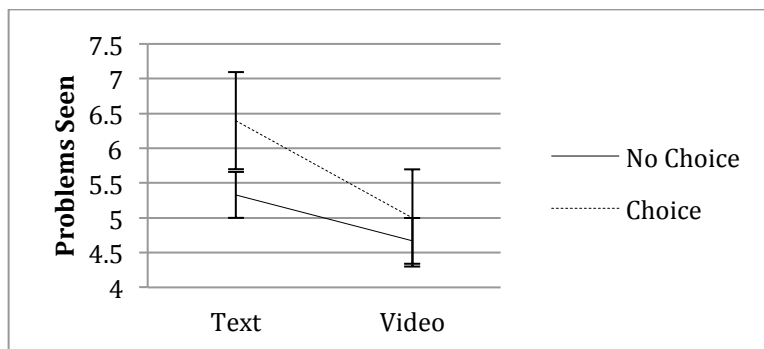
Overall Score
Considering all students in the sample, the main effect of choice was not significant, $F(1, 51) = 0.965$, $p > .05$. Students with choice did mildly worse ($M = 0.84$, $SD = 0.22$, $n = 31$) than students without choice ($M = 0.89$, $SD = 0.15$, $n = 22$) but these differences were not reliable, $t = 0.983$, $p > .05$. In treated students, the main effect of feedback medium was significant, $F(1, 10) = 5.287$, $p < .05$. An analysis of means revealed that those in text condition ($M = 0.68$, $SD = 0.09$, $n = 8$) did significantly worse than those in the video condition ($M = 0.79$, $SD = 0.04$, $n = 5$) as measured by average correctness across all problems in the assignment, $t = -2.492$, $p < .05$. When attempting to model the interaction between feedback medium and choice within the sample of treated students, the interaction effect was not significant, $F(1, 9) = 0.044$, $p > .05$. An analysis of means revealed that subtle differences do exist in performance, with students who chose video ($M = 0.79$, $SD = 0.06$) outperforming students in all other groups as shown Figure 6-7 below. However, these differences were not substantial enough to result in a significant interaction, as standard errors largely overlap.

Figure 6-7. Overall Score as a Function of Feedback Medium and Choice within Treated Sample
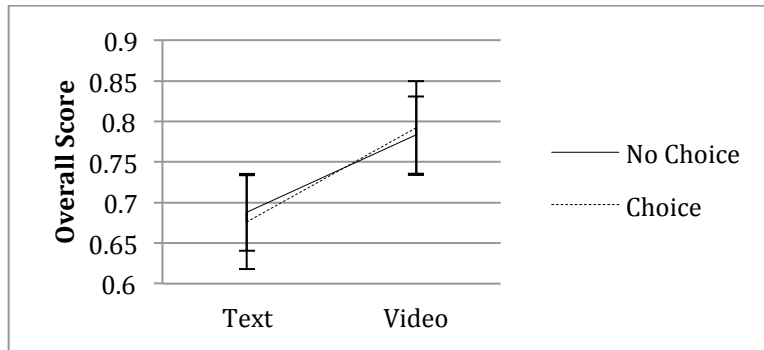
Partial Credit Score

When examining partial credit score, or the average of partial credit across problems solved, within the full sample the main effect of choice was not significant within the full sample, $F(1, 51) = 0.823$, $p > .05$. Students with choice (M = 0.94, SD = 0.12) underperformed students without choice (M = 0.97, SD = 0.05) but these differences were not reliable, $t = 0.907$, $p > .05$. In treated students, the main effect of feedback medium was not significant, $F(1, 10) = 0.268$, $p > .05$. Students in the text condition (M = 0.92, SD = 0.03) performed approximately the same as those in the video condition (M = 0.91, SD = 0.06), $t = 0.518$, $p > .05$. However, when modeling the interaction of choice and feedback medium within treated students, the effect was significant, $F(1, 9) = 6.563$, $p < .05$, $\eta^2 = 0.422$. An analysis of means revealed that those who chose video (M = 0.97, SD = 0.02) outperformed all other groups, as shown in Figure 6-8. When comparing these results to the overall score, it is clear that partial credit assessment would allow for more efficient group differentiation while simultaneously allowing students to earn more credit for their work, acting as a motivating force for learning. Scores are higher when partial credit is implemented, with students in the text condition seeing an average gain of 24%, and those in the video condition seeing an average gain of 12%. Further, partial credit allows us to differentiate between students who received choice in the video condition, as shown in Figure 6-8.

Figure 6-8. Partial Credit Score as a Function of Feedback Medium and Choice within Treated Sample
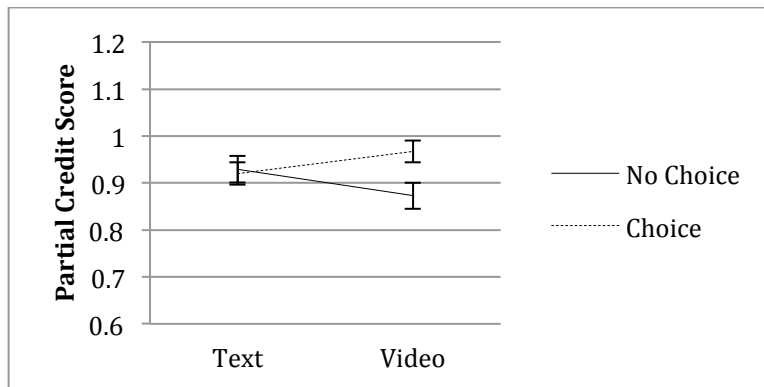
Total Hints Used

Considering the total number of hints used within the full sample, the main effect of choice was not significant, $F(1, 51) = 0.704$, $p > .05$. Students with choice (M = 0.26, SD = 0.51) used less hints than students without choice (M = 0.41, SD = 0.80) but these differences were not reliable, $t = 0.839$, $p > .05$. In treated students, the main effect of feedback medium was not significant, $F(1, 10) = 0.222$, $p > .05$. Students in the text condition (M = 1.38, SD = 0.74) used more hints than those in the video condition (M = 1.20, SD = 0.45), although this difference was not reliable, $t = 0.471$, $p > .05$. Within treated students,

there was no significant interaction of choice and feedback medium, $F(1, 9) = 0.028$, $p > .05$. Means for total hint usage within the treated sample are presented in Figure 6-9.

Figure 6-9. Total Hints as a Function of Feedback Medium and Choice within Treated Sample
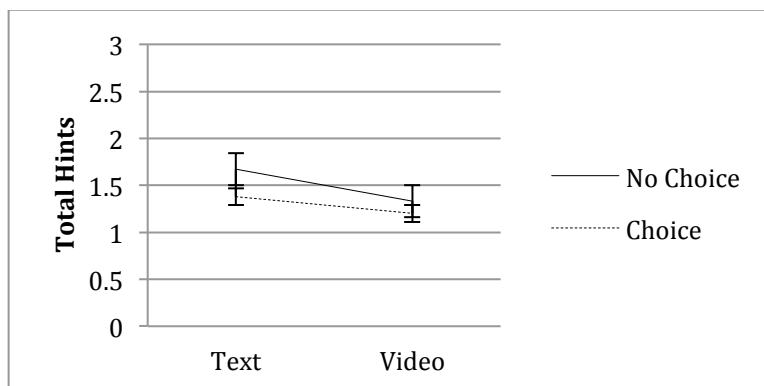
Average Hint Usage
Considering the total number of hints used within the full sample, the main effect of choice was not significant, $F(1, 51) = 1.069$, $p > .05$. Students with choice (M = 0.05, SD = 0.09) used less hints per problem than students without choice (M = 0.08, SD = 0.14) but these differences were not reliable, $t = 1.034$, $p > .05$. In treated students, the main effect of feedback medium was not significant, $F(1, 10) = 0.115$, $p > .05$. Students in the text condition (M = 0.24, SD = 0.10) used approximately the same amount of hints per problem as those in the video condition (M = 0.25, SD = 0.09), $t = -0.339$, $p > .05$. Within treated students, the main effect of choice neared significance, $F(1, 10) = 3.077$, $p = .110$. Those without choice used more hints on average (M = 0.29, SD = 0.10) than those with choice (M = 0.20, SD = 0.06), with results nearing significance, $t = 1.808$, $p = 0.087$. There was no significant interaction between choice condition and feedback medium, $F(1,9) = 0.026$, $p > .05$. Means for average hint usage within the treated sample are presented in Figure 6-10.

Figure 6-10. Average Hint Usage as a Function of Feedback Medium and Choice within Treated Sample
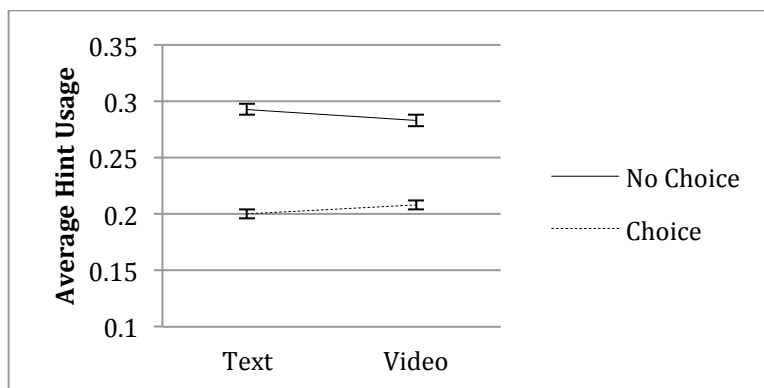
Total Attempts
Considering the total number of attempts used within the full sample, the main effect of choice was not significant, $F(1, 51) = 0.161$, $p > .05$. Students with choice (M = 5.52, SD = 3.58) made more attempts throughout their assignment than students without choice (M = 5.14, SD = 3.11) but these differences were not reliable, $t = -0.402$, $p > .05$. In treated students, the main effect of feedback medium was not significant, $F(1, 10) = 0.474$, $p > .05$. Students in the text condition (M = 9.00, SD = 3.21) used more attempts than those in the video condition (M = 7.80, SD = 2.77), although this difference was not found

to be reliable, t = .689, p > .05, possibly due to low sample sizes. Within treated students, the interaction effect for feedback medium crossed with choice condition trended toward significance, $F(1, 9) = 3.873$, p = .081, $\eta^2 = 0.301$. Analysis of means revealed that students who chose text feedback made the most attempts of all groups (M = 10.00, SD = 3.39), whereas those who chose video made the least attempts (M = 5.50, SD = 0.71). This relationship is presented in Figure 6-11.

Figure 6-11. Total Attempts as a Function of Feedback Medium and Choice within Treated Sample
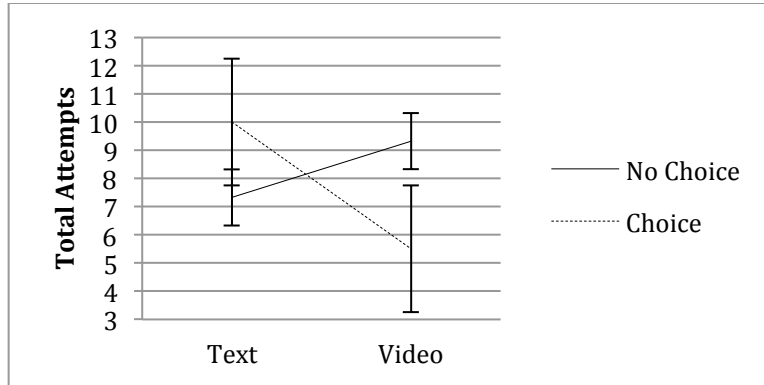
Average Attempts
Considering the average number of attempts used per problem within the full sample, the main effect of choice was not significant, $F(1, 51) = 0.262$, p > .05. Students with choice (M = 1.35, SD = 0.94) made more attempts on average than students without choice (M = 1.24, SD = 0.41) but these differences were not reliable, t = -0.512, p > .05. In treated students, the main effect of feedback medium was not significant, $F(1, 10) = 0.200$, p > .05. Students in the text condition (M = 1.53, SD = 0.36) used fewer attempts on average than those in the video condition (M = 1.64, SD = 0.54), but this difference was not found to be reliable, t = -.447, p > .05. Within treated students, the interaction effect for feedback medium crossed with choice condition was significant, $F(1, 9) = 6.166$, p < .05, $\eta^2 = 0.407$. Students who chose video used less attempts on average (M = 1.13, SD = 0.18) than those who were randomly assigned video (M = 1.60, SD = 0.28). Interestingly, those who chose text used more attempts (M = 1.98, SD = 0.36) than those who were randomly assigned to text (M = 1.42, SD = 0.52). An analysis of means is depicted in Figure 6-12.

Figure 6-12. Average Attempts as a Function of Feedback Medium and Choice within Treated Sample
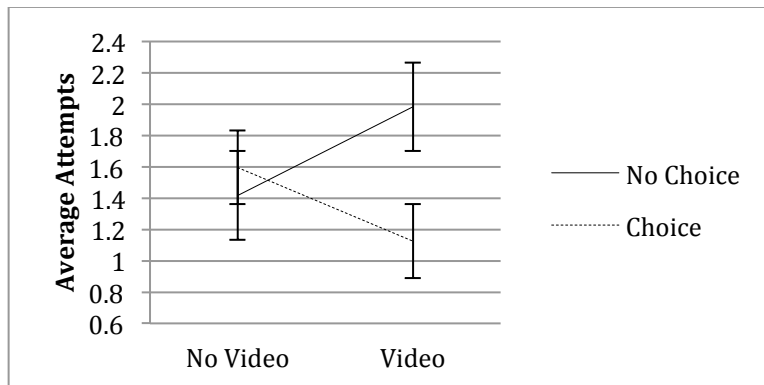
DISCUSSION

Results of this study are preliminary in nature, given that only 58 students have completed the assignment thus far. Further, the distribution of treated students is currently about as low as that observed in the pilot study presented in Chapter 5. However, it is interesting to note that observations made within this sample were slightly different than those presented in Chapter 5. Choice no longer plays a significant role in assignment performance. It is possible that for high performing students who are able to self-regulate well, the addition of choice to an assignment can be beneficial even if the effects of making that choice are not observed (i.e., observed in Chapter 5). It is also possible that for lower performing students who have difficulties self-regulating, the addition of choice could be deleterious. Larger samples and additional iterations of this work across new skill topics will be required before any reliable conclusions can be drawn.

Interestingly, a significant main effect of feedback medium was observed in this sample, channeling results similar to those observed in the study presented in Chapter 2. Video feedback was found to be significantly more effective that text feedback, with students earning enhanced scores while using fewer levels of feedback and ultimately requiring fewer problems. It is difficult to draw a clear understanding of this effect within a sample of 13 students, and the distribution of students across the 2x2 design is such that interaction effects between choice and feedback medium should not be considered until the sample size increases. It is also interesting to note however, that two thirds of students assigned to the choice condition opted for text feedback, a trend similar to that observed in the pilot.

The work presented in Chapter 5 took the approach of a MANOVA, considering all elements of performance together before breaking the analysis down into univariate effects. The present exploratory analysis used solitary ANOVAS to investigate each performance variable, as some of the variables examined were highly correlated (i.e., total attempts vs. average attempts used). The study presented in Chapter 5 did not consider total attempts, total hint usage, or partial credit scoring. Still, running multiple ANOVAs here may have increased the likelihood of Type I error, and the results observed may include false positives. A more sound analysis should be employed when data is recollected pending a large sample. Additionally, future work should include additional iterations of this study using different skill topics and video structures to determine if the effect of video is context dependent.

Chapter 7 – Blocking vs. Interleaving: Examining Single-Session Effects within Middle School Math Homework

The benefit of interleaving cognitive content has gained attention in recent years, specifically in mathematics education. The present study serves as a conceptual replication of previous work, documenting the interleaving effect within a middle school sample through brief homework assignments completed within ASSISTments, an adaptive tutoring platform. The results of a randomized controlled trial are presented, examining a practice session featuring interleaved or blocked content spanning three skills: Complementary and Supplementary Angles, Surface Area of a Pyramid, and Compound Probability without Replacement. A second homework session served as a delayed posttest. Tutor log files are analyzed to track student performance and to establish a metric of global mathematics skill for each student. Findings suggest that interleaving is beneficial in the context of adaptive tutoring systems when considering learning gains and average hint usage at posttest. These observations were especially relevant for low skill students.

This chapter has been published at the following venue:
Ostrow, K., Heffernan, N., Heffernan, C., & Peterson, Z. (In Press). Blocking vs. Interleaving: Examining Single-Session Effects within Middle School Math Homework. To be included in Conati, C., Heffernan,

N., Mitrovic, A., & Verdejo, M. (Eds.) Proceedings of the 17th International Conference for Artificial Intelligence in Educations (AIED). Madrid, Spain. pp. forthcoming.

INTRODUCTION

The benefit of interleaving cognitive content has gained attention in recent years. A simple intervention rooted in kinesthetic research pertaining to the acquisition of motor skills (Shea & Morgan, 1979), interleaving has since evolved into a powerful tool for the modern classroom. Specifically, significant effects have been verified in the realm of mathematics education in classroom trials and through simulated studies (Mayfield & Chase, 2002; Rohrer & Taylor, 2007; LeBlanc & Simon, 2008; Taylor & Rohrer, 2010; Li, Cohen, & Koedinger, 2012). Research within this realm has examined the interleaving effect by mixing or alternating the delivery of skill or problem content, such that similar problems are no longer 'blocked' or presented in uniform segments. The benefits observed when interleaving mathematics content are often credited to the discriminative-contrast hypothesis (Birnbaum, et al., 2013), which purports that the effect is rooted in a student's enhanced ability to pinpoint differences in problem content. As such, interleaving provides an obvious tool within a domain that relies largely on problem type identification and solution strategy choice (Rohrer, 2012).

Despite this clarity, the details of interleaving remain somewhat obscure. It is heavily documented that interleaving is confounded by an inherent spacing effect (Rohrer, 2012), yet few researchers effectively isolate interleaving by examining a single session or controlling for the spacing of content (Taylor & Rohrer, 2010). Researchers have also added complexity to the issue, questioning which dimension of cognitive content (i.e., the skill, the task type, the representation, etc.) to interleave for optimal results [(Rau, Aleven, & Rummel, 2013; Rau, et al., 2012). Further, despite continued reports of significant learning gains observed at posttest after interleaved practice, policymakers and educational designers fail to interleave mass-produced content, claiming that it is detrimental to the student's learning *experience* (Rohrer & Pashler, 2010; Taylor & Rohrer, 2010; Kornell & Bjork, 2008). Essentially, the practice has earned a bad reputation for making the learning process more complex, or for adding what Bjork terms 'desirable difficulty' (Bjork, 1994).

The present study serves as a conceptual replication of Rohrer & Taylor's (2007) work on shuffling mathematics practice problems. While replications are rare in general (Roediger, 2012), a recent analysis of leading education journals found that less than 0.13% of publications were replications (Makel & Plucker, 2014). However, repeated observations of significant educational findings, especially within different contexts, have the power to produce systemic change. While not a direct replication, we similarly aim to assess the interleaving effect within mathematics skills amidst a single practice session, considering delayed posttest measures as dependent metrics. More uniquely, we seek to document the effect using a brief homework assignment completed within ASSISTments, an online adaptive tutoring system. We also consider a global metric of mathematics skill for each student, in and attempt to gauge how the effect differs across skill level.

ASSISTments is fast growing platform offered as a free service of Worcester Polytechnic Institute and used for homework and classwork by over 50,000 students around the world (Heffernan & Heffernan, 2014). The system offers teachers a library of prebuilt content, primarily with a focus on mathematics skills aligned to the Common Core State Standards, as well as the ability to build content to match their curriculum or course goals. Simultaneously, students benefit from correctness feedback and tutoring strategies within an adaptive environment that advances skill practice beyond that achieved through traditional classroom practices. ASSISTments also serves as shared scientific tool for education research (Heffernan & Heffernan, 2014). Adaptive tutoring systems provide a natural learning environment from which to assess best practices, and yet, to our knowledge, little work has been done to examine interleaving within these settings. Thus, a randomized controlled trial was designed within ASSISTments to examine the subtleties of interleaving, as guided by the following research questions:

1. When controlling for student skill level, do learning gains (as measured by average posttest score) differ when practice session content is interleaved?

2. When controlling for student skill level, does interleaving practice session content lead students to interact differently with the system at posttest (as measured by average hint usage and average attempt count)?

It was hypothesized that interleaving skill content in the practice session would have a beneficial effect on student performance as measured at posttest, leading to increases in posttest score and reductions in the average number of hints and attempts used during posttest problems.

METHODS

This study was conducted with five classes spanning three teachers at a suburban middle school in Massachusetts. All teachers and students within the sample population were familiar with ASSISTments, having used the system for classwork and homework throughout the school year. Researchers worked with a participating teacher to design problem content for two homework assignments (i.e., a practice session and a delayed posttest). In April 2014, the teacher isolated three mathematics skills that her students had learned earlier in the year to serve as review while allowing for the observation of relearning via hint usage. The skills covered were Complementary/Supplementary Angles (Skill A; originally covered in February/March 2014), Surface Area of a Pyramid (Skill B; originally covered in November/December 2013), and Probability of Compound Events without Replacement (Skill C; originally covered in January 2014). A problem exemplifying Skill B with all available hint feedback is provided in Figure 7-1; problems exemplifying Skills A and C can be accessed at Ostrow (2015b) for further reference.

Figure 7-1. Example of Skill B, Surface Area of a Pyramid

For the practice session, four problems were created for each skill, resulting in a single assignment with twelve problems. These problems were isomorphic in structure, but designed such that problem difficulty would increase with each practice opportunity. Hence, a student's first experience with Skill A was relatively easy, while her fourth experience with the skill was more challenging. One additional problem was created for each skill, matching the highest difficulty level presented during practice, to establish a separate, three-problem assignment that would serve as a delayed posttest. Practice and posttest sessions were both assigned as homework, establishing an authentic learning experience and

reducing the potential for immediate assistance from and adult. Settings for homework completion were ultimately unknown and were likely differential across students.

Further, although straying from the conventions of a 'formal' posttest, permitting the use of hints and multiple attempts during the posttest assignment allowed researchers to investigate variables of student performance extending beyond average posttest score (i.e., an average of the student's accuracy on their first attempt at solving each problem).

PROCEDURE

After the creation and release of content, five teachers assigned this study to an initial sample of 226 7th grade students. Students were randomly assigned to either the experimental condition, in which skill problems within the practice assignment were presented in an interleaved or mixed pattern, or to the control condition, in which skill problems within the practice assignment were presented using a blocked approach. Random assignment was accomplished using a pseudo-random number generator within the ASSISTments tutor, and occurred at the student level rather than the class level to control for potential teacher and class effects.

Regardless of condition, students received the same twelve problems during the practice session, with the only difference being presentation order. Problem delivery patterns for each group are depicted in Figure 7-2. Using this design, the effects of interleaving were not specifically isolated from the effects of spacing. For instance, students in the interleaved condition experienced problem A_4 at a later point in time than students in the blocked condition. However, the practice session was delivered as a single assignment in an attempt to minimize the effects of spacing (Rohrer & Pashler, 2010; Cepeda et al., 2006).

Regardless of condition, all students received a second homework assignment consisting of three problems in a static delivery pattern, serving as a delayed posttest. Participating teachers assigned this posttest anywhere from two to five days following the practice session. Details pertaining to the design of this study, including access to question content and the student experience can be found at Ostrow (2015b).

Figure 7-2. Experimental Design: Skill Problem Delivery Across Groups

Blocked	$A_1, A_2, A_3, A_4, B_1, B_2, B_3, B_4, C_1, C_2, C_3, C_4$
Interleaved	$A_1, A_2, B_1, B_2, C_1, C_2, A_3, B_3, C_3, B_4, C_4, A_4$
Posttest	A_5, B_5, C_5

Tutor log files were retrieved from the ASSISTments database and problem level data, including correctness, hint usage, and attempt count was isolated for each student. Using previously logged data, it was also possible to calculate a global metric of mathematics skill for each student based on the average accuracy of all problems he or she had ever completed within the system. This measure was then discretized using a median split to bin students as generally 'high' or 'low' skill.

Within the initial sample of 226 students assigned the practice session, one participating teacher failed to assign the posttest, resulting in the removal of 68 students from final analysis. Of the remaining 158 students, three students failed to complete enough of the practice session to verify their condition based on logged data, and were therefore excluded from analysis. Additionally, nine low-skill students failed to start the posttest assignment. Further assessment of these nine students revealed that six had experienced the blocked condition during the practice session, while three had experienced the interleaved condition. Only five of these students completed the practice session, with four students failing to complete the blocked session and one student failing to complete the interleaved session. A two-tailed independent t-test was performed to compare the number of practice session problems completed by these students across groups, revealing that condition was not a significant factor in disparate completion rate, t

= 0.048, p = .963. These nine students were therefore excluded from posttest analysis without introducing an obvious bias.

A Chi-squared test of independence of the remaining 146 students did not indicate a significant relationship between condition and student skill level, χ^2 (1, N = 146) = 0.195, p > .05. However, the distribution across conditions was not equivalent (Blocked, n = 60; Interleaved, n = 86) due to the pseudo-random number generator that conducted student level randomization. Given the successful use of this assignment method in previous research, the authors had no reason to believe that a selection effect had occurred or that this process was in any way biased (i.e., affected by specific student characteristics). Thus, the skewed distribution observed here was not regarded as a threat to validity. The log files discussed herein have been stripped of identifiers and are available at Ostrow (2015b) for further reference.

RESULTS

To examine our first research question, an ANCOVA was performed to analyze average posttest score across conditions when controlling for student skill level. Within 146 students, after controlling for the effect of student skill level, the effect of condition on posttest score trended toward significance, $F(1,143)$ = 2.69, p = 0.103, η^2 = 0.02, Hedge's g = 0.22. As a covariate, student skill level was significantly related to posttest score, $F(1, 143)$ = 29.308, p < .001, η^2 = 0.17. Levene's test was not significant, p > .05, and thus error variance was assumed to be equal across conditions. A summary of the effects of condition on average posttest score is depicted in Table 7-1. Analysis of means revealed that students in the interleaved condition **(**M = 0.67, SD = 0.27, n = 86) outperformed those in the blocked condition **(**M = 0.61, SD = 0.27, n = 60).

Split file ANOVAs were conducted to further examine the effect of condition across student skill level. For low skill students, condition had a significant effect on average posttest score, $F(1, 62)$ = 5.59, p < .05, η^2 = 0.08, Hedge's g = 0.60. Levene's test was significant, $F(1, 62)$ = 5.16, p < .05 suggesting the assumption of equivalent variance has been violated. Analysis of means revealed that students in the interleaved condition (M = 0.58, SD = 0.29, n = 39) significantly out performed those in the blocked condition (M = 0.42, SD = 0.23, n = 25). Within high skill students, condition no longer had a significant effect on posttest score, $F(1, 80)$ = 0.01, p > .05. Students in the interleaved condition (M = 0.74, SD = 0.23, n = 47) performed quite similarly to those in the blocked condition (M = 0.74, SD = 0.21, n = 35). Summaries of the effects of condition on average posttest score for both skill levels are presented in Table 7-2. Figure 7-3 depicts the interaction of condition and student skill level observed in average posttest score.

To examine our second research question, a MANCOVA was used to analyze the dependent measures of average posttest hint usage and average posttest attempt count as a function of condition after controlling for student skill level. Pillai's Trace is reported throughout, as the assumption of equality of covariance matrices was violated and this parameter offers a more robust understanding of variance. Within 146 students, after controlling for the effect of student skill level, there was a significant main effect of condition, Pillai's Trace = 0.06, F (2, 142) = 4.81, p < 0.05. At the multivariate level, student skill level was significant as a covariate, Pillai's Trace = 0.36, F (2, 142) = 39.25, p < .001, explaining approximately 36% of the total variance. Tests of between subjects effects revealed that condition had a significant effect on average posttest hint usage, $F(1, 143)$ = 6.24, p < .05, η^2 = 0.03, Hedge's g = -0.29. Students in the interleaved condition used significantly less hints on average (M = 0.33, SD = 0.57, n = 86) than those in the blocked condition (M = 0.50, SD = 0.64, n = 60). However, condition did not significantly affect average posttest attempt count, $F(1, 143)$ = 0.10, p > .05**,** with those in the interleaved condition (M = 1.75, SD = 1.08, n = 86) and those in the blocked condition (M = 1.68, SD = 0.57, n = 60) using a similar amount of attempts. A summary of univariate results is presented in Table 7-3.

Table 7-1. ANCOVA of the Effects of Condition on Average Posttest Score

Source	df	SS	MS	F	p	η^2
Skill Level	1	1.80	1.80	29.21	.000	0.17
Condition	1	0.17	0.17	2.69	.103	0.02
Error	143	8.80	0.06			
Total	146	71.25				

Table 7-2. ANOVA of the Effects of Condition on Average Posttest Score by Skill Level

Source	df	SS	MS	F	p	η^2
Low Skill						
Condition	1	0.41	0.41	5.59	0.021	0.08
Error	62	4.51	0.07			
Total	64	22.19				
High Skill						
Condition	1	0.00	0.00	0.01	0.945	0.00
Error	80	4.06	0.05			
Total	82	49.06				

Figure 7-3. Means for Average Posttest Score as a Function of Condition and Student Skill Level
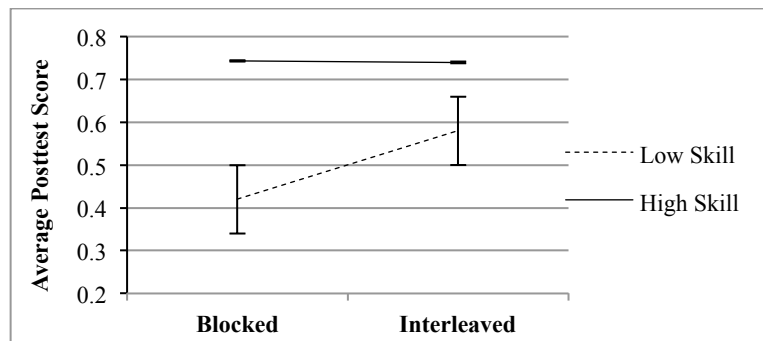
Note. Standard Error for High Skill students is not visible at this scale.

Split file analyses revealed that the effects of interleaving were more impressive when low skill students were considered in isolation. Within 64 low skill students, condition had a significant multivariate effect, Pillai's Trace = 0.12, F(2, 61) = 4.20, p < .05. Univariate analyses revealed that condition had a significant effect on posttest hint usage, F(1, 62) = 5.38, p < .05, η^2 = 0.08, Hedge's g = -0.59, with students in the interleaved condition using less hints on average (M = 0.64, SD = 0.70, n = 39) than those in the blocked condition (M = 1.04, SD = 0.62, n = 25). Condition did not significantly affect posttest attempts, F(1, 62) = 0.08, p > .05, with those in the interleaved condition (M = 2.12, SD = 1.42, n = 39) and those in the blocked condition (M = 2.04, SD = 0.58, n = 25) using a similar amount of attempts.

Within high skill students, condition no longer had a significant multivariate effect, Pillai's Trace = 0.02, F(2, 79) = 0.84, p > 0.05. Summaries of the effects of condition on the dependent variables for both skill levels are presented in Table 7-4. Figure 7-4 depicts the interaction of condition and student skill level observed in average posttest hint usage.

Table 7-3. Univariate Summaries of the Effects of Condition on Dependent Variables

Source	df	Ave. Posttest Hints					Ave. Posttest Attempts				
		SS	MS	F	p	η^2	SS	MS	F	p	η^2
Skill Level	1	18.24	18.24	79.06	.000	0.35	15.13	15.13	21.07	.000	0.13
Condition	1	1.44	1.44	6.24	.014	0.03	0.07	0.07	0.10	.749	0.00
Error	143	32.98	0.23				102.67	0.72			
Total	146	75.75					552.06				

Table 7-4. ANOVA of the Effects of Condition on Dependent Variables by Skill Level

Source	df	Ave. Posttest Hints					Ave. Posttest Attempts				
		SS	MS	F	p	η^2	SS	MS	F	p	η^2
Low Skill											
Condition	1	2.43	2.43	5.38	.024	0.08	0.10	0.10	0.08	.785	0.00
Error	62	27.93	0.45				84.32	1.36			
Total	64	71.00					363.94				
High Skill											
Condition	1	0.06	0.06	1.09	.299	0.01	0.01	0.01	0.03	.865	0.00
Error	80	4.01	0.05				18.31	0.23			
Total	82	4.75					188.13				

Figure 7-4. Means for Average Posttest Hint Usage as a Function of Condition and Student Skill Level
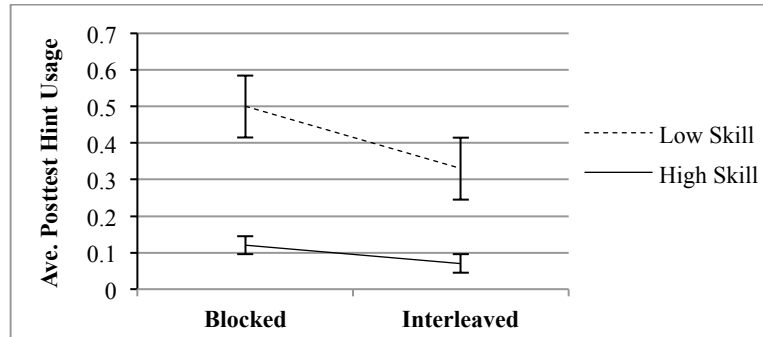
DISCUSSION

The findings herein highlight the promising effects of interleaving skill content within brief mathematics homework assignments in the context of adaptive tutoring systems. Despite failing to achieve an effect size as large as that observed by Rohrer & Taylor (2007) (Cohen's d = 1.34), we observed trends toward significance aligning with past work, serving as further evidence that interleaving skill content enhances learning gains as measured at a delayed posttest. This study also expanded upon interleaving literature to examine how these learning gains differ across student skill level. Further, the findings of the present study extended beyond binary measures of correctness to consider students' differential use of hints and attempts within an informal posttest setting. While this approach was somewhat novel, adaptive tutoring systems allow for the comparison of a variety of rich features within the learning experience that may provide deeper insight than accuracy alone. The observation of significantly different hint usage across conditions suggested that the consideration of feedback utilization, perhaps through a partial credit metric, may offer a more robust explanation for differential learning gains in future research.

The findings observed for low skill students were especially impressive and could prove groundbreaking for future design of adaptive tutoring content. Systems like ASSISTments already

provide educational resources in a manner that has been shown to produce significantly greater learning gains than those found using traditional classroom practices (Mendicino, Razzaq, & Heffernan, 2009). This study suggests that learning outcomes can be further enhanced simply by adding support for a dynamic approach to content delivery through interleaving.

A major limitation of this study was the loss of a large portion of the original sample due to the failure of a participating teacher to assign the posttest to her students. It is possible that a larger sample would better reveal subtleties in the interaction between condition and student skill level. The sample distribution was also suboptimal, with random assignment resulting in more students in the interleaved condition than in the blocked condition. Further, analyses may have been weakened by the discretization of students as generally 'high' or 'low' skill. Departing from the use of a median split should be examined in future work.

Future iterations of this work should incorporate a pretest assignment and use novel skill content rather than skills intended for review. Future work should also examine variables pertaining to student performance within the *practice session* (i.e., average problem time, hint usage, and attempt count) to investigate Bjork's theory of desirable difficulties (Bjork, 1994). Additionally, future research should investigate more robust measures of learning, including extended retention rates following interleaved assignments and the effects on far transfer application.

CONTRIBUTION
While many studies have examined the effect of interleaving, we offer a significant contribution to the field of artificial intelligence in education in that our work replicates the effect of interleaving within a brief homework assignment delivered using an adaptive tutoring system. Emphasized significance was observed for low skill students. Further, the use of homework assignments as both intervention and posttest resulted in the observation that rich features common to adaptive tutoring systems may allow researchers to pinpoint effects in variables other than correctness. The ease with which interleaving can be conducted within adaptive tutoring systems offers a low-cost, high-benefit approach to enhancing student learning outcomes.

Chapter 8 – Improving Student Modeling Through Partial Credit and Problem Difficulty

Student modeling within intelligent tutoring systems is a task largely driven by binary models that predict student knowledge or next problem correctness (i.e., Knowledge Tracing (KT)). However, using a binary construct for student assessment often causes researchers to overlook the feedback innate to these platforms. The present study considers a novel method of tabling an algorithmically determined partial credit score and problem difficulty bin for each student's current problem to predict both binary and partial next problem correctness. This study was conducted using log files from ASSISTments, an adaptive mathematics tutor, from the 2012-2013 school year. The dataset consisted of 338,297 problem logs linked to 15,253 unique student identification numbers. Findings suggest that an efficiently tabled model considering partial credit and problem difficulty performs about as well as KT on binary predictions of next problem correctness. This method provides the groundwork for modifying KT in an attempt to optimize student modeling.

INTRODUCTION

Modeling student learning within an intelligent tutoring system can be a daunting task. In order to make predictions about a student's knowledge or their next problem correctness, models must decipher noisy input and isolate only those features that define the probability of knowledge or learning. As such, designers of intelligent tutoring systems have largely relied on Knowledge Tracing (KT), as presented by Corbett & Anderson (1995), to model the probability of student learning at real time within popular systems such as Cognitive Tutor (Koedinger & Corbett, 2006). Other methods, such Performance Factors Analysis, seek to model learning when considering overlapping knowledge components (i.e., skills) and individualized student metrics (Pavlik, Cen, & Koedinger, 2009), offering an alternative to KT in certain circumstances.

Despite the popularity of KT and PFA, the standard models rely on binary input to establish predictions of students' knowledge state or performance, failing to consider continuous metrics that would better individualize the model across students or skills. Expansion in the field educational data mining has since lead to a number of alternative or supplementary learning models. For instance, researchers have attempted to impart individualized prior knowledge nodes for each student (Pardos & Heffernan, 2010), to supplement KT with a flexible metric for item difficulty (Pardos & Heffernan, 2011), to ensemble various methods of binning student performance (i.e., partial credit) with standard KT models (Wang & Heffernan, 2011), and to consider the sequence of a student's actions within the tutor to help predict next problem correctness (Duong, Zhu, Wang, & Heffernan, 2013).

Without modifying KT or PFA directly, adding parameters to student learning models can be a limited approach. Tabling methods that quickly establish maximum likelihood probabilities have previously been used by Wang & Heffernan (Wang & Heffernan, 2011, Wang & Heffernan, 2013) to test and optimize various potential adaptations to KT. Following in this process, the present study uses a tabling method to lay the groundwork for future modifications to KT that will allow for predictions of next problem correctness using the partial credit score and difficulty estimate of the current item. While previous work has shown the benefit of ensembling tabling methods with KT, we hope to use the findings presented herein to modify KT directly, as it has previously been suggested that ensembling can be a rather sensitive approach (Gowda, Baker, Pardos, & Heffernan, 2011).

Perhaps standard learning models rely on binary correctness as measured by a student's first response at each skill opportunity (i.e., a sequence of correct and incorrect responses based on a student's first action within each problem) due to the complexity of accurately and universally defining an algorithm that validates partial credit scores within intelligent tutoring systems. Within the majority of current learning models, a student would be penalized with a score of zero for taking advantage of the tutoring that plays an integral role in these platforms. Yet the primary goal of most intelligent tutoring systems is not solely to assess student knowledge, but to simultaneously promote student learning through adaptive feedback, making binary correctness a stale concept. Students often require multiple attempts to solve a problem or request system feedback for guidance, thus assigning value to the concept of partial credit. Attali and Powers (2010) suggested the benefits of considering partial credit when predicting learning outcomes in adaptive environments, as evidenced by the modification of standardized tests to allow partial credit when predicting GRE scores.

Within ASSISTments, an adaptive mathematics tutor, a naïve model of partial credit scoring was previously established by Wang and Heffernan (2011), termed the "Assistance" Model. This method calculated maximum likelihood probabilities for next problem correctness using a twelve-parameter table built from binning students' hint usage and attempt count. In this manner, the authors used system features to indirectly gauge a partial credit metric that would help predict binary performance.

The present study provides methodological evidence that student modeling can be enhanced through the use of algorithmically derived partial credit scores and a binned metric of problem difficulty. We first use tabling method (a probabilistic approach employing maximum likelihood estimations) that considers the partial credit score of the current problem to predict both binary and partial next problem correctness. We also establish a more complex prediction table that considers both partial credit and problem difficulty. Through this novel concept, we hope to show that students can ultimately gain knowledge from

a problem even if they fail to earn full credit. Our findings argue for the design of a modified KT model that is sensitive to a continuous measure of partial credit rather than binary input, and that isolates a known level of problem difficulty for each question. We seek to answer the following research questions:

1. Does an algorithmically determined partial credit score outperform binary metrics when used to predict next problem correctness?
2. Does a binned metric of current problem difficulty (e.g., Low, Medium, or High difficulty) provide a valid prediction of next problem correctness?
3. Can current problem difficulty supplement partial credit score to outperform similar modeling techniques?

DATASET

The dataset used for this analysis was compiled from problem logs from the ASSISTments platform during the 2012-2013 school year. The original file included roughly 1.5 million rows of problem level data (i.e., each row detailed all logged actions for one problem for one student). For this study, we chose to analyze only the top ten most densely populated knowledge components. Attributes of these skills are further explained in Table 8-1. The dataset examined here has been made publicly available at (Ostrow, 2014a).

In order to properly calculate partial credit, approximately 5,000 rows were removed due to a lack of logged end time, meaning that these problems had never been properly completed. Using the platform's current grading method, which is based on the students' first response, these logs carry binary correctness scores. However, as the problem was ultimately considered incomplete, partial credit could not be determined with certainty and the logs were therefore excluded from analysis. Further, the analysis presented herein reports only on main problems. Scaffolding problems, a feedback style within the ASSISTments platform typically used to break a problem down into steps or to provide worked examples, were excluded from the final dataset. The decision to work with main problems was based in part on the justification made by Pardos & Heffernan (2011) when using a similar dataset from the ASSISTments platform. As scaffolding problems are guided, they offer a less accurate view of skill knowledge and skew performance data within an opportunity based analysis. An analysis of the remaining dataset revealed that only 0.3% of first actions were scaffold requests, further supporting the intuition that the removal of scaffolding data was appropriate.

Due to the time constraints involved in running multiple models with five fold cross-validation (explained further in the Compared Models and Model Testing and Training sections), we chose to restrict the dataset to a maximum of 15 opportunities per student per skill. This reduced the dataset by 46,680 rows, primarily removing students who were excessively struggling and those gaming the system; the majority of students were unaffected by this refinement.

The resulting dataset consisted of 338,297 problem logs representative of 15,253 unique student identification numbers. On average, each student identifier linked to approximately 3.3 skills. Further exploration of this dataset revealed that it was comprised of 7,363 unique problems. A total of 3,787 unique assignments were made by 417 teachers spanning 231 schools. The skill content ranged from grades 6-8 as shown in Table 8-1. The majority of logged problems (over 90%) were completed by students who 'mastered' or finished the full assignment from which the problem originated.

Three types of questions were represented in the dataset. The majority of problems logged, 84.3%, were 'mathematical expressions,' a problem type that accepts any answer that is mathematically equivalent to the correct answer (i.e., answers of 1/2 and 0.5 are both accurate). In contrast, 12.5% of problems logged were 'fill-in,' a problem type that requires the student to input an *exact* string matching the preset correct response (i.e., if 1/2 was the preset answer, 0.5 would be incorrect). The remaining 3.2% of problems logged in the dataset were 'multiple choice,' featuring two or more answers available for selection.

Table 8-1. Skill Details and Distribution in Resulting Dataset

Skill ID	Definition	Grade Level	# Logs (Rows)	% Resulting Dataset
277	Addition and Subtraction of Integers	7	44,731	13.2
311	Equation Solving with Two or Fewer Steps	7	44,005	13.0
280	Addition and Subtraction of Fractions	6	42,550	12.6
276	Multiplication and Division of Positive Decimals	6	37,033	10.9
47	Conversion of Fractions, Decimals, and Percentages	6	32,741	9.7
67	Multiplication of Fractions	6	31,716	9.4
61	Division of Fractions	6	28,809	8.5
278	Addition and Subtraction of Positive Decimals	6	27,301	8.1
310	Order of Operations	8	25,132	7.4
79	Proportions	7	24,279	7.2

Further assessment of students' responses provided insight into their first actions, attempt counts, and hint usage. For 95.5% of logged problems, the student's first action was to make an answer attempt. Using ASSISTments' current scoring scheme, these attempts would receive binary scores of either correct (1) or incorrect (0). Within this subgroup of logged problems, 24% of the problems were marked as incorrect while 76% were marked as correct. This suggests that a partial credit metric could provide benefit for approximately one quarter of attempted questions. Of the remaining logged problems, 4.2% represented first action hint requests, and 0.3% represented first action scaffolding requests.

Given that partial credit scores for the present study are algorithmically derived from an assessment of the student's attempt count and hint usage for each logged problem, these variables were examined thoroughly. Analysis of attempt counts across logged problems revealed a minimum of 0 and a maximum of 496, with a mean of 1.47 and a standard deviation of 2.23. For logs that were marked as incorrect based on first action, mean attempts rose to 2.70 with a standard deviation of 3.99. Within the full dataset, students made a total of 496,533 attempts.

Hint counts were also analyzed across logged problems and compared to the total number of hints available for each problem. Each problem had at least one hint, usually serving as the bottom out hint (i.e., it provided the answer). The average number of hints available per problem was 3.38, with a standard deviation of 0.88. The majority of problems had three hints (38.9%) or four hints (33.4%), with the maximum number of hints available in any problem topping off at seven. Across all logged problems, a total of 1,090,225 hints were available. Of the available hints, students only used a total of 167,371, or roughly 15.4%. The average number of hints used was 0.49 with a standard deviation of 1.20. For problem logs in which students answered incorrectly on their first attempt, 55.8% of available hints were utilized. Information particular to the bottom out hint showed that within problems initially answered incorrectly, only 14.5% of students proceeded to the bottom out hint. Thus, when struggling, the majority of students used the adaptive feedback inherent to the tutoring system in an appropriate manner. This provides further evidence for consideration of valid partial credit metrics.

Figure 8-1 provides a screenshot of a typical problem within the ASSISTments tutor. Specifically, this problem is a representation of the second most densely populated skill in the 2012-2013 ASSISTments log file: "Equation Solving with Two or Fewer Steps." This skill is exemplified, rather than highlighting the top skill, "Addition and Subtraction of Integers," as the problem provides a more robust example of the system's tutoring feedback. As shown in Figure 8-1, the student is presented with the equation and asked to solve for the missing variable. He or she can make an attempt to solve the problem, or may ask for the first of three hints. The hints increase in specificity, in an attempt to guide the student without providing excess assistance. The first hint shown in Figure 8-1 provides a worked example of a similar problem solving for the missing variable, x. If the student is unable to proceed using only the worked example, he or she can request the second and third hints as needed. The third hint in Figure 8-1 is the bottom out hint; it provides the correct answer ("-24") in an attempt to keep the student

from getting stuck in the assignment, as it is not possible to skip problems and return at a later point as one can with traditional bookwork.

Figure 8-1. An Example Problem Featuring Three Hints for the Skill "Equation Solving with Two or Fewer Steps"

COMPARED MODELS
The following subsections explain the design and brief history (when appropriate) of the five models compared in the current study. All five models are primarily designed to predict binary next problem correctness. For permitting models, we present predictions of partial credit next problem correctness using continuous probabilities for additional consideration.

Partial Credit Predicting Next Problem Correctness
A naïve partial credit algorithm was derived by the ASSISTments design team in hopes of providing the system with partial credit scoring capabilities based on students' attempt count and feedback usage. Scores were determined subjectively based on teacher input and a conceptual understanding of how students typically behave within the tutoring platform. For this study, the algorithm was altered slightly to consider multiple problem types and to account for the students' first action. For instance, if a

student asked for tutoring feedback without making an attempt to solve the problem, we felt that a larger penalty was merited.

The resulting algorithm used to define partial credit scores is depicted in Figure 8-2. Rather than establishing a deduction method on a per hint or per attempt basis as shown in previous work (Wang & Heffernan, 2011), the algorithm presented in Figure 8-2 places each logged problem into one of five partial credit bins (0, .03, 0.6, 0.7, 0.8, 1.0) by considering the logged data pertaining to first response type (attempt = 0, hint request = 1, scaffold request = 2), attempt count, and hint count.

For example, if a student makes only one attempt and is correct without requiring feedback, they earn full credit (a score of 1). This is similar to the notion of binary correctness on first response that ASSISTments currently employs. However, in the current method, all other first actions equate to an incorrect answer (i.e., requesting a hint or scaffold, or making a first attempt that is incorrect would both earn the student a score of 0).

As shown in Figure 8-2, after ruling out a 'correct' response, the partial credit algorithm considers whether the student requested a scaffold before even making an attempt. This behavior would suggest that either the student was not actually trying to answer the problem, or that he or she was struggling conceptually. Thus, rather than earning no credit, the student is only discounted to a score of 0.6.

Regardless of the student's first action, if he or she uses less than three attempts and does not request any hints, they earn slightly more, with a score of 0.8. The next bin is marked by students who have three or fewer attempts and have not used a hint, or those who have asked for only one hint and were not provided the answer (i.e., if a student's first action is to request a hint that is not the bottom out hint, they would fall into this bin). These students earn a score of 0.7. If the student can solve the problem within 5 attempts without seeing the bottom out hint, or if he or she uses multiple hints without ultimately reaching the bottom out hint, their partial credit score is 0.3. Finally, for students who use five or more attempts, or for those that see the answer, the problem is marked incorrect (a score of 0).

For multiple-choice questions the algorithm reverts to binary correctness because this type of problem does not usually provide feedback and guessing can be far more prevalent and consequential. Thus, if a student fails to get the correct answer on their first attempt, he or she receives a score of 0. This method was employed to keep the problem type from gaining an unfair advantage within the dataset. For instance, using the algorithm applied to other problem types, a student guessing through a multiple-choice problem with only four responses would still receive a score of 0.3.

Figure 8-2. Algorithm Used to Determine Partial Credit Score Based on First Response, Attempt Count, and Hint Usage

```
IF type = algebra OR type = fill_in
    IF attempt = 1 AND correct = 1 AND hint_count = 0
        THEN 1
    ELSIF first_action = 2
        THEN .6
    ELSIF attempt < 3 AND hint_count = 0
        THEN .8
    ELSIF (attempt <= 3 AND hint=0)
        OR (hint_count = 1 AND bottom_hint != 1)
        THEN .7
    ELSIF (attempt < 5 AND bottom_hint != 1)
        OR (hint_count > 1 AND bottom_hint != 1)
        THEN .3
    ELSE 0

IF type = multiple_choice
    IF correct = 1
        THEN 1
    ELSE 0
```

The full algorithm was run across the dataset and partial credit scores were obtained for each logged problem. These partial credit scores were then used to define a pivot table to predict averages for both binary and partial next problem correctness, using maximum likelihood estimation. Results are presented in Table 8-2. For all parameter Tables, the number of logged problems falling into respective bins is depicted by sample size, n. The distribution of the data suggests that slight improvements could be made to the partial credit algorithm as few students fell into the 0.6 bin. Of all available 'next problem' data, only 14.7% of logs had partial credit values between 0 and 1. Thus, 85.3% of students would be insured by the platform's current method of binary correctness. This suggests that any significant finding among the models considered in the present study would be quite intriguing, as only a small portion of the sample is actually receiving the 'partial credit' treatment.

It should be noted that a potential problem inherent to this tabling method (apparent in all tabled models in the present study) is the inability to predict correctness on a student's first opportunity within a skill, as there is no preceding problem data. This essentially causes the loss of 49,990 rows of data representing first problem predictions. Thus, sample sizes in Tables 8-2, 8-3, and 8-4 total 288,307 logs rather than 338,297.

Table 8-2. Parameters for Predicting Binary and Partial Next Problem Correctness from Current Problem Partial Credit

Partial Credit	n	Binary	Partial
0	45,735	0.5062	0.5634
0.3	6,471	0.5902	0.7438
0.6	940	0.3660	0.7948
0.7	12,077	0.6921	0.8396
0.8	22,797	0.7085	0.8668
1	200,287	0.8050	0.8785

Problem Difficulty Predicting Next Problem Correctness
A continuous metric of problem difficulty was calculated by retrieving data from all problems logged in the platform that were created before August 2012 (i.e., prior to the first timestamp in the modeling dataset). For each unique problem, all existing logs were averaged and a percentage of correct responses were determined. The resulting value offers an inverse metric of the problem's difficulty level. For instance, a problem on which students averaged 80% on all previous opportunities would not be considered very difficult. This metric was then binned into Low, Medium, and High difficulties by defining Medium difficulty as scores falling within +/- 0.5 standard deviations from the mean. Considering the inverse nature of the metric, High difficulty problems therefore had continuous values *below* this cut off, and Low difficulty problems had continuous values *above* this cutoff.

The bins for current problem difficulty were used in a maximum likelihood probability table to predict averages for both binary and partial scores for next problem correctness. Resulting parameters are presented in Table 8-3.

Table 8-3. Parameters Predicting Binary and Partial Next Problem Correctness from Current Problem Difficulty

Difficulty	n	Binary	Partial
Low	91,712	0.7764	0.8465
Medium	107,901	0.7452	0.8297
High	88,694	0.6928	0.7895

Partial Credit and Problem Difficulty Predicting Next Problem Correctness
Based on the definitions of partial credit and problem difficulty defined in the singular models above, our goal was to create a novel model that used a tabling approach to consider partial credit together with problem difficulty to make predictions about next problem correctness. For each logged problem, partial

credit score and problem difficulty were referenced to determine parameters for both binary and partial credit next problem correctness. Resulting probabilities are presented in Table 8-4.

Table 8-4. Parameters Predicting Next Problem Correctness from Partial Credit and Problem Difficulty

| Partial Credit | High | | | Medium | | | Low | | |
	n	Binary	Partial	n	Binary	Partial	n	Binary	Partial
0	8,357	0.5130	0.5621	16,307	0.5027	0.5622	21,071	0.5062	0.5650
0.3	1,107	0.6035	0.7401	2,332	0.6017	0.7548	3,032	0.5766	0.7367
0.6	29	0.5902	0.8508	236	0.3388	0.7897	675	0.3661	0.7943
0.7	2,829	0.6971	0.8288	4,888	0.6987	0.8463	4,360	0.6816	0.8391
0.8	5,094	0.7770	0.8753	8,342	0.7354	0.8712	9,361	0.6473	0.8581
1	74,296	0.8116	0.8787	75,796	0.8072	0.8841	50,195	0.7921	0.8697

Knowledge Tracing

Knowledge Tracing (KT) is perhaps the most common method for modeling student performance. The standard KT model (Corbett & Anderson, 1995) has successfully proven itself as the basis for modeling student knowledge within intelligent tutoring systems (Koedinger & Corbett, 2006) and thereby serves as a stable comparison for new work.

As shown in Figure 8-3, the standard model of KT is a Bayesian Network comprised of four learned parameters. Two parameters represent student knowledge (*prior knowledge* and *learn rate*) and two parameters represent student performance (*guess rate* and *slip rate*). The standard KT model is binary in that skills can only be in a 'learned' or 'unlearned' state, and questions can only be 'correct' or 'incorrect.' The model is updated with each skill opportunity based on the student's performance by using the following equation as defined by Corbett & Anderson (1995):

$$p(L_n) = p(L_{n-1}|evidence) + (1 - p(L_{n-1}|evidence)) * p(T)$$

Forgetting does not factor into the standard KT model when observing individual skills, as *guess* and *slip* parameters are thought to account for incorrect answers within the students' sequence of opportunities. For further information regarding the details of KT, refer to (Corbett & Anderson, 1995). For this study, KT analysis was performed using the Bayes Net Toolbox (BNT), a popular open-source code for fitting directed graphical models within MATLAB (Murphy, 2001).

Figure 8-3. The Standard Knowledge Tracing Model with All Learned Parameters and Nodes Explained

Standard KT Model

$P_{(L_0)}$ = Prior Knowledge $P_{(T)}$ = Learn Rate
$P_{(G)}$ = Guess Rate $P_{(S)}$ = Slip Rate

Knowledge nodes (K_n) & Question nodes (Q_n)
have binary states (0 or 1)
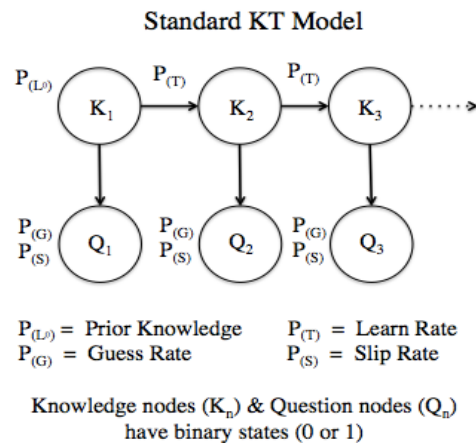
Performance Factors Analysis

Performance Factors Analysis (PFA) was proposed as an alternative to KT by Pavlik, Cen, and Koedinger (2009). The method can model problems with multiple skills and has been shown to accurately model and select practice within adaptive systems. PFA was derived from Learning Factors Analysis (LFA), an approach that considers a parameter for student ability, a parameter for the skill's difficulty, and a learning rate for each skill. While PFA still considers skill difficulty, β, the model improves upon LFA by considering the frequency of both correct and incorrect answers in a student's response pattern, rather than simply assessing the frequency of skill practice. Thus, PFA predictions are updated with each skill opportunity based on a cumulative history of the student's successes (weighted by γ) and failures (weighted by ρ), as depicted in the following equation defined by Pavlik, Cen, and Koedinger (2009):

$$m(i, j \in KCs, s, f) = \sum_{j \in KCs} (\beta_j + \gamma_j s_{i,j} + \rho_j f_{i,j})$$

The log-likelihood (*m*) attained through this equation can then be passed through an exponential function to find the probability that the student will get the item correct. This model suggests that learning is defined by more than just skill practice, and that performance is strongly tied to skill acquisition. For this study, PFA was performed using unpublished code within MATLAB. With properly formatted data, the analysis can also be performed using logistic regression in common statistical packages like IBM's SPSS.

MODEL TRAINING AND TESTING

Five-fold cross validation was used to train and test each model. In order to perform five-fold cross validation within our tabled models, the dataset was divided using a modulo operation on each student's unique identification number. Thus, for every student in the file, student id mod 5 was called, returning a remainder falling into bins from 0 to 4, thereby assigning students to folds. The distribution of the resulting folds was roughly equivalent, as shown in Table 8-5. With 15,253 unique student identification numbers in the dataset, the largest fold had 3,082 student ids and the smallest fold had 2,996 student ids, leaving a range of 86 and a standard deviation of 33.7.

Within each iteration of the cross-validation process, the model was trained on approximately 80% of the data and tested on the 20% that had been held out. Thus, when trained on folds 1, 2, 3, and 4 (80% of the data) the model would impart predictions on fold 0 (the held out 20%). In this manner, for each tabling method described in the previous section, table parameters were learned using four training folds and predictions were made on the held out fold. The process was repeated for all folds, thus resulting in five probability tables for each prediction type (i.e., five 'training' tables for partial credit predicting binary next problem correctness). Using an extensive formula in Microsoft Excel, the predicted averages were then applied back to each logged problem respective of test fold. For predictions of binary next problem correctness, rather than arbitrarily selecting a cutoff point for classifying binary correctness (e.g., simply using values greater than 0.5 to convey '1'), we instead subtracted the prediction directly from the actual binary result. Thus, when predicting next problem correctness using partial credit alone, if the next problem is actually correct using binary standard, the resulting residual is calculated as: 1.0000 - 0.7085 = 0.2915. In this manner, residuals were calculated for each log entry in each test fold that contained data for next problem correctness.

A similar method of five fold cross-validation was coded into the KT and PFA analyses within MATLAB. Without modification, KT and PFA are not intended to accurately predict partial credit next question correctness, and as such we have saved these analyses for future work.

Table 8-5. Distribution of Data Across Five Folds

Fold	Unique Students	# Logs	% Dataset
0	3074	67,715	20.0
1	3046	68,081	20.1
2	3082	68,739	20.3
3	3054	67,996	20.1
4	2996	65,766	19.4

To compare our tabled models with KT and PFA, slight modifications were made to the standard modeling procedures. Unlike tabling, these models carry the benefit of being able to predict performance on a student's first opportunity within a skill. Based on a 'prior knowledge' parameter, KT is able to predict the student's initial knowledge state, K_1, and therefore their performance on the first question, Q_1. Similarly, the equation for PFA defaults a prediction of the skill's difficulty parameter, β, as the student's initial state. These values essentially define a baseline for the student's knowledge, prior to any practice. Thus, in order to provide a fair comparison to tabled models, these first opportunity predictions were removed by shifting predictions to align with our 'next problem' analysis. Within KT, all subsequent skill opportunities were predicted using Expectation Maximization, a standard method for parameter learning within KT. The model was supplied the following initial parameters as 'ground truths' to begin the hill climbing process: *prior knowledge* = 0.30, *learn rate* = 0.20, *forget rate* = 0.00, *guess rate* = 0.20, and *slip rate* = 0.08. Within PFA, all subsequent skill opportunities were predicted by updating the equation presented in the previous section. These modifications resulted in the same number of data points for each model, providing grounds for fair comparison of the models.

Further, as noted briefly in the Dataset section, all models were restricted to 15 predicted opportunities per student per skill. This method was chosen largely to reduce the computation time required to fit KT using five-fold cross validation on such an extensive dataset. By capping the opportunity count, analysis time was reduced to approximately 20 hours. Other models were far less time intensive, all taking under three hours to arrive at predictions. Setting this restriction also served to reduce potential skewing in student level analyses by removing outliers with extensive opportunity counts.

RESULTS

All models were compared using the fit statistics of RMSE, R^2, AUC, and model accuracy. As the tabled models were not restricted to binary input, fit statistics were also found for consideration of modeling partial credit next problem correctness.

For each model, these statistics were found at the problem log level, the skill level, and the student level where merited. These statistics were then averaged across the level of analysis, resulting in the findings presented in Table 8-6, Table 8-7, and Table 8-8, respectively. Thus, at the problem log level, fit statistics were determined overall for the 288,307 predictions, without consideration of student or skill before being averaged across all problems. At the skill level, ten sets of fit statistics were determined (one set for each skill), which were then averaged across skills. At the student level, 15,253 sets of fit statistics were determined (one set for each student), which were then averaged across students. The latter two procedures were intended to properly weight skill and students based on their contribution to the dataset, thereby improving measures of model fit.

Student level statistics of RMSE were calculated based on all predictions. However, it should be noted that measures of R^2, AUC, and model accuracy could not be calculated for students with less than three skill opportunities. This discrepancy should affect all models equally, and thus we provide these measures for comparison in Table 8-8 with the caveat that they should not be directly compared to measures of student level RMSE.

Table 8-6. Problem Level Average RMSE, R^2, AUC, and Accuracy for Models Predicting Next Problem Correctness

Model	Binary NPC				Partial NPC			
	RMSE	R^2	AUC	Accuracy	RMSE	R^2	AUC	Accuracy
Partial Credit + Problem	0.4241	0.0674	0.6365	0.7310	0.3326	0.1062	0.5395	0.7298
Partial Credit	0.4244	0.0660	0.6309	0.7309	0.3327	0.1060	0.5351	0.7298
Problem Difficulty	0.4379	0.0057	0.5464	0.7300	0.3511	0.0043	0.3953	0.7298
Knowledge Tracing	0.4240	0.0680	0.6621	0.7298	--	--	--	--
Performance Factors Analysis	0.4227	0.0738	0.6644	0.7485	--	--	--	--

Table 8-7. Skill Level Average RMSE, R^2, AUC, and Accuracy for Models Predicting Next Problem Correctness

Model	Binary NPC				Partial NPC			
	RMSE	R^2	AUC	Accuracy	RMSE	R^2	AUC	Accuracy
Partial Credit + Problem	0.4224	0.0670	0.6300	0.7414	0.3284	0.1032	0.5130	0.7399
Partial Credit	0.4229	0.0656	0.6290	0.7414	0.3284	0.1031	0.5103	0.7399
Problem Difficulty	0.4364	0.0046	0.5323	0.7402	0.3473	0.0037	0.3560	0.7399
Knowledge Tracing	0.4225	0.0602	0.6500	0.7466	--	--	--	--
Performance Factors Analysis	0.4212	0.0664	0.6506	0.7499	--	--	--	--

Table 8-8. Student Level Average RMSE, R^2, AUC, and Accuracy for Models Predicting Next Problem Correctness

Model	Binary NPC				Partial NPC			
	RMSE*	R^2	AUC	Accuracy	RMSE*	R^2	AUC	Accuracy
Partial Credit + Problem	0.3864	0.1027	0.5431	0.7684	0.2702	0.1108	0.3593	0.7674
Partial Credit	0.3866	0.0994	0.5392	0.7683	0.2701	0.1057	0.3619	0.7674
Problem Difficulty	0.4064	0.0829	0.5219	0.7676	0.2941	0.0851	0.3145	0.7674
Knowledge Tracing	0.3897	0.1057	0.4425	0.7729	--	--	--	--
Performance Factors Analysis	0.3882	0.0970	0.5003	0.7754	--	--	--	--

*R^2, AUC, and Accuracy are reported with less data than RMSE due to the nature of student level data

DISCUSSION

The fit statistics for both the problem log and skill level generalizations paint very similar pictures of the relative success of our tabling method. The combined Partial Credit and Problem Difficulty model performs about as well as KT at both levels of analysis. At these levels, PFA appears to be the 'best' model for predicting binary next problem correctness, showing the lowest RMSE and highest AUC and model accuracy. However, we feel that a simple tabling method that can be performed with extreme efficiency yet still meets the standards of KT is well worth discussion.

Our first research question, "Does an algorithmically determined partial credit score outperform binary metrics when used to predict next problem correctness?" was answered with mixed results for binary predictions. Considering problem log level analysis, while KT and PFA attained fit statistics relative to those accepted in the field, our tabling method for partial credit considered alone only slightly underperformed these standards (RMSE = 0.4244, R^2 = 0.0660, AUC = 0.6309, Accuracy = 0.7309). However, when considering student level analysis, our partial credit tabling method outperformed both KT and PFA in terms of RMSE and AUC. To confirm that these findings were significantly different, we used a two-tailed paired samples t-test for RMSE comparison at both the student level and skill level. RMSEs obtained using our tabling method with partial credit alone were significantly different from those found using KT at the student level, t = 5.65, p < .001, but were not significantly different at the skill level, t = -1.65, p = 0.133. Thus, it is difficult to tell if this finding is truly significant.

Our second research question, "Does a binned metric of current problem difficulty (e.g., Low, Medium, or High difficulty) provide a valid prediction of next problem correctness?" was answered by assessing the "Problem Difficulty" model. When taken alone, problem difficulty is not very helpful in predicting next problem correctness. This was the worst performing model across all granularities of analysis. A paired samples t-test was again used to compare student level and skill level RMSEs to those observed using the KT model. RMSEs obtained using our tabling method for Problem Difficulty were significantly worse than those found using KT at the student level, t = -41.27, p < .001, as well as those found using KT at the skill level, t = -9.93, p < .001. Of the tabled models, this model was also the lowest performing model when considering predictions of partial next credit correctness, drastically underperforming models that considered current problem partial credit score. Thus, we argue that problem difficulty alone is a poor metric for modeling student performance.

Our final research question, "Can current problem difficulty supplement partial credit score to outperform similar modeling techniques?" was answered by assessing the fit statistics for the combined "Partial Credit + Problem Difficulty" model. At the student level, this model outperformed both KT and PFA on predictions of binary next problem correctness as measured by RMSE (0.3864) and AUC (0.5431). This finding was significant using a two-tailed paired samples t-test comparing student level RMSEs, t = 6.50, p < .001, but was not significant when considering skill level RMSEs, t = -1.34, p = 0.214. Despite the low performance of the Problem Difficulty model, this combined model consistently outperformed partial credit when modeled alone, suggesting possible mediation effects. Using a paired t-test comparison, this difference was significant at the student level, t = -4.55, p < .001, but was not significantly reliable at the skill level, t = -1.03, p = .310. As such, it is difficult to quantify the potentially negative impact of considering problem difficulty when using partial credit to model next problem correctness.

Model fit indices for the prediction of partial credit scores for next problem correctness are provided for further consideration, but do not specifically link to our research questions. Drastic improvements in model fit suggest that intelligent tutoring systems should incorporate partial credit scoring as it has the potential to enhance the precision of student modeling. In the current study, these findings cannot be compared to standard KT and PFA models that utilize binary input and essentially predict binary performance. Future research will incorporate modifying these models to predict continuous partial credit metrics, thus allowing for further comparison.

CONTRIBUTION

The results from the present study suggest that considering partial credit for each skill opportunity can enhance the accuracy of student modeling. While the concept of using a tabling method to establish partial credit metrics that predict binary correctness is not novel (Wang & Heffernan, 2011), tabling a model based on algorithmically determined partial credit is, to the best of our awareness, a unique approach. This method was shown to perform about as well as KT when predicting binary next problem correctness. We feel that this finding still provides a significant contribution to the field, as KT is far more computationally expensive. Our KT analysis took approximately 20 hours to run, while all tabling methods were conducted by hand in less than three hours. While this is impressive in and of itself, the second author was then able to implement the tabling method presented here within the ASSISTments test database, arriving at a replication of our predictions in less than two minutes. If automated in such a manner, our Partial Credit + Problem Difficulty model could predict next problem performance on par with KT in approximately one 600th of the time. This increase in efficiency could prove essential for intelligent tutoring systems that currently incorporate KT models to adaptively control student skill practice.

Further, the partial credit model was novel in its ability to predict partial credit scores for next problem correctness, thereby enhancing model fit even further. In future research, we hope to modify the standard KT and PFA models to allow for the prediction of continuous variables for comparison. We also anticipate directly comparing our partial credit model to the "Assistance" Model established in previous research (Wang & Heffernan, 2011). The "Assistance" Model cited a clear cut, albeit subjective, method

for the provision of partial credit scores. As the tabling technique employed made predictions on a continuous scale rather than by binning partial credit as we have shown in the present study, comparison was not presently possible without ensembling our findings with standard KT measures (Wang & Heffernan, 2011). However, as alternating ensembling techniques lead to inconsistent results (Gowda, Baker, Pardos & Heffernan, 2011), we argue for direct modifications within KT that will allow the model to learn partial credit scores at each opportunity and to gauge a student's knowledge state on a continuum. A similar model was previously suggested by Pardos & Heffernan (2012), but to our knowledge has never been implemented. Thus, the present study lays the groundwork for future research in modifying KT.

The assessment of models considering problem difficulty also provides a contribution to the modeling literature. It seems intuitive that problem difficulty should influence a students' ability to answer the current problem correctly, and that it likely influences their knowledge state and next problem correctness. The findings here suggest that problem difficulty alone, when binned into generic groups of Low, Medium, and High difficulty, does not provide accurate models of next problem correctness. However, problem difficulty appeared to enhance modeling when coupled with partial credit in comparison to partial credit modeled alone, although this difference was not shown to be significant. Still, we believe that some measure of problem difficulty is important to consider when modeling student learning. Future research should investigate using a continuous metric or designing an alternative binning approach for this feature. Future work should also consider devising an approach to remedy the issue of being unable to predict a student's first opportunity within a skill when using tabled models. Possible solutions include per student estimates of prior knowledge based on performance on other skills within the tutor, or simply implementing problem difficulty as a measure of likelihood for accuracy.

Despite the impressive performance of our partial credit model, we retain skepticism in regards to the subjective nature of our partial credit algorithm. As multiple arbitrary partial credit models have now been designed to assess log data from the ASSISTments platform (Wang & Heffernan, 2011), we argue for the design of a data driven algorithm that considers and compares a myriad of logged features. Future work will examine a grid search of possible hint and attempt penalties to examine the sensitivity of the approach described herein. The data files of intelligent tutoring system are rich with information pertaining to students' actions, including the time required for first response, their sequence of actions within each problem, and the specific misconceptions that are driving incorrect responses. These features may provide critical information for the scoring of partial credit. When considering the approach used in the present study, using an algorithm to establish partial credit scores prior to tabling provides the leeway for tabled models to consider these additional features. Future research could easily replicate similar models, combining partial credit with novel features for additional exploration of the observed effect.

Chapter 9 – Optimizing Partial Credit Algorithms to Predict Student Performance

As adaptive tutoring systems grow increasingly popular for the completion of classwork and homework, it is crucial to assess the manner in which students are scored within these platforms. The majority of systems, including ASSISTments, return the binary correctness of a student's first attempt at solving each problem. Yet for many teachers, partial credit is a valuable practice when common wrong answers, especially in the presence of effort, deserve acknowledgement. We present a grid search to analyze 441 partial credit models within ASSISTments in an attempt to optimize per unit penalization weights for hints and attempts. For each model, algorithmically determined partial credit scores are used to bin problem performance within a maximum likelihood table, using partial credit to predict binary correctness on the next question. An optimal range for penalization is discussed and limitations are considered.

INTRODUCTION

Adaptive tutoring systems provide rich feedback and an interactive learning environment in which students can excel, while teachers can maintain data-driven classrooms, using the systems as powerful assessment tools. Simultaneously, these platforms have opened the door for researchers to conduct minimally invasive educational research at scale while offering new opportunities for student modeling. Still, they are commonly restricted to measuring performance through binary correctness at the problem level. Arguably the most popular form of student modeling within computerized learning environments, Knowledge Tracing, is rooted in the binary correctness of each opportunity or problem a student experiences within a given skill (Corbett & Anderson, 1995). Knowledge Tracing (KT) drives the mastery-learning component of renowned tutoring systems including the Cognitive Tutor series, allowing for real time predictions of student knowledge, skill mastery, or next problem correctness (Koedinger & Corbett, 2006). Similar modeling methods consider variables that extend beyond correctness but rarely escape the binary nature of the construct, including Item Response Theory (Drasgow & Hulin, 1990) and Performance Factors Analysis (Pavlik, Cen, & Koedinger, 2009). By restricting input to a single binary metric across questions, these modeling techniques fail to consider a continuous metric that is commonplace for many teachers: partial credit.

Partial credit scoring used within adaptive tutoring systems could provide more individualized prediction and thus establish models with better fit. It is likely that binary correctness has remained the default for learning models due to the inherent difficulty of defining a universal algorithm to generalize partial credit scoring across platforms. Additional onus may fall on users' familiarity with current system protocol; students tend to avoid using system feedback regardless of the benefits it may provide as requesting feedback results in score penalization. However, the primary goal of these platforms is to promote student learning rather than simply offering assessment, and thus binary correctness becomes a weakness.

The present study considers data from ASSISTments, an online adaptive tutoring system that provides assistance and assessment to over 50,000 users around the world as a free service of Worcester Polytechnic Institute. Researchers have previously used ASSISTments data to modify student-modeling techniques in a variety of ways including through student level individualization (Pardos & Heffernan, 2010), item level individualization (Pardos & Heffernan, 2011), and the sequence of student response attempts (Duong, Zhu, Wang, & Heffernan, 2013). Previous work has also shown that naïve algorithms and maximum likelihood tabling methods that consider hints and attempts to predict next problem correctness can be successful in establishing partial credit models meant to supplement KT (Wang & Heffernan, 2011; Wang & Heffernan, 2013). More recently, algorithmically derived partial credit scoring has resulted in stand-alone tabled models relying on data from only the most recent question, showing goodness of fit measures on par with KT at lower processing costs (Ostrow, Donnelly, Adjei, Heffernan, 2015). However, we hypothesize that some conceptualizations of partial credit may lead to better predictive models than others. Rather than subjectively defining tables or algorithms, a data driven approach should be considered. Thus, considering student performance within the ASSISTments platform, the current study employs a grid search on per unit penalizations of hints and attempts to ask:

1. Based on penalties for hints and attempts dealt per unit, is it possible to algorithmically define partial credit scoring that optimizes the prediction of next problem correctness?
2. Does the optimal model of partial credit differ across different granularities of dataset analysis?

Establishing an optimal partial credit metric within ASSISTments would allow teachers to gain more accurate assessment of student knowledge and learning, while allowing students to alter their approach to system usage to take advantage of adaptive feedback. The optimization of partial credit scoring would also enhance student-modeling techniques and offer a new approach to answering complex questions within the domain of educational data mining.

DATA

The ASSISTments dataset used for the present study is comprised solely of assignments known as Skill Builders. This type of problem set requires students to correctly answer three consecutive questions to complete their assignment. Questions are randomly pulled from a large pool of skill content and are typically presented with tutoring feedback, most commonly in the form of hints. The dataset has been de-identified and is available at (Ostrow, 2014b) for further investigation.

The dataset used in the present study is a compilation of Skill Builder data from the 2012-2013 school year, containing data for 866,862 solved problems. Data recorded includes a student's performance on the problem (i.e., binary correctness, hint count, attempt count), variables that identify the problem itself (i.e., problem type, unique problem identification number) and information pertaining to the assignment housing the problem (i.e., unique identifiers for assignments, skill type, teachers, and schools). This dataset was representative of 120 unique skills and 24,912 unique problems, solved by 20,206 students.

On average, students made 1.53 attempts per problem (SD = 15.08). The minimum number of attempts was 0 (i.e., a student who opened the problem and then left the tutor), while the maximum number of attempts was a daunting 12,246 (i.e., a student who hit enter repeatedly for a prolonged period of time, likely out of frustration or boredom). Students made a total of 1,324,226 attempts across all problems. The majority of problems (74.9%) had just one logged attempt per student (typically correct answers), while 15.1% of problems carried only two logged attempts.

Hint usage among all students averaged 0.61 hints per problem (SD = 1.29). The minimum number of hints used was 0 (i.e., no feedback requested), while the maximum number of hints used was 10. Interestingly, the maximum number of hints available for any particular problem was 7. Thus, a handful of students who logged more than 7 hints were accessing the tutor in multiple browser windows (i.e., cheating). On average there were 3.22 hints available per problem (SD = 0.89). The majority of problems contained 3 hints (44.6%), 4 hints (28.9%) or 2 hints (18.2%). Although there were 2,768,299 hints available across all problems, students only used 529,394 hints, or approximately 19% of available feedback. Bottom out hints, or those providing the problem's solution, were only used on 146,742 (16.9%) of problems.

Additional analysis was performed on the 261,787 problems that students answered incorrectly out of the original 866,862 problems solved. Within this subset of data, students made an average of 2.75 attempts per problem (SD = 27.40). Students also used an average of 2.02 hints (SD = 1.63). This subset of problems had 860,131 total hints available, of which students used 528,644 hints (61.5%).

Hint usage would likely increase if partial credit scoring was implemented within the ASSISTments platform. Binary first attempt scoring has created an environment in many classrooms where students are afraid to use hints although they would benefit from feedback, as they know they will receive no credit. Further, the dataset suggests that once students are marked wrong, they are more likely to jump through all available hints and seek out the answer (56% of incorrect first attempts led to bottom out hinting). This reflects another substantial downfall in the system's current protocol; once the risk has passed, so has the drive to learn. The implementation of partial credit scoring has the potential to alleviate this misuse.

METHODS

The present study presents an extensive grid search of potential per hint and per attempt penalizations. The full dataset described above was used to algorithmically define partial credit scores based on a per unit penalization scale ranging from 0 to 1 in increments of 0.05 for both variables. Thus, for each problem in the dataset, 441 potential partial credit scores were established based on each possible combination of per unit penalization. For instance, in a model in which each attempt earned a penalization of 0.05, and each hint earned a penalization of 0.1, a student who made three attempts and used one hint would receive a penalty of 0.25 ((3 x 0.05) + (1 x 0.1)), effectively scoring 75% on that problem. This process was used to score each problem in the dataset for each possible penalty combination, with a minimum per problem score of 0 set as the floor (students could not receive negative scores). This method was similar to that presented by Wang & Heffernan in the Assistance Model (2011) which established a

tabling method to calculate probabilities of next problem correctness based on combinations of hints and attempts that resulted in twelve possible bins or parameters.

For each of the 441 partial credit models, a maximum likelihood tabling method was employed using five fold cross validation. Within each model, a modulo operation was used on each student's unique identification number to assign students to one of five folds. Note that this method results in folds that all represent approximately 20% of students in the dataset. Maximum likelihood probabilities for next problem correctness were then calculated for each partial credit score within each model. Table 9-1 presents an average of test fold probabilities for the model in which each attempt hint are penalized 0.1. This model is depicted as the penalization structure results in eleven clean and understandable bins for partial credit scores. For instance, a student using two attempts and one hint would be penalized 0.3, thus falling into the score bin of 0.7 (PC Score). Following through with this example, based on 11,174 logged problems that fit this scoring structure, the average of known binary performance on the following problem (performance at time t +1) was 0.599. This value becomes the prediction for next problem correctness.

Table 9-1. Probabilities Averaged Across Test Folds for the Model in Which Penalization Per Hint and Per Attempt is 0.1

PC Score	n	Max. Likelihood NPC
0	149,504	0.467
0.1	422	0.571
0.2	685	0.581
0.3	1,055	0.578
0.4	1,784	0.574
0.5	3,442	0.583
0.6	6,623	0.585
0.7	11,174	0.599
0.8	18,679	0.662
0.9	49,972	0.725
1.0	476,523	0.802

Using the maximum likelihood probabilities for next problem correctness within each test fold as predicted values, residuals were then calculated by subtracting predictions directly from actual next problem binary correctness (i.e., $1 - 0.725 = 0.275$; $0 - 0.571 = -0.571$). This approach was used rather than selecting an arbitrary cutoff point to classify a prediction as correct or incorrect in the binary sense (i.e., values greater than or equal to 0.6 serve as predictions of correctness) because it reduced the potential for researcher bias.

RESULTS

For each model, residuals were used to calculate RMSE, R^2 & AUC at three levels of granularity: problem level, student level, and skill level. Heat maps are presented here only for RMSE, as the other metrics established almost identical maps. Metrics representing greater model fit are depicted using the purple end of the spectrum, while those representing poor fit are represented using the red end of the spectrum. Further, a series of ANOVAs are used to compare each set of models within the same penalization level for attempts and hints. For instance, the 21 models in which attempt penalty was set to 0.2 were compared to all other sets of attempt penalty models to investigate significant differences occurred across penalties. This method was used rather than attempting to compare each model with all other models using paired samples t-tests, as the resulting 194,481 analyses (441^2) would greatly inflate the rate of Type I error without vast and unrealistic correction strategies.

Initial analysis was performed at the problem level; residuals were calculated for each row of data that contained next problem correctness metrics and goodness of fit measures were averaged across the dataset. Each metric followed a similar structure in which low attempt penalties appear to result in better

fitting models, while hint penalty does not appear to be significant. Thus, partial credit scoring algorithms using lower penalties for attempts were better at predicting next problem performance, as depicted in Figure 9-1. The ANOVA results depicted in Table 9-2 suggest that differences in attempt penalty models were significant. Thus, the set of models with per attempt penalties of 0.1 differed significantly from the set of models with per attempt penalties of 0.8. Differences among hint penalty models were not reliably significant. Figure 9-1 also suggests that the current binary scoring protocol used by ASSISTments results in predictive models that are inadequate. First attempt binary correctnes is the equivalent of the model in which per attempt and per hint penalty are both set to 1, or the upper right corner of each heatmap). This model resulted in consistently poor fit metrics, suggesting that modeling techniques such as KT should use continuous or binned partial credit values as input as they enhance next problem prediction ability. It has not yet been investigated how this alteration would change the prediction of other variables commonly predicted through KT, such as latent student knowledge or mastery.

Student level analysis was undertaken using a subset of the original data file. At this granularity, goodness of fit metrics were calculated for each student and averaged across students to obtain final metrics for each of the 441 models. As the ASSISTments system measures completion of a Skill Builder as three consecutive correct answers, a number of high performing students had limited opportunity counts within skills. For students with too few data points, it was not possible to calculate R^2 and AUC given the nature of these metrics. Therefore, student level analysis incorporated 7,429 students from the original dataset, or 651,849 problem logs. Answering our second research question, it appears as though the region of optimal partial credit values observed at the problem level remains consistent at the student level, as shown in Figure 9-2. ANOVA results depicted in Table 9-2 show reliably significant differences across attempt penalty models but not across hint penalty models.

Skill level analysis was also undertaken using a subset of the original data file. One skill did not have enough data based on a low number of users and high mastery within those users, and was excluded from skill level analysis, resulting in a file with 119 skills. At this granularity, goodness of fit metrics were calculated for each skill and averaged across all skills to obtain final metrics for each of the 441 models. Results are depicted in Figure 9-3. The heat map shows that the region of optimal penalization has grown more concise, showing optimal fit among models with low per hint and per attempt penalties (< 0.3). ANOVA results depicted in Table 9-2 suggest reliably significant differences in all metrics across attempt penalty models but not across hint penalty models.

Table 9-2. ANOVA Results for Groups of Attempt and Hint Penalty Models at Each Level of Analysis

Level	Min	Max	Attempt Penalty			Hint Penalty		
			F	p	R^2	F	p	R^2
Problem								
RMSE	.430	.435	302.70	.000	.935	0.95	.519	.043
AUC	.626	.655	295.46	.000	.934	1.14	.304	.052
R^2	.070	.091	304.34	.000	.935	0.95	.525	.043
Student								
RMSE	.424	.429	222.49	.000	.914	1.34	.149	.060
AUC	.578	.593	208.19	.000	.908	1.42	.106	.063
R^2	.096	.110	374.52	.000	.947	0.80	.715	.037
Skill								
RMSE	.423	.429	517.85	.000	.961	0.55	.944	.026
AUC	.624	.647	250.17	.000	.923	0.72	.805	.033
R^2	.073	.090	510.96	.000	.961	0.49	.971	.023

Note. For all models, df = (20, 420).

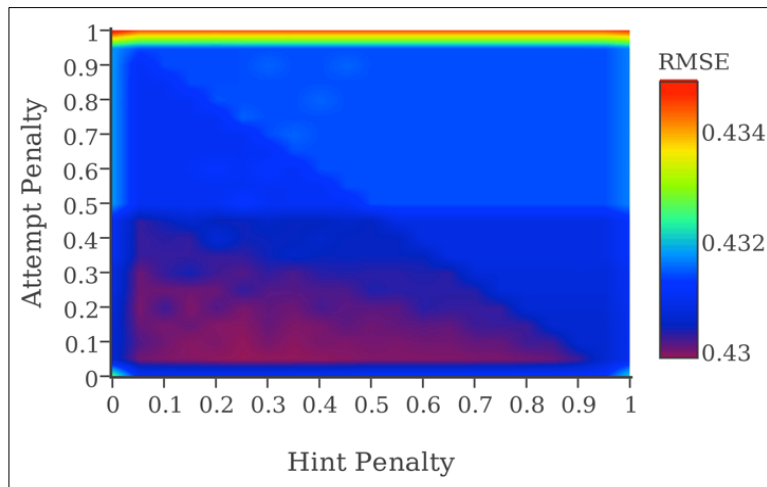**Figure 9-1. Problem Level RMSE**

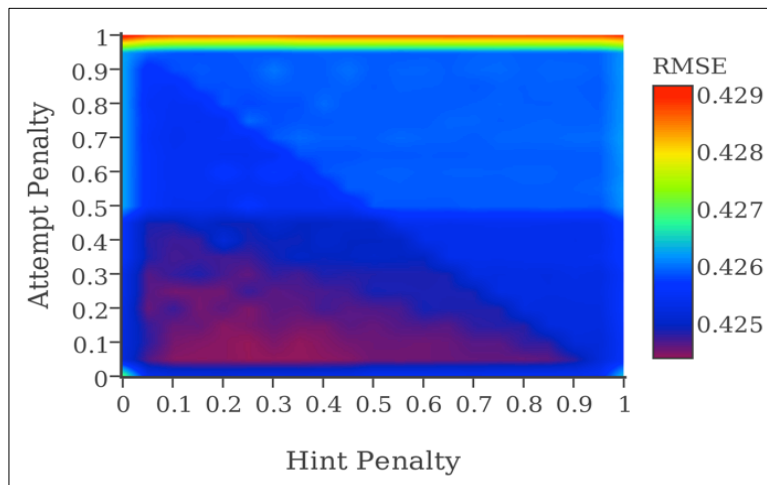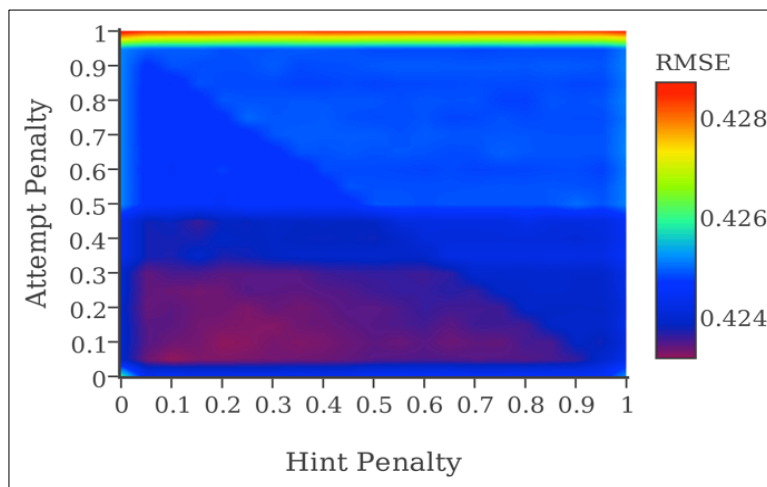**Figure 9-2. Student Level RMSE**

Figure 9-3. Skill Level RMSE

Post-hoc analyses were conducted on ANOVA results using multiple comparisons to examine significant differences between attempt penalty and hint penalty model groups when considering problem level AUC. Using a Bonferroni correction to reduce Type I error, this process resulted in a series of significance estimates for penalty group comparisons (i.e., all models where attempt penalty is 0.1 compared to all models where attempt penalty is 0.3 results in a non-significant difference, p = 0.88). Results suggest that models close in penalty are less likely to differ significantly than models with greater difference in penalty. For instance, models with an attempt penalty of 0.1 are significantly different than those with an attempt penalty of 0.4, but are not significantly different than those with an attempt penalty of 0.2. This information can be used to help select optimal partial credit penalizations, as it may be more motivating and productive for students to receive smaller penalizations. Such information would allow systems like ASSISTments to define a range of possible penalizations that can then be refined by the teacher, providing all users with a greater sense of control.

DISCUSSION
The initial findings of a grid search on partial credit penalization through per unit hint and attempt docking suggest that the implementation of partial credit within adaptive tutoring systems can be established using a data driven approach that will ultimately produce stronger predictive models of student performance while ultimately enhancing the way these systems are used by students and teachers.

Our first research question was answered with a resounding "Yes," certain algorithmically derived combinations of partial credit penalization are better than others when used to predict next problem performance. Optimal partial credit models were visible in heat maps spanning three levels of data granularity and remained relatively consistent across granularities, thus answering our second research question. ANOVAs revealed that differences in attempt penalty models were consistently significant across dataset granularities, while differences in hint penalty models were not reliable. This finding is likely due to the fact that hint usage is lower and less distributed than attempt count across problems in the dataset, and it is possible that this finding would diminish in a system that more readily promoted the use of tutoring feedback without penalization, or a system already employing partial credit scoring.

The partial credit models that we define here as optimal, based on their ability to predict next problem performance, were models with per hint and per attempt penalties of 0.3 or less. Additional analyses revealed that at the problem level, there should be no reliable difference in predictive ability of a model penalizing 0.3 per attempt from a model penalizing 0.1 per attempt, with variable hint penalization. This finding suggests that less penalization is just as effective, offering an opportunity to consider student motivation and affect when defining a partial credit algorithm. This grid search also revealed that partial credit metrics outperform binary metrics when predicting next problem performance, as previously shown in (Ostrow, Donnelly, Adjei, & Heffernan, 2015). Thus, it is possible to improve prediction of student performance within adaptive tutoring systems simply by implementing partial credit scoring. It should also be noted that a leading limitation of the approach presented here is that we have only been predicting next problem correctness, rather than latent variables such as skill mastery or student knowledge. It is possible that optimizing partial credit would also provide benefits for the prediction of latent effects, but further research is necessary in this domain.

Chapter 10 – Conclusions and Future Work

Conclusions
The work presented in this thesis has combined a set of publications and prepublications pertaining to student motivation and learning within ASSISTments. For many students, adaptive learning platforms will enhance the future of education. For others, these systems may serve to make learning even more frustrating than it already is. The goal of my work is to leverage the capabilities of technology to expand the beneficial effects of adaptive platforms to as many students as possible. Essentially, this goal is the

premise behind Bror Saxberg's (2015) concept of allowing education to thrive using technology at scale through learning engineering. Enhancing student engagement and performance within adaptive tutoring platforms can be accomplished in a variety of ways, and the secret may not be the same for all students. A multifaceted approach must be employed along with intelligent and adaptive content and feedback delivery before we will see a universal benefit. It is not clear that systemic change can arise from a handful of randomized controlled trials. However, the work presented in this thesis combines to form talking points for future work within the ASSISTments Platform:

- Video shows the potential for benefit when used as feedback within the platform. Multiple analyses have revealed that matched content video feedback, delivered through 15-30 second snippets can increase student performance. Observed improvements include requiring fewer interactions with feedback, spending longer within the feedback experience, and answering the next question more accurately and more efficiently. These findings may not be true across all students, but establishing that there is a time and a place to employ this type of feedback is a contribution to the field.

- Growth Mindset may be somewhat validated in adaptive tutoring systems, as students who self-report a growth mindset approach to learning show better overall performance. However, it is not yet clear that mindset can be altered through motivational messages presented alongside questions content or feedback. Dweck's work (2006) resorts to longer and more formal mindset training that requires full lessons alongside the type of messaging used here before effectiveness of the intervention is observed. Additional work will be required using adaptive tutoring systems to understand how these platforms can be leveraged to intervene on mindset.

- Student choice seems to play a significant role in assignment performance for some students but not for others. Thus far, the effect of choice has only been examined using feedback medium within a single skill topic. Future work is necessary to pinpoint the subtleties of this effect.

- Simple alterations in content delivery can lead to stronger performance at posttest; with low performing student impacted more significantly by the effects of interleaving skill content. Little research has been conducted to consider the proper order and spacing of skill content within adaptive platforms. The work presented in this thesis suggests that while the effect is differential across students, it is not deleterious to any particular group. Therefore, current advice is to incorporate interleaving more frequently while observing additional aspects of performance. Further work is required to investigate the in-assignment implications of interleaving (i.e., desirable difficulties). Given that the use of detectors to pinpoint student frustration within these platforms is a common approach within the field, future work should also employ these tools to examine the qualitative effect on student behavior.

- Partial credit is an important assessment approach, which although difficult to properly define, should be implemented more often in adaptive tutoring platforms. Traditional education is not commonly restricted to binary performance, and yet the majority of adaptive learning systems take this approach. The use of partial credit scoring can help to motivate students, enhance proper system usage (reduce fear or anxiety surround the use of feedback), and can greatly enhance predictions from student models. Partial credit can be established using a data driven approach, although the process should be iterative (i.e., transferring from a binary system to a partial credit system may require an initial data driven of how students use the platform, followed by updates to the partial credit algorithm based on how students adapt when partial credit is implemented).

As my work transitions toward a dissertation, I expect my research to grow and meld into a unified construct surrounding the enhancement of student motivation and learning within adaptive tutoring systems. Through continued investigation of the facets presented in this thesis, I hope to establish a unique line of research that promotes learning engineering. The immediate impact of my research is already evident through continued improvements to the ASSISTments platform. The work presented here has inspired content expansion as well as infrastructure changes to enhance future research design. Within the next three years I expect that my research will continue to refine ASSISTments while increasing

intellectual merit in my field. The broader impact of my work will be measured in long-term achievements that affect systemic change in education and promote data driven practices and individualized learning via adaptive tutoring platforms.

Future Work

It is difficult to consider the future of my research without understanding the directions being sought by the ASSISTments platform. One of the most promising goals that ASSISTments is currently seeking to achieve is crowdsourcing of tutoring feedback from teachers and students. This endeavor intends to establish the platform as a 'Wikipedia' for educational questions and feedback. While teachers have always been able to build their own content, the system is moving toward an environment in which teachers and students will be asked to explain their solution to previously established content. This crowdsourcing of hints and common wrong answers will allow new and unique research to investigate questions surrounding how content is best taught. These explanations will use both text and video to expand upon the collection of feedback already provided by the platform. It is thought that various teachers will have differing explanations and that struggling students may respond more effectively to versatile solutions or worked examples rather than a repetitive approach. A select group of teachers and students have already recorded video feedback for use in experiments examining the potential benefits of crowdsourcing, as noted earlier. This goal of scaling this design naturally gives way to another goal for the future of ASSISTments: establishing an automated process for selecting optimal feedback using contextual k-armed bandits. Implementing k-armed bandits will allow us to test system content and feedback automatically, allowing the new versions of material designed by researchers to be adaptively delivered to the right students at the right times. This will provide opportunities for dynamic versioning of materials, and truly adaptive personalization to student users. Another feature working with the k-armed bandits concept will be the capability for storing user variables for lasting personalization. Storing user variables based on initial performance, student responses, or specific student characteristics will allow for the ultimate in adaptive content and feedback delivery, as well as ideal personalization in later assignments built by researchers and teachers. These new features will allow endless hypotheses to form around the effects of student preference and the proper delivery of system content and feedback. ASSISTments is also designing a partial credit algorithm for implementation within the system. This will allow additional research pertaining to the motivational effects of partial credit scoring and the benefits that continuous assessment will have for student modeling.

References

Arroyo, I., Burleson, W., Tai, M., Muldner, K., Woolf, B.P. (2013). Gender differences in the use and benefit of advanced learning technologies for mathematics. Journal of Educational Psychology. 105, 4, 957-969.

Attali, Y. & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-end questions. *Educational and Psychological Measures*, 70 (1), 22-35.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A. & Koedinger, K. (2008). Why students engage in "Gaming the System" behavior in interactive learning environments. *Journal of Interactive Learning Research*. 19, 2, 185-224.

Bernacki, M. L., Nokes-Malach, T. J., & Aleven, V. (2013). Fine-grained assessment of motivation over long periods of learning with an intelligent tutoring system: Methodology, advantages, and preliminary results. In *International handbook of metacognition and learning technologies*. Springer New York. 629-644.

Birnbaum M.S., Kornell N., Bjork E.L., Bjork R.A. (2013). Why interleaving enhances inductive learning: the roles of discrimination and retrieval. *Mem Cogn,* 41, 392–402.

Bjork, R.A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalf & A.P. Shimamura (eds.) Metacognition: Knowing about knowing. 185-295. Cambridge, MA: MIT Press.

CEM. (2013). Effect Size Calculator. Centre for Evaluation & Monitoring, Durham University. Accessed 11/8/2013 at https://tinyurl.com/nt4snvo.

Cepeda, N.J., Pashler, H., Vul, E., Wixted, J.T. & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psy Bulletin*, 132, 354-380.

Clark, R.C. & Mayer, R. E. (2003). *e-Learning and the science of instruction: proven guidelines for consumers and designers of multimedia learning.* San Francisco, CA: Pfeiffer.

Corbett, A.T., Anderson, J.R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4: 253-278.

Cordova, D.I. & Lepper, M.R. (1996). Intrinsic Motivation and the Process of Learning: Beneficial Effects of Contextualization, Personalization, and Choice. Journal of Educational Pyschology, 88 (4). Pp. 715-730.

Drasgow, F. & Hulin, C.L. (1990). Item response theory. In M.D. Dunnette & L.M. Hough (Eds.), Handbook of Industrial and Organizational Psychology, Vol. 1, pp 577-636. Palo Alto, CA: Consulting Psychologists Press.

Duong, H.D., Zhu, L., Wang, Y., & Heffernan, N.T. (2013). A Prediction Model Uses the Sequence of Attempts and Hints to Better Predict Knowledge: Better to Attempt the Problem First, Rather Than Ask for a Hint. In S. D'Mello, R. Calvo, & A. Olney (Eds.) Proceedings of the 6th International Conference on Educational Data Mining. 316-317.

Dweck, C.S. (2002). Messages that motivate: how praise molds students' beliefs, motivation, and performance (in surprising ways). Improving Academic Achievement: Impact of Psychological Factors in Education. Ed. Joshua Aronson. New York.

Dweck, C.S. (2006). Mindset: The new psychology of success. Random House.

Dweck, C.S. (2013). Mindsets: Helping Students Fulfill Their Potential. Smith College Lecture Series, North Hampton, MA. September 19.

Educreations, Inc. (2014). Educreations Interactive Whiteboard. (Version 2.0.3) [iPad application]. Retrieved from http://itunes.apple.com.

Frenzel, A.C., Pekrun, R. & Goetz, T. (2007). Girls and mathematics – A "hopeless" issue? A control-value approach to gender differences in emotions towards mathematics. European Journal of Psychology of Education. 22 (4). Pp. 497-514.

Goodwin, B. & Miller, K. (2013). Research says/evidence on flipped classrooms is still coming in. *Educational Leadership: Technology-Rich Learning*. 70, 6, 78-80.

Gowda, S.M., Baker, R.S.J.D, Pardos, Z., & Heffernan, N.T. (2011). The Sum is Greater than the Parts: Ensembling Student Knowledge Models in ASSISTments. Proceedings of the KDD 2011 Workshop on KDD in Educational Data.

Graesser, A., Chipman, P., King, B., McDaniel, B., & D'Mello, S. (2007). Emotions and learning with autotutor. *Frontiers in Artificial Intelligence Applications*, *158*, 569.

Heffernan N. & Heffernan C. (2014) The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. Int J Art Intel in Ed. 24 (4), pp. 470-497.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. International Journal of Human-Computer Studies. 65, 724-736.

Kelly, K., Heffernan, N., D'Mello, S., Namais, J. & Strain, A.C. (2013). Adding Teacher-Created Motivational Video to an ITS. In Proceedings of the Florida Artificial Intelligence Research Society Conference (FLAIRS). pp. 503-508.

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* (61-78). New York: Cambridge University Press.

Kornell, N. & Bjork, R.A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, 19, 585-592.

LeBlanc, K., & Simon, D. (2008). *Mixed practice enhances retention and JOL accuracy for mathematical skills*. 49th Annual Meeting of the Psychonomic Society, Chicago, IL.

Li, N., Cohen, W., & Koedinger, K. R. (2012). Problem order implications for learning transfer. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), Proceedings of the 11th Int Conf on ITS, 2012. pp. 185-194. Berlin: Springer.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M.W., Roberts, M., et al. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*. Washington, DC.

Makel, M.C., & Plucker, J.A. (2014). Facts Are More Important Than Novelty: Replication in the Education Sciences. *Educational Researcher*. AERA.

Mayer, R.E. (Ed). (2005). *The Cambridge handbook of multimedia learning*. New York: Cambridge University Press.

Mayer, R.E. (2014). Incorporating motivation into multimedia learning. Learning and Instruction. Volume 29, 171-173.

Mayfield, K.H., & Chase, P.N. (2002). The effects of cumulative practice on mathematics problem solving. *J of Applied Behavior Analysis*, 35, 105-123.

Mendicino, M., Razzaq, L. & Heffernan, N.T. (2009). Comparison of Traditional Homework with Computer Supported Homework. *J of Research on Tech in Ed*. 41 (3), 331-358.

Mueller, C. & Dweck, C. (1998). Praise for Intelligence Can Undermine Children's Motivation and Performance. Journal of Personality and Social Psychology, Vol. 75, No. 1, 33-52.

Murayama, K., Pekrun, R., Lichtenfeld, S. & vom Hofe, R. (2013). Predicting Long-Term Growth in Students' Mathematics Achievement: The Unique Contributions of Motivation and Cognitive Strategies. Child Development. 84 (4) pp. 1475-1490.

Murphy, K. (2001). The Bayes Net Toolbox for MATLAB. Computing Science and Statistics, 33(2), 1024-1034.

Ostrow, K.S. (2013a). Motivational Message Study. Accessed 12/12/2013. Student Experience, RCT & All Data: https://sites.google.com/site/korinnostrow/research

Ostrow, K. (2013b). Pythagorean Theorem Math Study. Accessed 11/8/2013. Student Experience by Group, Randomized Controlled Trial, Data: https://tinyurl.com/lr529jp. Video Scaffolds: https://tinyurl.com/mcc4w8z

Ostrow, K. (2014a). L@S 2015 Submission: Dataset. Retrieved 10/14/14, http://tiny.cc/LaS2015Submission

Ostrow, K. (2014b). Optimizing Partial Credit Data. Accessed 12/8/14. https://tiny.cc/OptimizingPartialCredit

Ostrow, Korinn. (2015a). Materials for Study on Student Choice within Adaptive Tutoring. Retrieved 1/14/15 from http://tiny.cc/AIED-2015-Choice.

Ostrow, Korinn. (2015b). Materials for Study on Blocking vs. Interleaving. Retrieved on January 13, 2015 from http://tiny.cc/AIED-2015-Interleaving.

Ostrow, K., Donnelly, C., Adjei, S. & Heffernan, N. (2015). Improving Student Modeling Through Partial Credit and Problem Difficulty. In Russell, Woolf & Kiczales (Eds.), Proceedings of the 2nd ACM Conf on L@S. pp. 11-20.

Ostrow, K., Donnelly, C., & Heffernan, N. (In Press). Optimizing Partial Credit Algorithms to Predict Student Performance. To be included in Romero, C., Pechenizkiy, M., Boticario, J.G., & Santos, O.C. (Eds.), Proceedings of the 8th International Conference on Educational Data Mining. Madrid, Spain.

Ostrow, K.S. & Heffernan, N.T. (2014). Testing the Multimedia Principled in the Real World: A Comparison of Video vs. Text Feedback in Authentic Middle School Math Assignments. In Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B.M. (eds) Proceedings of the 7th International Conference on Educational Data Mining. pp. 296-299.

Ostrow, K. S. & Heffernan, N. T. (In Press). The Role of Student Choice Within Adaptive Tutoring. . To be included in Conati, C., Heffernan, N., Mitrovic, A., & Verdejo, M. (Eds.) Proceedings of the 17th International Conference for Artificial Intelligence in Educations (AIED). Madrid, Spain. pp. forthcoming.

Ostrow, K., Heffernan, N., Heffernan, C., & Peterson, Z. (In Press). Blocking vs. Interleaving: Examining Single-Session Effects within Middle School Math Homework. To be included in Conati, C., Heffernan, N., Mitrovic, A., & Verdejo, M. (Eds.) Proceedings of the 17th International Conference for Artificial Intelligence in Educations (AIED). Madrid, Spain. pp. forthcoming.

Ostrow, K. S., Schultz, S. E., & Arroyo, I. (2014). Promoting Growth Mindset Within Intelligent Tutoring Systems. In CEUR-WS (1183), Gutierrez-Santos, S., & Santos, O.C. (eds.) EDM 2014 Extended Proceedings. In Ritter & Fancsali (eds.) NCFPAL Workshop. London, United Kingdom, July 4-7. pp. 88-93.

Pane, J.F. (1994). Assessment of the ACSE science learning environment and the impact of movies and simulations. Carnegie Mellon University, *School of Computer Science Technical Report CMU-CS-94-162*, Pittsburgh, PA.

Pardos, Z.A. & Heffernan, N.T. (2010). Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization. 255-266.

Pardos, Z.A., & Heffernan, N.T. (2011). KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. In Joseph A. Konstan et al. (Eds.): UMAP 2011, LNCS 6787, 243-254.

Pardos, Z.A. & Heffernan, N.T. (2012). Tutor Modeling vs. Student Modeling. Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, 420-425.

Patall, E.A., Cooper, H., & Robinson, J.C. (2008). The Effects of Choice on Intrinsic Motivation and Related Outcomes: A Meta-Analysis of Research Findings. Psychology Bulletin. 134 (2), pp 270-300.

Pavlik, P.I., Cen, H., Koedinger, K.R. (2009). Performance Factors Analysis - A New Alternative to Knowledge Tracing. In: Proceedings of the 14th International Conference on Artificial Intelligence in Education, Brighton, UK, 531-538.

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. Educational Psychology Review, 18 (4), Emotion Research in Education: Theoretical and Methodological Perspectives on the Integration of Affect, Motivation, and Cognition. Pp. 315-341.

Rau, M.A., Aleven, A., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? *Learning and Instr*. 23, 98-114.

Rau, M.A., Aleven, V., Rummel, N., Pacilio, L., & Tunc-Pekkan, Z. (2012). How to schedule multiple graphical representations? A classroom experiment with an intelligent tutoring system for fractions.*The future of learning: Proceedings of the 10th ICLS*.

Razzaq, L. & Heffernan, N.T. (2006). Scaffolding vs. hints in the ASSISTments system. In Ikeda, Ashley & Chan (Eds). *Proceedings of the Eighth International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. 635-644.

Reeve, J. (2009). Understanding motivation and emotion. (5th ed.). Hoboken, NJ: Wiley

Roediger, H.L. (2012). Psychology's woes and a partial cure: the value of replication. The Academic Observer, The Association for Psychological Science. Retrieved on 9/30 from http://tiny.cc/RoedigerReplication.

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, 24, 355–367.

Rohrer, D. & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, 39 (5), pp. 406–412.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics practice problems boosts learning. *Instructional Science*, 35, 481-498.

Saxberg, B. (2015). Learning Engineering: The Art of Using Learning Science at Scale to Lift Performance. Worcester Polytechnic Institute, Learning Sciences and Technologies Invited Speaker. Worcester, MA. April 24.

Shea, J.B. & Morgan, R.L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, 5(2), 179-187.

Singh, R., Saleem, M., Pradhan, P., Heffernan, C., Heffernan, N., Razzaq, L. Dailey, M. O'Connor, C. & Mulchay, C. (2011). Feedback during web-based homework: the role of hints. In Biswas et al. (Eds). *Proceedings of the Artificial Intelligence in Education Conference*. 328–336.

Taylor, K. & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, 24, 837-848.

U.S. Department of Education. (2010a). *Transforming American Education: Learning Powered by Technology*. Office of Educational Technology, Washington, D.C.

U.S. Department of Education. (2010b). *Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies*. Office of Planning, Evaluation, and Policy Development. Washington, D.C.

van Seters, J.R., Ossevoort, M.A., Tramper, J & Goedhart, M.J. (2012). The influence of student characteristics on the use of adaptive e-learning material. Computers & Education, 58. Pp. 942-952.

Wang, Y. & Heffernan, N. (2011). The "assistance" model: leveraging how many hints and attempts a student needs. In Proceedings of the *Florida Artificial Intelligence Research Society Conference* (FLAIRS 2011).

Wang, Y. & Heffernan, N. (2013). Extending Knowledge Tracing to Allow Partial Credit: Using Continuous versus Binary Nodes. In K. Yacef et al. (Eds.) AIED 2013, LNAI 7926, 181-188.