A FOUNDATION FOR EDUCATIONAL RESEARCH AT SCALE: EVOLUTION AND APPLICATION

by

Korinn S. Ostrow

A Dissertation

Submitted to the Faculty

of

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

in

Learning Sciences & Technology

April 23, 2018

APPROVED:

____

Neil T. Heffernan, Ph.D., Major Advisor

____

Erin R. Ottmar, Ph.D.

Joe Beck

____

Joseph E. Beck, Ph.D.

____

Ryan S. Baker, Ph.D., University of Pennsylvania, Graduate School of Education

Abstract

The complexities of how people learn have plagued researchers for centuries. A range of experimental and non-experimental methodologies have been used to isolate and implement positive interventions for students' cognitive, meta-cognitive, behavioral, and socio-emotional successes in learning. But the face of learning is changing in the digital age. The value of accrued knowledge, popular throughout the industrial age, is being overpowered by the value of curiosity and the ability to ask critical questions. Most students can access the largest free collection of human knowledge (and cat videos) with ease using their phones or laptops and omnipresent cellular and Wi-Fi networks. Viewing this new-age capacity for connection as an opportunity, educational stakeholders have delegated many traditional learning tasks to online environments. With this influx of online learning, student errors can be corrected with immediacy, student data is more prevalent and actionable, and teachers can intervene with efficiency and efficacy. As such, endeavors in educational data mining, learning analytics, and authentic educational research at scale have grown popular in recent years; fields afforded by the luxuries of technology and driven by the age-old goal of understanding how people learn.

This dissertation explores the evolution and application of ASSISTments Research, an approach to authentic educational research at scale that leverages ASSISTments, a popular online learning platform, to better understand how people learn. Part I details the evolution and advocacy of two tools that form the research arm of ASSISTments: the ASSISTments TestBed and the Assessment of Learning Infrastructure (ALI). An NSF funded Data Infrastructure Building Blocks grant (#1724889, $494,644 2017-2020), outlines goals for the new age of ASSISTments Research as a result of lessons learned in recent years. Part II details a personal application of these research tools with a focus on the framework of Self Determination Theory. The primary facets of this theory, thought to positively affect learning and intrinsic motivation, are investigated in depth through randomized controlled trials targeting Autonomy, Belonging, and Competence. Finally, a synthesis chapter highlights important connections between Parts I & II, offering lessons learned regarding ASSISTments Research and suggesting additional guidance for its future development, while broadly defining contributions to the Learning Sciences community.

Acknowledgements

First and foremost, I would like to extend my gratitude to the many people who have made this dissertation possible. Most are listed below, by their role or level of support in their patronage. I love and appreciate every last one of them and would not be at this finish line without their love, support, patience, efforts, collaboration, or belief in me. My sincere gratitude to all listed below, and to anyone who helped along the way but I have managed to omit, thank you.

Loves of My Life

Sam & Louie

Family

Fern & Gary
Kim & Ryan
Brooke
Logan & Nick
Brud* & Judy
Chrissy* & Jenny
Lisa
Suzanne

Marcy & Matthew
Rachel, Chris, Desmond & Petra
Emma, Chris & Margot
Naomi
Madeline
Natalie & Bob
Zoë

Friends

Heidi & Steven
Aaron
Heath & Kait
Lauren, Evan, & Grayson
Chris & Becca
Anthony & Paige
Taylyn & Logan

Avery & Cole
Kim & Dan
Sabrina
Seth
Dan
Sarah

Co-Authors

Neil
Sarah
Ivon
Joseph
Xiaolu
Chris
Seth
Cristina

Zoe
Charles
Doug
Kim
Eric
Yan
Joe
Erin

Gbeton
Paul
Peter
Shihfen
Craig
Mary Ellin
Noboru
Yeping

Pat
Erik
Ryan
Mia
Jaclyn
Stefan

Benefactors & Sages

Neil & Cristina
Tricia
Andrew
Erin

Joe
Ivon
Ryan
Sabrina & Amy

*Lost during my candidacy but never forgotten.

Funding

The work presented in the following pages has been supported by a number of funding avenues. Specifically, my work has been funded by a Partnership in Math and Science Education (PIMSE) Fellowship (NSF DGE-0742503), a Graduate Assistance in Areas of National Need (GAANN) Fellowship from the U.S. Department of Education (P200A120238), and the Office of Naval Research (N00014-13-C-0127). More broadly, funding is acknowledged from NSF (grant #'s DRL-1316736, DRL-1252297, DRL-1109483, DRL-1031398, 0742503, ACI-1440753, DGE-1535428, 0231773, 0448319, OAC-1724889, and OAC-1636782), ONR's "STEM Grand Challenges," IES (grant #'s R305A120125, R305C100024, R305K03140, R305A07440), The Spencer Foundation, and The Bill and Melinda Gates Foundation. The views presented in this work are not necessarily endorsed by these entities.

Table of Contents

List of Tables

List of Figures

FOREWORD

It has been my privilege to spend the last five years harnessing the powers of the digital age to study how students learn. This dissertation offers a collection of work that epitomizes my efforts, delivered in two parts. Part I describes the infrastructures that form ASSISTments Research and how my work has guided their iterative refinement and advocated for their use within the community. Part I also includes the NSF funded DIBBs grant proposal that will shape the continued evolution of ASSISTments Research. Part II highlights my personal application of ASSISTments Researcher, featuring a set of randomized controlled trials that target basic psychological needs as defined by Self-Determination Theory. A closing synthesis integrates Parts I & II and describes the broader contributions this work has made to the Learning Sciences community.

As this dissertation culminates my time at WPI, the story of my graduate experience helps to properly introduce included works. When I began my research career with the Learning Sciences & Technologies program in 2013, ASSISTments was used primarily for private research examining the efficacy of learning interventions that had been hypothesized by members of the ASSISTments team or their immediate collaborators. The platform allowed for the creation of randomized controlled trials to help isolate best practices in online classwork and homework. In my first research project with ASSISTments, I tested multimedia principles by comparing two simple scenarios – whether students were offered text or video hints while working on an assignment about the Pythagorean theorem (Ostrow & Heffernan, 2014). I recorded videos of myself explaining hint content on a whiteboard and a counterbalanced design was used to allow all students the opportunity to receive both text and video hints as they completed a series of math problems. Following their assignment, students were asked four survey questions to confirm that they had been able to access video content and to learn how they felt about the videos if they were accessed (i.e., if the student got an answer wrong and/or asked for help). There were modest significant findings and students reacted positively to the addition of video, prompting an influx of video content to ASSISTments. All data was collected from the ASSISTments database and compiled by hand using a problem-level file format (i.e., raw tutor log files in which each row represented the characteristics of a single problem). Without programming skills, data retrieval, preprocessing, and analysis tasks took more than 120 hours. The process was tedious and there were obstacles that many other researchers would not have been willing to tackle.

The scope and accessibility of research within ASSISTments was limited. It became clear that the power of ASSISTments as a research tool was something that other researchers may find useful, but only if the process could be streamlined. Was it feasible? The team began presenting the idea of scaled educational research at conferences focused on learning analytics and educational datamining and continued to publish the results of our "in-house" research. The push to get external researchers involved spurred the proper evolution of "ASSISTments Research." The ASSISTments TestBed was born and quickly grew to host the research of external researchers interested in learning at scale. While not solely responsible for the development of the research infrastructure presented in this dissertation, I played a critical role in its evolution and advocacy (many thanks to the programmers in the ASSISTments lab that crafted these tools and considered my input along the way).

The NSF funded SI2 grant was a driving force in upending early research practices to make them scalable to researchers around the world. The grant (SI2-SSE&SSI: 1440753, Adding Research Accounts to the ASSISTments' Platform: Helping Researchers Do Randomized Controlled Studies with Thousands of Students) provided nearly half a million dollars toward establishing ASSISTments as a shared scientific instrument for the collaborative efforts of researchers in fields like educational psychology and mathematics education. Funding began in Fall 2014 as a three-year project to establish the ASSISTments TestBed (Ostrow & Heffernan, 2016; Ostrow, Heffernan & Williams, 2017) and the early stages of the Assessment of Learning Infrastructure (ALI). The resulting journal paper advocating for other online learning platforms to support randomized controlled experimentation, as prototyped by the ASSISTments TestBed, is presented in Chapter 1. Similarly, the seminal paper on the capabilities and analytic power of ALI is presented in Chapter 2. The grant enhanced the speed, quality, and reliability of conducting K-12 educational research using tools that leveraged the teacher relationships and subject pool already established by ASSISTments' decade of positive growth. It became easier to study the effects of learning interventions using minimally invasive approaches that would not disrupt classroom practices. The grant listed its ten-year goal as routinely hosting the work of hundreds of researchers, while effectively streamlining the translation of research to implementation by getting successful interventions directly into the hands of teachers.

The push to broaden external research perspectives impacted ASSISTments itself as new features were created to support experimental designs. For instance, the creation of the If-Then-Else section type in 2015 was prompted by a study I conducted to investigate the effects of providing students with a choice in their hint medium. Developed by Chris Donnelly, the new section type allowed for the development of a much more robust study design that original pilot work had employed and has since paved the way for a wide variety of research ideas that involve similar path routing (Ostrow & Heffernan, 2015).

Around the same time, the "Data Dumper" was developed as the earliest implementation of ALI. I was a novice when it came to databases and programming. In order to retrieve my data from the database, an ASSISTments programmer would help by implementing a redundant query to the database to provide problem-level CSV files that I would format and manipulate in Microsoft Excel. I learned how to connect to the database and reuse the query myself, but the process was complicated without strong programming skills. The ASSISTments Research team arrived at the goal of automating data retrieval for researchers who, like me, were interested in accessing data but lacked required skills. These researchers could soon access the database without actually touching it by providing a few pieces of information in a Google form; this was the birth of the Data Dumper. I used this form frequently and my work contributed dramatically to its evolution.

At this stage, the Data Dumper provided a single file type. I recognized that researchers would not be interested in spending time repeatedly formatting their data and I knew that we could do even better in supporting their efforts. I enrolled myself in an introductory Python programming class and began the processing of automating my approaches to data formatting. My course project resulted in the student-level file (see A.1 in the Appendix for my original Python code). My colleague Doug Selent then translated my Python program into Java to integrate it into the Data Dumper, allowing output of two file types. In working to translate my code he realized his own set of needs with a student-level file differed. Given his background with computational thinking, his version of the file relied more on filtering features and variables to produce analyses. He composed his own transformer and established the student+problem-level file. Finally, as external researchers began to work with the TestBed, interest in educational data mining and the fine-grained task of parsing action-level data led to the development of the action-level file. Within about a year, what began as a way to provide a single, rudimentary response to a researcher's call to the database became a much more complex and unique opportunity to cater to the myriad of research needs within the ASSISTments TestBed.

Simultaneously, partial credit work (Ostrow, Donnelly, Adjei & Heffernan, 2015; Ostrow, Donnelly & Heffernan, 2015; Wang, Ostrow, Beck & Heffernan, 2016; Ostrow, Wang & Heffernan, 2017) refined reports by introducing new variables and creating the "covariate" file that provided students' data gathered prior to their participation in a particular RCT. The team realized the value of partial credit through my work on blocking vs. interleaving (Ostrow, Heffernan, Heffernan, & Peterson, 2015). The study compared how the delivery pattern of three skills (A, B, C) altered learning outcomes at posttest. When skill problems were presented using an interleaved pattern (ABCABCABC), students performed significantly better at posttest than when content was delivered using a blocked pattern (AAABBBCCC). Findings revealed that when considering only posttest performance, no significant differences were observed between groups. However, when an impromptu partial credit metric was formed by using a Multivariate Analysis of Covariance (MANCOVA) that combined performance with hint usage and attempt count, significant differences were observed. Partial credit scoring was added to ASSISTments as a teacher preference and data quality and flexibility became critical concerns for the team.

The Data Dumper, now more robust with multiple reporting formats (or data granularities), came to be known as the Assessment of Learning Infrastructure (ALI) through the addition of automated reporting. ALI serves the TestBed as a data delivery tool. Run through a series of Google services, ALI allows researchers to register their studies, add tags to their design to automate analysis (i.e., experimental vs. control conditions), and access on-demand data reports. I served as a consultant for the analytics and formatting of ALI reports, working with developers Yan Wang and Doug Selent. The goal of ALI was to automate analysis and reporting via weekly emails to researchers (Ostrow, Selent, Wang, Van Inwegen, Heffernan & Williams, 2016; Ostrow & Heffernan, 2016). Researchers are provided multiple data formats (action-level, problem-level, student-level, and student+problem-level) that have been carefully crafted for ease of use. Early stages of ALI involved attempts to automate logical analyses for researchers using tags in the building structure to label conditions and posttest. Unfortunately, running these tests weekly threatened to inflate error and opened the door to cherry picking. Weekly reporting has since ceased, and reports are now delivered using on-demand requests. Future work with ALI will attempt to address this concern in a more robust way.

I also helped to brand ALI by creating its logo, “Ali” the *lab* assistant (a cartoon Labrador Retriever in a lab coat). While it was a fun part of the project that let my creativity flourish, I felt that ALI’s branding was actually an important part of the work because the tool offered a direct link to new researchers working with the ASSISTments TestBed and the logo helped to promote it as a user-friendly tool.

The SI2 grant involved approximately 20 researchers interested in designing, conducting, and analyzing studies within the ASSISTments TestBed. An early issue the team noticed with the scale of this work was the bottleneck created by the need for involvement of an ASSISTments team member at each step in the research process. Cristina Heffernan and I worked directly with external researchers to help them integrate their ideas into the system and to later analyze their data. Many of these collaborations and how they informed the evolution of the tools that form ASSISTments Research are discussed in Chapter 3.

The future of ASSISTments Research will be informed by lessons learned from four years of working with external researchers. Chapter 4 concludes Part I by offering the grant proposal I organized for the NSF funded Data Infrastructure Building Blocks (DIBBs) grant (1724889 - CIF21 DIBBs PD: Enhancing and Personalizing Educational Resources through Tools for Experimentation, $494,644) that was awarded to PI Neil Heffernan. The grant is projected to span three years (2017-2020) and describes my vision for the future of ASSISTments Research by enhancing how researchers leverage ASSISTments as a research tool. The general idea of proposed work is that by restricting the number of experimental designs readily supported by the TestBed through use of a visual interface and experimental “templates” that researchers can fill with their manipulated content, the bottleneck observed in past collaborations would be dramatically reduced in pursuit of scalable automation. Researchers would be able to build their studies by applying modified content or feedback to a template, and templates would integrate intelligently with ALI to automate analyses after a preset threshold for sample size had been met (using power analyses), or when p-value entropy had stabilized (to reduce the likelihood of error). Automated analyses would be supported by raw data files offering a series of researcher friendly formats for in-depth exploration.

Part II offers a personal application of the tools underlying ASSISTments research. While much of my time at WPI has involved helping external researchers or advocating for research at scale, I have also brought my own research interests to the table. When I began at WPI in 2013, I brought interest in cognitive science and positive psychology that had flourished during my time at Rensselaer Polytechnic Institute, while earning a B.S. in Psychology. My undergraduate thesis had examined how signature strengths and personality traits could affect learning outcomes. This interest evolved when I began conducting research with K-12 students using ASSISTments. My master's thesis focused on elements of motivation and engagement, and how interventions crafted in various studies, or factors considered in datamining endeavors, could alter students' learning outcomes. Since then, my interests have evolved further (Ostrow, 2015a; Ostrow, 2015b; Ostrow, 2016), growing even more focused toward motivational interventions. In the Spring of 2017, Self-Determination Theory (SDT) became the primary framework for my research, linking past work on motivation and choice with a well-developed and strongly supported underlying theory (Ryan & Deci, 2002).

Part II of this dissertation offers my personal application of the tools that form ASSISTments Research through a set of randomized controlled trials and supplementary content aligned to the framework of SDT. Chapter 5 offers background supporting three randomized controlled trials that target the basic psychological needs defined by SDT: autonomy, belonging, and competence. Then, Chapter 6 assesses the reliability of subscales of the Intrinsic Motivation Inventory within ASSISTments. These subscales measure motivational constructs aligning to the basic needs and serve as the primary self-report measures used in later chapters. Chapters 7-9 offer conference or journal length papers featuring randomized controlled trials that focus on motivational interventions targeting the three basic needs. Chapter 7 focuses on an intervention intended to support autonomy by providing students with choice in hint medium. Chapter 8 focuses on an intervention intended to support relatedness or belonging by promoting the human condition through a value affirmation exercise. Finally, Chapter 9 focuses on an intervention intended to support competence by assessing students' confidence in a skill and intervening with a skill lesson before or during assessment. This dissertation closes with a synthesis that summarizes Parts I & II and provides commentary on lessons learned and guidance for the future of ASSISTments Research while contextualizing the broader contributions of this work within the Learning Sciences community.

References

Ostrow, K. (2015a). Enhancing Student Motivation and Learning Within Adaptive Tutors. In Santos, Boticario, Romero, Pechenizkiy, Merceron, Mitros, Luna, Mihaescu, Moreno, Hershkovitz, Ventura, & Desmarais (eds.) Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015). Madrid, Spain, June 26-29. 668-670.

Ostrow, K. (2015b). Motivating Learning in the Age of the Adaptive Tutor. In Conati, Heffernan, Mitrovic & Verdejo (eds.) Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015). Springer International Publishing. Madrid, Spain. June 22-26. 852-855.

Ostrow, K.S. (2016). Toward a Sound Environment for Robust Learning Analytics. Learning Analytics and Knowledge Doctoral Consortium.

Ostrow, K., Donnelly, C., Adjei, S., & Heffernan, N. (2015). Improving Student Modeling Through Partial Credit and Problem Difficulty. In Russell, D.M., Woolf, B., & Kiczales, G. (eds), Proceedings of the 2nd ACM Conference on Learning at Scale (L@S 2015). Vancouver, British Columbia, March 14-15. 11-20.

Ostrow, K., & Donnelly, C., & Heffernan, N. (2015). Optimizing Partial Credit Algorithms to Predict Student Performance. In Santos, Boticario, Romero, Pechenizkiy, Merceron, Mitros, Luna, Mihaescu, Moreno, Hershkovitz, Ventura, & Desmarais (eds.) Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015). Madrid, Spain. June 26-29. 404-407.

Ostrow, K.S. & Heffernan, N.T. (2014). Testing the multimedia principle in the real world: A comparison of video vs. text feedback in authentic middle school math assignments. In Stamper, Pardos, Mavrikis, McLaren (eds.) Proceedings of the 7th International Conference on Educational Data Mining, 296-299.

Ostrow, K. & Heffernan, N. (2015). The Role of Student Choice Within Adaptive Tutoring. In Conati, Heffernan, Mitrovic & Verdejo (eds.) Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015). Springer International Publishing. Madrid, Spain. June 22-26. 752-755.

Ostrow, K.S., & Heffernan, N.T. (2016). The ASSISTments TestBed and The Assessment of Learning Infrastructure: The Evolution of Educational Research at Scale. CODE@MIT.

Ostrow, K., Heffernan, N., Heffernan, C., & Peterson, Z. (2015). Blocking vs. Interleaving: Examining Single-Session Effects within Middle School Math Homework. In Conati, Heffernan, Mitrovic, & Verdejo (eds.) Proc of the 17th Int Conf on AIED. pp. 388-347.

Ostrow, K.S., Heffernan, N.T., & Williams, J.J. (2017). Tomorrow's EdTech Today: Establishing a Learning Platform as a Collaborative Research Tool for Sound Science. *Teachers College Record.* 119(3): 1-36.

Ostrow, K., Selent, D., Wang, Y., Van Inwegen, E., Heffernan, N., & Williams, J.J. (2016). Assessment of Learning Infrastructure (ALI): The Theory, Practice, and Scalability of Automated Assessment. To be included in the *Proceedings of the 6th International Conference on Learning Analytics and Knowledge*, 2016. Edinburgh, Scotland, April 25-29.

Ostrow, K.S., Wang, Y., & Heffernan, N.T. (2017). How Flexible Is Your Data? A Comparative Analysis of Scoring Methodologies Across Learning Platforms in the Context of Group Differentiation. *Journal of Learning Analytics*.

Ryan, R. M., & Deci, E. L. (2002). An overview of self-determination theory. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 3-33). Rochester, NY: University of Rochester Press.

Wang, Y., Ostrow, K., Beck, J., & Heffernan, N. (2016). Enhancing the Efficiency and Reliability of Group Differentiation through Partial Credit. To be included in the Proceedings of the 6th International Conference on Learning Analytics and Knowledge, 2016. Edinburgh, Scotland, April 25-29.

PART I: EVOLUTION

1 ASSISTments TestBed Design & Advocacy

This chapter was originally published as a journal article in a special edition of the *Teachers College Record*. The article is included here because it was the seminal work outlining the establishment of the ASSISTments TestBed, an organized approach for leveraging ASSISTments as a collaborative research platform. The article advocated for educational research at scale and the perpetual benefits allotted by collaborative research within online learning platforms. Publication of this work mirrored the completion of an NSF SI2 grant (SI2-SSE&SSI: 1440753, Adding Research Accounts to the ASSISTments' Platform: Helping Researchers Do Randomized Controlled Studies with Thousands of Students) and served to further advocate for uptake of the TestBed by researchers around the country.

Proper citation of this chapter is as follows:

Ostrow, K. S., Heffernan, N. T., & Williams, J. J. (2017). Tomorrow's EdTech Today: Establishing a Learning Platform as a Collaborative Research Tool for Sound Science. *Teachers College Record.* 119 (3): 1-36.

1.1 Introduction

Educational psychologists, researchers, and practitioners have grown accustomed to the complex and time-consuming nature of studying effective classroom practices. When studying learning interventions, seasoned experts turn to the gold standard in determining causality: the randomized controlled experiment (RCE). Yet despite a recent call encouraging the use of RCEs within authentic learning environments (IES, 2013), and despite the nearly infinite array of complexities to be examined within the context of instruction (Koedinger, Booth, & Klahr, 2013), RCEs can be difficult to conduct in real-world classrooms (National Research Council, 2002). Common complications include IRB restrictions, lengthy and invasive pre- and post-tests, curriculum restrictions for the design of strict controls, and large sample populations required to detect significantly reliable results. Further, experimental designs must be carefully vetted prior to implementation in an attempt to account for as much variance as possible. Thorough organization is also necessary when recording and maintaining anonymized student data. With so many moving parts, traditional classroom RCEs leave numerous windows for error and bias. Even when reporting findings, publication bias and the cherry picking of results can lead to the inability for replication, contributing to a growing crisis of faith in RCEs spanning numerous scientific fields (Ioannidis, 2005; Achenbach, 2015; Open Science Collaboration, 2015). Additionally, while a handful of traditional classroom RCEs *have* led to significant implications for educational practice and policy, most lack the statistical power necessary to observe reliable improvements in student achievement because they are restricted by class- or school-level randomization (i.e., all students within a particular class or school fall within the same experimental condition, resulting in drastically reduced sample sizes). High stakes explorations at scale (i.e., stressful 'make or break' longitudinal studies costing millions of dollars) often include thousands of students and span multiple years but still fall short of identifying learning interventions that reliably enhance student achievement.

While it is crucial that high standards exist for educational research, the present work investigates the use of educational technologies to simplify the process of conducting RCEs within authentic learning environments, making research at scale more feasible and more accessible to researchers. Infusing popular learning platforms with the capacity to support collaborative research environments has the potential to lower the stakes by drastically reducing costs, promoting validated universal measures of achievement, and assisting researchers through the process of designing, implementing, and analyzing RCEs conducted at scale within real-world classrooms. Supplementing educational technologies with environments for sound, collaborative science can result in a broad range of benefits for students, researchers, platforms, and educational practice and policy.

1.1.1 The Growth of Educational Technologies

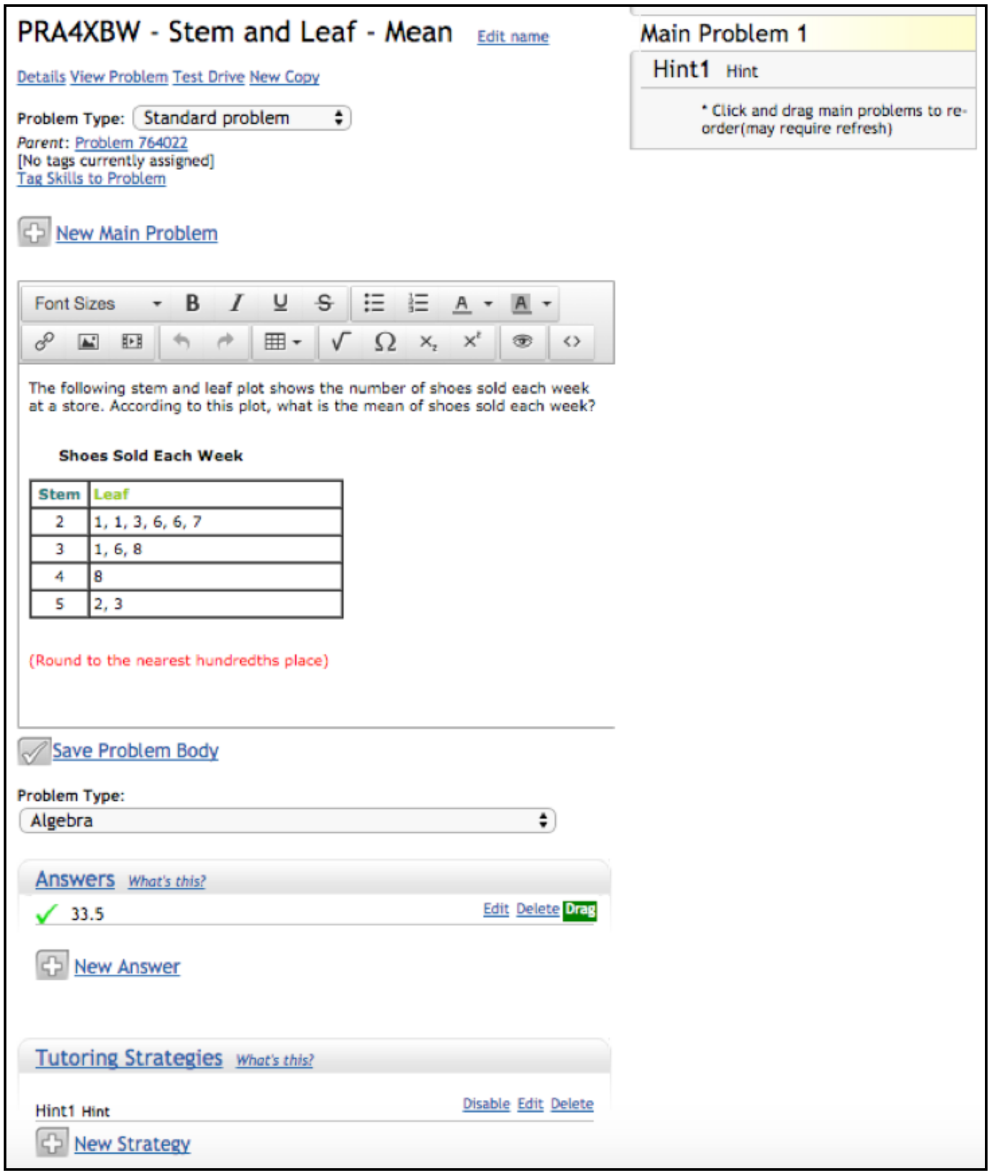
Educational technologies offer the novel opportunity to drive best practices in K-12 education by testing what works in authentic learning environments while simultaneously simplifying the process of educational research. Technology is gaining acceptance in the modern classroom, with intelligent tutoring systems, computer-aided testing platforms, and adaptive learning applications offering new and unique approaches to learning, beckoning a transition from ‘teaching’ based practices to ‘learning’ based practices (Bush & Mott, 2009), and producing exponential growth in the availability of educational data. Educational technologies commonly include immediate feedback, adaptive assistance, elements that enhance student motivation and engagement, and assessment tools for teachers and administrators that help to drive data-driven classroom practices. As such, the National Education Technology Plan predicted that these platforms would play a key role in personalizing educational interventions (U.S. Dept. of Ed., 2010). However, less focus has been devoted to one of the primary forces driving successful personalization: the use of adaptive learning technologies to conduct educational research.

These platforms and applications already have great promise for extending the accessibility of educational materials and improving learning outcomes across diverse populations. At scale, the data collected from these technologies can be leveraged in dynamic ways that may reveal revolutionary insights about learning. Entire fields of research are growing alongside educational technologies in hopes of better understanding how these tools and their data can be used to improve education (e.g., Learning Analytics, Educational Data Mining). However, despite significant growth in researcher interest, few platforms currently available to teachers and students allow for real-time hypothesis testing. In lieu of in vivo experimentation, researchers often turn to logged data to model student performance, make predictions regarding learning, and determine the effectiveness of system features (Koedinger, Baker, et al., 2010). “Big Data” in education has grown synonymous with solutions that enhance educational practices, platforms, and theories. Still, a critical link is missing: causality. Examining the causal effects of specific learning interventions through “Big Experimentation” would allow researchers to begin answering three questions to truly drive personalized education: “What works best? For whom? When?” By determining the interventions that work best for particular students and the optimal time to deliver those interventions, controlled experimentation conducted within these platforms has the potential to revolutionize the future of education.

1.1.2 The ASSISTments Platform

Despite expanse in the availability of adaptive learning technologies in recent years, popular platforms have been very slow to mobilize, support, and leverage randomized controlled experimentation (Williams, Ostrow, et al., 2015; Williams, Maldonado, et al., 2015). ASSISTments is an online learning platform that was designed with the flexibility to house RCEs and has supported the publication of more than two dozen peer-reviewed articles on learning since its inception in 2002 (Heffernan & Heffernan, 2014). The platform, offered as a free service of Worcester Polytechnic Institute (WPI), is an increasingly powerful tool that provides students with *assistance* while offering teachers *assessment*. Over $14 million in grant funding from the IES and the NSF has supported twelve years of co-development with teachers and researchers to establish a unique tool for educational research at scale. Historically, the primary investigators of these studies have had close connections to WPI (e.g., graduate students or other researchers working closely with the ASSISTments Team). However, a recent NSF grant (SI2-SSE&SSI: 1440753) has helped to launch a formal infrastructure that allows external researchers to use ASSISTments as a shared scientific tool. This supplementary infrastructure is called the ASSISTments TestBed (www.ASSISTmentsTestBed.org). While other systems have the potential to provide many of the same classroom benefits as ASSISTments, none promote an infrastructure allowing educational researchers to design and implement content-based experimentation, and to do so with ease.

Doubling its user population each year for almost a decade, ASSISTments is used by hundreds of teachers and over 50,000 students around the world, with over 10 million problems solved in the 2013-2014 school year. Although most content pertains to middle school mathematics, teachers from alternative domains like history, biology, and statistics have also built material to harness the powers of the platform in their own classrooms. Content is built at the problem level, as shown in Figure 1.1. The problem builder allows teachers and researchers to design questions and tutorial strategies using a simple interface that allows for the inclusion of text, graphics, and hypermedia elements. The builder is unique in that it allows for efficient content design without extensive knowledge of computer programming. Questions can then be combined to form problem sets for assignment to students. Teachers commonly use ASSISTments to assign classwork and homework with immediate feedback and rich tutoring, but can also turn off feedback elements to assign content as a test or quiz. Use of ASSISTments has been shown to reliably improve students' learning in comparison to traditional paper and pencil approaches (Mendicino, Razzaq, Heffernan, 2009; Koedinger, McLaughlin & Heffernan, 2010; Singh et al., 2011; Kelly, Heffernan, Heffernan, et al., 2013; Miller, et al., 2013; Soffer, et al., 2014). Most recently, SRI International reported



PRA4XBW - Stem and Leaf - Mean

Problem Type: Standard problem

Parent: Problem 764022

[No tags currently assigned]

The following stem and leaf plot shows the number of shoes sold each week at a store. According to this plot, what is the mean of shoes sold each week?

Shoes Sold Each Week

Stem	Leaf
2	1, 1, 3, 6, 6, 7
3	1, 6, 8
4	8
5	2, 3

(Round to the nearest hundredths place)

Problem Type: Algebra

✓ 33.5

Hint1 Hint

Main Problem 1

Hint1 Hint

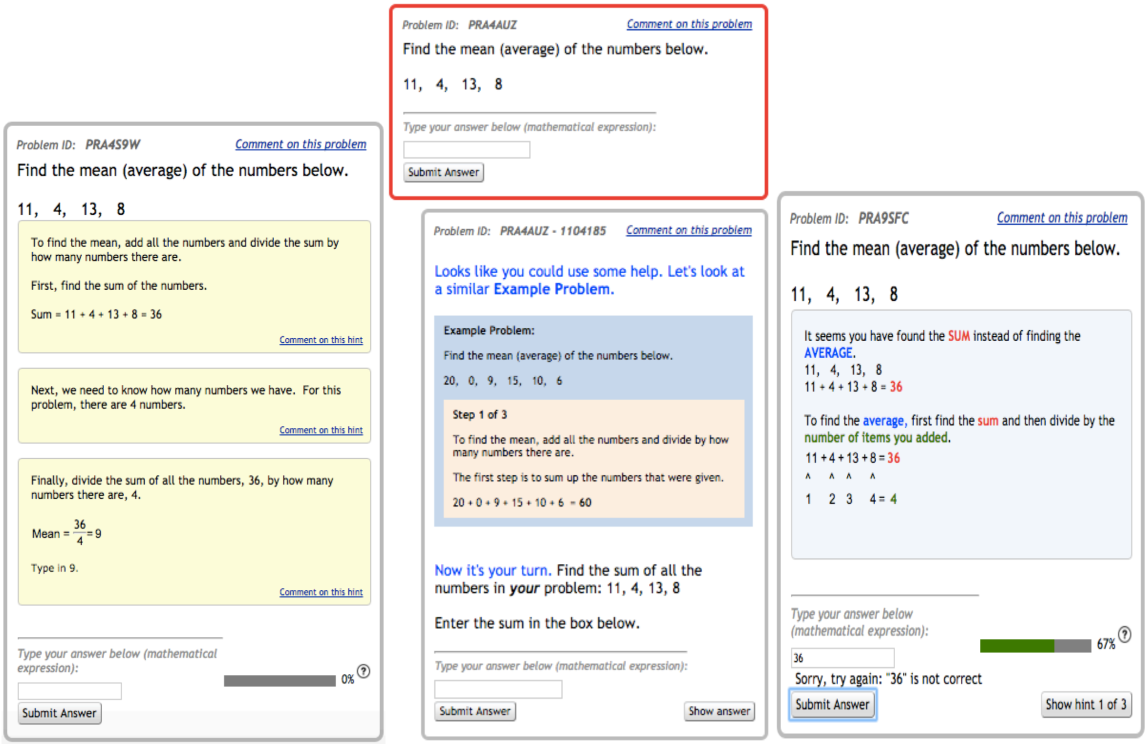
* Click and drag main problems to re-order(may require refresh)

Figure 1.1. An example of a problem viewed within the builder. Notice the interface allows creation of the problem itself, answers (both correct and incorrect), and tutoring strategies. The navigation menu in the top right corner allows the user to navigate from editing a main problem to editing feedback.

early results of an efficacy trial of ASSISTments, showing that the platform caused large, reliable, learning gains on standardized assessments (Roschelle, et al., 2015).

In addition to building content, teachers and researchers are able to access an extensive library of prebuilt content and textbook material. Full problem content is available for more than 20 of the top 7th grade mathematics texts in the United States, delivered without infringing on copyright. Teachers can select from prebuilt problem sets, or use and alter copies of content to develop their own problem sets. There are two primary types of problem sets within ASSISTments. A *linear* problem set has a predetermined number of problems and the assignment is considered complete when the student has finished all problems, whether or not the answers are accurate. Alternatively, in a *Skill Builder* problem set, students must solve problems selected at random from a skill pool until reaching a predetermined threshold of mastery (i.e., answering three consecutive questions accurately on first attempts). Although the system default is three problems, mastery can be redefined to include any number of consecutive accurate problems. In both types of problem sets, assistance can vary to include correctness feedback, tutoring specific to particular problems, or worked examples depicting solutions to isomorphic problems. Tutoring strategies include hint messages, scaffolding problems (used to break a problem down into steps), and mistake messages (feedback tailored to common wrong answers). Hints, scaffolds, and mistake messages are compared in Figure 1.2. If researchers do not wish to design their own content, over 300 'Certified' Skill Builders (tailored by the ASSISTments Team to the Common Core State Standards for Mathematics (NGACBP & CCSSO, 2010)), can be manipulated to incorporate experimental modifications.

ASSISTments also offers optional features like the *Automatic Reassessment and Relearning System* (ARRS), which helps to reassess student retention following Skill Builder mastery (Xiong & Beck, 2014), and PLACEments, a prerequisite skill training system that allows teachers to create skill tests that pinpoint and help to alleviate knowledge gaps (Whorton, 2013). When a teacher


Problem ID: PRA4AUZ

Find the mean (average) of the numbers below.

11, 4, 13, 8

Type your answer below (mathematical expression):

Problem ID: PRA4S9W

Find the mean (average) of the numbers below.

11, 4, 13, 8

To find the mean, add all the numbers and divide the sum by how many numbers there are.

First, find the sum of the numbers.

Sum = 11 + 4 + 13 + 8 = 36

Next, we need to know how many numbers we have. For this problem, there are 4 numbers.

Finally, divide the sum of all the numbers, 36, by how many numbers there are, 4.

Mean = $\frac{36}{4}$ = 9

Type in 9.

Type your answer below (mathematical expression):

Problem ID: PRA4AUZ - 1104185

Looks like you could use some help. Let's look at a similar **Example Problem.**

Example Problem:

Find the mean (average) of the numbers below.

20, 0, 9, 15, 10, 6

Step 1 of 3

To find the mean, add all the numbers and divide by how many numbers there are.

The first step is to sum up the numbers that were given.

20 + 0 + 9 + 15 + 10 + 6 = 60

Now it's your turn. Find the sum of all the numbers in *your* problem: 11, 4, 13, 8

Enter the sum in the box below.

Type your answer below (mathematical expression):

Problem ID: PRA9SFC

Find the mean (average) of the numbers below.

11, 4, 13, 8

It seems you have found the SUM instead of finding the AVERAGE.
11, 4, 13, 8
11 + 4 + 13 + 8 = 36

To find the average, first find the sum and then divide by the number of items you added.
11 + 4 + 13 + 8 = 36
^ ^ ^ ^
1 2 3 4 = 4

Type your answer below (mathematical expression):

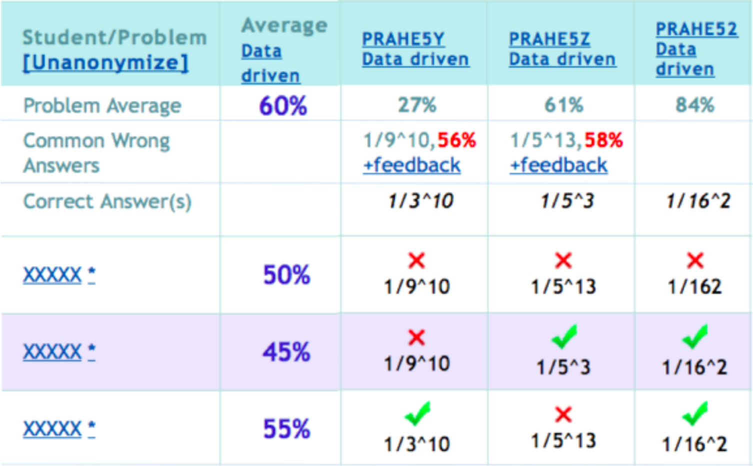
36

Sorry, try again: "36" is not correct

Figure 1.2. A comparison of hints, a scaffold problem, and a mistake message in response to the same problem content. Three hints are shown on the left, as requested by the student. In the middle, the student provided an incorrect response and was automatically given a scaffold with a worked example on how to solve a similar problem. On the right, a mistake message is provided in response to a specific wrong answer, with detailed tutoring on strategy revision.

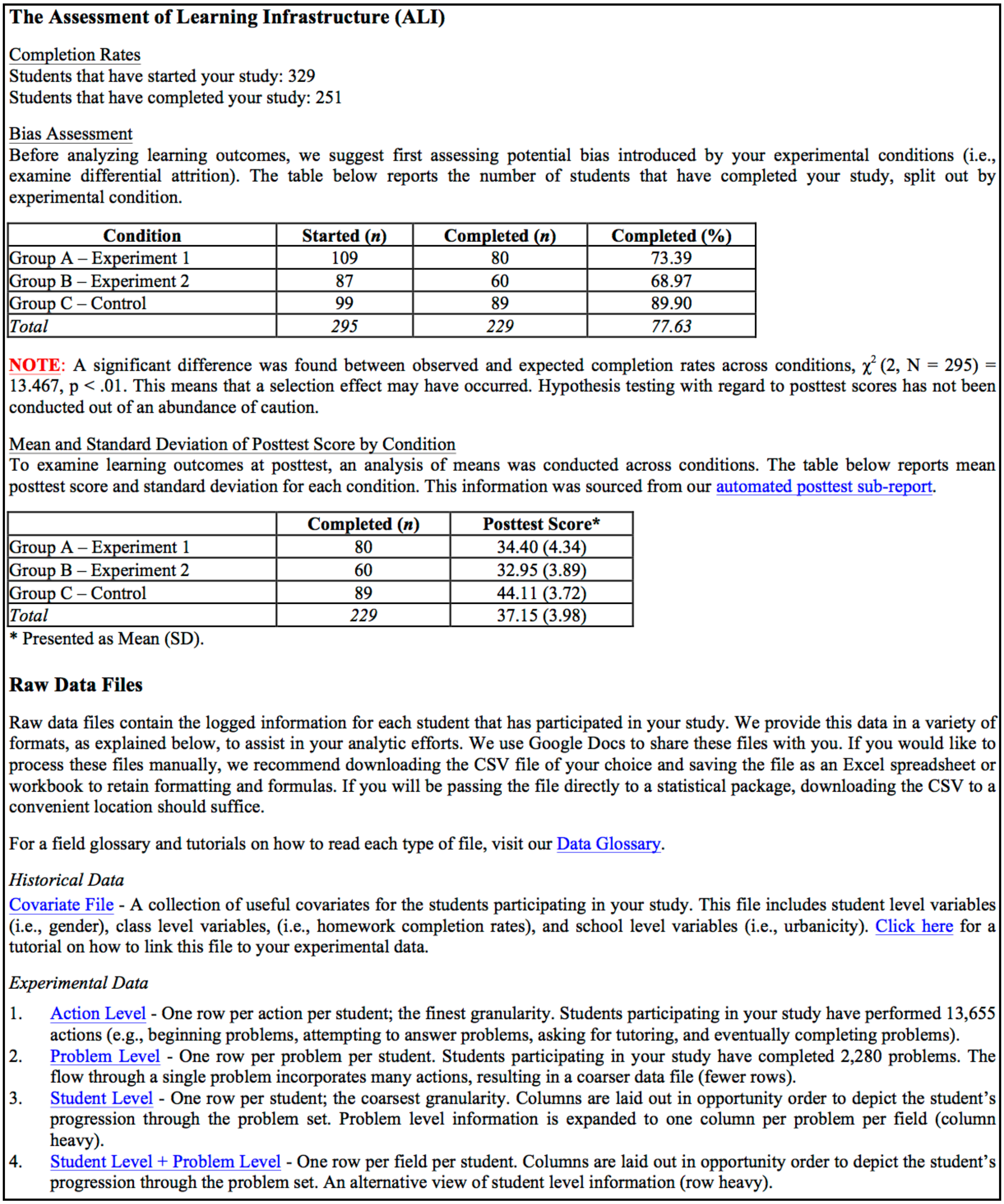
elects to use ARRS, after completing a Skill Builder, students are given a series of single question reassessment tests, scheduled 7, 14, 28, and finally 56 days after the initial learning experience to estimate skill retention. If students fail to answer the reassessment question accurately, they are provided support to relearn the material through a secondary Skill Builder. Research has shown that ARRS significantly enhances longitudinal skill understanding and student assessment (Soffer, et al., 2014; Wang & Heffernan, 2014). Like ARRS, PLACEments is also connected to Skill Builder content. PLACEments acts as a computer adaptive test that taps into a hierarchy of prerequisite skills to personalize the remediations a student should receive based on performance in an initial skill test. Research has shown that PLACEments is a useful tool for isolating learning gaps that can also help to strengthen curriculum through a stronger understanding of prerequisite skill relationships (Adjei & Heffernan, 2015).

As an assessment tool, ASSISTments offers teachers a myriad of student and class reports that allow an expanse in classroom practices through actionable data. An example of an Item Report, the most commonly used report, is shown in Figure 1.3. This report has a column for each problem and a row for each student, as well as various summaries of student and problem performance. The report can be made anonymous (as shown in Figure 1.3) for teachers to use in the classroom to facilitate discussion. This report also allows teachers to pinpoint areas of struggle through common wrong answers (errors that were made by at least 10% of students in the class). In Figure 1.3, only 27% of students answered the first problem accurately, with 56% of students sharing the common wrong answer of 1/9^10. This offers an opportunity for discussion that may be lost on students grading their own homework using traditional classroom methods. Teachers can also work with students to design mistake messages (like that shown in Figure 1.2) for future students who attempt the problem and share the same misconception.



Student/Problem [Unanonymize]	Average Data driven	PRAHE5Y Data driven	PRAHE5Z Data driven	PRAHE52 Data driven
Problem Average	60%	27%	61%	84%
Common Wrong Answers		1/9^10,56% +feedback	1/5^13,58% +feedback	
Correct Answer(s)		*1/3^10*	*1/5^3*	*1/16^2*
XXXXX *	50%	✗ 1/9^10	✗ 1/5^13	✗ 1/162
XXXXX *	45%	✗ 1/9^10	✓ 1/5^3	✓ 1/16^2
XXXXX *	55%	✓ 1/3^10	✗ 1/5^13	✓ 1/16^2

Figure 1.3. An excerpt from an anonymized Item Report. Students are listed in the first column, followed by average performance and then specific performance on each question within the problem set. Teachers can see if the student answered correctly or incorrectly, the response given, whether a tutoring strategy was used, and common wrong answers as measured across the entire class. Common wrong answers are actionable; teachers and students can work together to provide a mistake message for future students.



The Assessment of Learning Infrastructure (ALI)

Completion Rates
Students that have started your study: 329
Students that have completed your study: 251

Bias Assessment
Before analyzing learning outcomes, we suggest first assessing potential bias introduced by your experimental conditions (i.e., examine differential attrition). The table below reports the number of students that have completed your study, split out by experimental condition.

Condition	Started (*n*)	Completed (*n*)	Completed (%)
Group A – Experiment 1	109	80	73.39
Group B – Experiment 2	87	60	68.97
Group C – Control	99	89	89.90
Total	*295*	*229*	*77.63*

NOTE: A significant difference was found between observed and expected completion rates across conditions, χ^2 (2, N = 295) = 13.467, p < .01. This means that a selection effect may have occurred. Hypothesis testing with regard to posttest scores has not been conducted out of an abundance of caution.

Mean and Standard Deviation of Posttest Score by Condition
To examine learning outcomes at posttest, an analysis of means was conducted across conditions. The table below reports mean posttest score and standard deviation for each condition. This information was sourced from our automated posttest sub-report.

	Completed (*n*)	Posttest Score*
Group A – Experiment 1	80	34.40 (4.34)
Group B – Experiment 2	60	32.95 (3.89)
Group C – Control	89	44.11 (3.72)
Total	*229*	37.15 (3.98)

* Presented as Mean (SD).

Raw Data Files

Raw data files contain the logged information for each student that has participated in your study. We provide this data in a variety of formats, as explained below, to assist in your analytic efforts. We use Google Docs to share these files with you. If you would like to process these files manually, we recommend downloading the CSV file of your choice and saving the file as an Excel spreadsheet or workbook to retain formatting and formulas. If you will be passing the file directly to a statistical package, downloading the CSV to a convenient location should suffice.

For a field glossary and tutorials on how to read each type of file, visit our Data Glossary.

Historical Data

Covariate File - A collection of useful covariates for the students participating in your study. This file includes student level variables (i.e., gender), class level variables, (i.e., homework completion rates), and school level variables (i.e., urbanicity). Click here for a tutorial on how to link this file to your experimental data.

Experimental Data

1. Action Level - One row per action per student; the finest granularity. Students participating in your study have performed 13,655 actions (e.g., beginning problems, attempting to answer problems, asking for tutoring, and eventually completing problems).
2. Problem Level - One row per problem per student. Students participating in your study have completed 2,280 problems. The flow through a single problem incorporates many actions, resulting in a coarser data file (fewer rows).
3. Student Level - One row per student; the coarsest granularity. Columns are laid out in opportunity order to depict the student's progression through the problem set. Problem level information is expanded to one column per problem per field (column heavy).
4. Student Level + Problem Level - One row per field per student. Columns are laid out in opportunity order to depict the student's progression through the problem set. An alternative view of student level information (row heavy).

Figure 1.4. The Assessment of Learning Infrastructure (ALI) provides researchers with logged data from students participating in RCEs within the ASSISTments TestBed (Ostrow, et al., In Press). This automated report is generated weekly, and/or at the request of the researcher, and presents analyses and raw data. Analyses include a Chi-squared test comparing the observed and expected sample distributions, simple hypothesis testing, and an analysis of means on posttest performance.

Through NSF funding (SI2-SSE&SSI: 1440753), reports for researchers have grown far more complex than teacher reports, providing numerous formats of raw performance data with rich student, class, and school level covariates, as well as a number of automated analyses. Through the ASSISTments TestBed, and specifically through the Assessment of Learning Infrastructure (ALI), researchers are provided weekly automated reports detailing anonymized study participation (Ostrow, et al., In Press). These reports, as shown in Figure 1.4, provide basic analyses including bias assessment (examining attrition across experimental conditions) and simple hypothesis testing on posttest performance. Researchers are also provided a student covariate file, detailing student information collected prior to study participation (i.e., prior performance average), and four formats of raw data logged by the ASSISTments tutor as students work through the assignment. ALI's reporting and researcher communications make the TestBed easier for researchers to use, streamlining research at scale.

1.1.3 Technology Supported Randomized Controlled Experimentation

Through the ASSISTments TestBed, researchers are able to design minimally invasive RCEs within easily accessible and highly used educational content delivered by ASSISTments, while receiving organized reports detailing student performance to streamline the analysis of learning interventions. This type of open research environment is rare within learning technologies. The common use for RCEs or A/B testing within popular technologies is to optimize user experience or prolong user interaction. For instance, Google experiments with advertisement location to maximize ad traffic without diminishing the user experience. Similarly, gaming application creators like Zynga conduct A/B testing to optimize their games in a way that will retain users while promoting ad space. Although these approaches are consistent in marketing, few large-scale education platforms show an outward interest in examining learning interactions and optimizing learning gains. Massive Open Online Course (MOOC) platforms and large-scale learning tools like Coursera, EdX, Udacity, openHPI, and Google's "Course Builder" focus on delivering content, while spending little time or money thoroughly examining the effects of what they deliver. This argument is not intended to suggest a complete deficit of sound research, but instead to point out that few researchers have access to course data from these platforms to improve user interfaces or curriculum delivery. Even commercialized educational technologies lack open and easily accessible avenues for empirical research. For instance, the popular Khan Academy provides resources and support for select researchers to work through a process requiring substantial time and effort to understand the dynamics of the system. Creating and running an experiment within Khan Academy requires knowledge of the platform's open-source code, the coding skills necessary to make modifications to implement experimentation, and progression through a standard code review process working alongside Khan Academy developers. Obtaining data files following an experiment is also heavily reliant on system programmers. To our knowledge, none of the A/B experiments that researchers *have* patiently conducted on Khan Academy have been formally published (e.g., see Williams, Paunesku, Haley & Sohl-Dickstein, 2013; Williams, 2014). Instead, work with less regard for improving specific interventions has evaluated the platform's efficacy in schools (Murphy, et al., 2014) and prediction models for large-scale but secondary data (Piech, et al., 2015). Such major platforms should be reframed with a focus on open educational research at scale or should at least support the open collection of anonymized data through APIs to inform EdTech policy.

The application of stringent research methodologies to improve learning technologies and educational outcomes is severely lagging. This deficit is what makes the ASSISTments TestBed so unique. The TestBed guides researchers through the process of running practical RCEs by leveraging ASSISTments' content and user population. There are currently over 130 RCEs running within the ASSISTments TestBed. These studies are directed at solving practical problems within education and understanding best practices within technology driven learning. While these studies help researchers to identify evidence-based instructional improvements, findings also lead to the generation of new hypotheses that expand investigation or reroute postulated theories. Results from a single study may generate four new hypotheses, with the potential for exponential expansion as a line of research evolves. The results of these studies can also benefit ASSISTments: findings regarding best practices continuously improve the system's content and delivery, while pinpointing areas for broad change through infrastructure improvements. Thus, a collaborative and open research infrastructure supports perpetual evolution on a small scale within the system and on a large scale across research communities.

1.2 Developing Collaboratives Around Shared Scientific Tools

To get the most out of educational technologies, learning platforms must be revolutionized into shared scientific instruments. Through the ASSISTments TestBed, ASSISTments is attempting to initiate this movement by stepping forward as the 'Hubble Telescope' of learning science. Unlike a static piece of equipment, the platform can be used to run multiple experiments simultaneously and researchers are able to improve the instrument for others through their experiences. Through this collaborative approach, as shown in Figure 1.5, researchers bring many ideas and hypotheses to the TestBed. Some of the studies designed around these hypotheses result in reliably positive

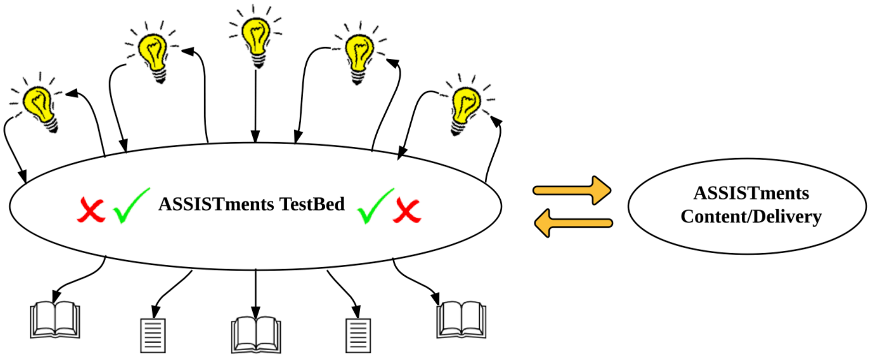


Figure 1.5. Research within the ASSISTments TestBed leads to knowledge of best practices, enhancements to student learning outcomes, and peer reviewed publications. Multiple iterations of hypotheses may arise, enhancing system content and strengthening content delivery as work progresses.

effects while others are extended to form stronger research questions. Through this process, researchers alter and enhance content and feedback within ASSISTments. Students and teachers benefit from stronger content while researchers expand their fields through refereed publications.

Realization of the platform's value as a shared scientific tool has encouraged research at scale from universities including Boston College, Freiburg University, Harvard University, Indiana University, Northwestern University, Southern Methodist University, Texas A&M, University of Colorado - Colorado Springs, University of California - Berkeley, University of Maine, University of Wisconsin, and Vanderbilt. Since its inception, interest in the TestBed has continued to expand through a kickoff webinar, an AERA seminar, and well documented support for researchers made possible by NSF funding (SI2-SSE&SSI: 1440753).

By articulating specific challenges for improving K-12 mathematics education to a broad and multidisciplinary community of psychology, education, and computer science researchers, this funding allows researchers to collaboratively (and perhaps competitively) propose and conduct RCEs, at an unprecedentedly precise level and large scale. The following list highlights the broad spectrum of work that researchers have shown interest in examining further within the TestBed:

Types of Feedback

- Immediate vs. delayed feedback (Fyfe, Rittle-Johnson, & DeCaro, 2012).
- Comparing the type of hints provided adaptively to learners (Stamper, et al., 2013).
- Comparing levels of feedback from guided to open (Sweller, Kirschner, & Clark, 2007).
- Comparing 'what you see is what you get' vs. interaction (Keehner, et al., 2008).
- Prompting for comparison of analogous problems and worked examples (Jee, et al., 2013).

Sequencing and Spacing

- Changing schedules and procedures for practice sessions & quizzes (Roediger & Karpicke, 2006).
- Testing the effectiveness of pre-testing prior to instruction (Richland, Kornell, & Kao, 2009).
- Spacing skill content (Pashler, et al., 2007).
- Examining testing effects (Butler & Roediger, 2007).

Self-Regulated Learning & Metacognition

- Testing interventions to increase motivation and teach strategies (Ehrlinger & Shain, 2014).
- Examining how task framing changes what students learn (Belenky & Nokes-Malach, 2013).
- Examining metacognitive scaffolding provided in problem solving (Roll, et al., 2012).
- Testing the value of free recall (Arnold & McDermott, 2013).

Social Context and Interaction

- Adapting instructional materials to students' personal & peer interests (Walkington, 2013).
- Embedding software & dynamics for peer assistance (Walker, Rummel, & Koedinger, 2011).

- Examining how confidence impacts performance in early algebra (Mazzocco, et al., 2013).

Assessment

- Examining computational models used to diagnose learner state (Rafferty & Griffiths, 2014).
- Examining computational methods for assessing affective states (Ocumpagh, et al., 2014).
- Examining forgetting (Storm, et al., 2006).

Motivation

- Embedding motivational videos from teachers (Kelly, Heffernan, D'Mello, et al, 2013).
- Incorporating messages to foster growth mindset (Williams, 2013).
- Examining the effects of goal setting (Bernacki, Byrnes, & Cromley, 2012).
- Examining the effects of student choice (Chernyak & Kushnir, 2013).
- Inserting quizzes and tests to maintain and guide student focus (Szpunar, et al., 2013).

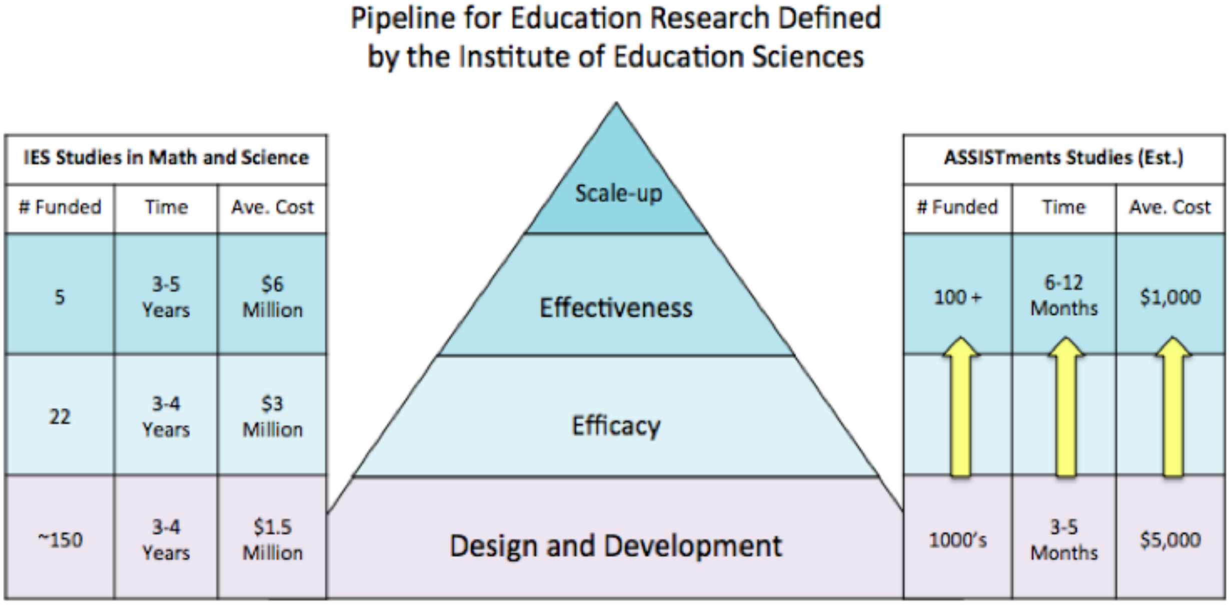
Mathematics Education

- Comparing representational formats in supporting mathematics learning (Rau, et al., 2012).
- Investigating effective presentations of worked examples in mathematics (Booth, et al., 2013).
- Examining strategies for learning fractions (Cordes, et al., 2007).
- Testing images of manipulatives vs. virtual manipulatives (Mendiburo, et al., 2012).

By building these types of collaborative scientific tools, the cost of funding educational research could be drastically reduced. For instance, the Institute of Education Sciences (IES) currently funds "Efficacy Trials" for promising interventions that cost an average of $3M and can involve more than 50 schools. Larger and more stringent "Effectiveness Trials" carry a median cost of $6M. In the math and science domains, the IES has funded 22 Efficacy Trials and five Effectiveness Trials. Despite the high cost of funding this work, reliable positive implications for educational practice are rarely observed. Using adaptive technologies geared toward research, large-scale trials could be expedited at a fraction of the cost. The IES funding pipeline (IES, 2015) and the ASSISTments TestBed equivalent are depicted in Figure 1.6. Studies that were once restricted by the availability of funding could be considered through learning technologies.

Much of the efficacy attained through use of the TestBed is due to student-level randomization (rather than traditional class- or school-level randomization), allowing experiments to be conducted within classrooms rather than across classrooms. This accrues drastically larger samples, increasing the power of analyses in order to better detect the reliable effects of interventions. The unique ability for student-level randomization, coupled with the scalability inherent to manipulating pre-built content of interest to a large user base, allows in vivo educational research to gain the minimally invasive A/B flavor often used in marketing. Studies within the TestBed also align with typical educational practice (i.e., students are never intentionally disadvantaged by a study design). This approach allows students to access and complete assignments, often without awareness that they are participating in research. Teachers are made aware of experimentation through a conventional assignment naming procedure that 'tags' experiments with 'Ex.' As data dissemination is carefully preprocessed to protect students' identifies, and students receive assignments that are within the definition of 'normal instructional practice,' this passive approach to research is IRB approved.

While low cost procedures may not hold for all educational investigations (i.e., the design of full learning programs or platforms that require significant funding), there are many benefits to cost-effective, efficient, and rigorous experimentation that can be conducted using educational technologies. Many unique features make ASSISTments capable of serving researchers as a shared scientific tool. However, ASSISTments is not the only platform with the power to drive a collaborative like the TestBed. The majority of learning applications have the capacity for data collection, and many could be restructured to offer the flexibility required for experimental content manipulation. Other platforms may also be capable of establishing an API to deliver preprocessed data, anonymized for student protection, to researchers conducting RCEs or even wishing to mine data. With similar research-based platforms in the field, it would also be possible for researchers to compare learning interventions across platforms to better measure the reliability and generalizability of results. Collaborative research goals that crosscut platforms may finally usher in the tipping point of educational technologies (Bush & Mott, 2009; Gladwell, 2002) as researchers grow to understand "What works best? For Whom? When?"


Pipeline for Education Research Defined by the Institute of Education Sciences
Scale-up
Effectiveness
Efficacy
Design and Development

IES Studies in Math and Science		
# Funded	Time	Ave. Cost
5	3-5 Years	$6 Million
22	3-4 Years	$3 Million
~150	3-4 Years	$1.5 Million

ASSISTments Studies (Est.)		
# Funded	Time	Ave. Cost
100 +	6-12 Months	$1,000
1000's	3-5 Months	$5,000

Figure 1.6. The pipeline for education research as defined by the IES compared to a similar timeline for research within the ASSISTments TestBed. Educational technologies can be used as shared scientific tools to drastically reduce costs and enhance the efficiency with which educational research is conducted.

1.3 Collaborative Research at Scale Offers Perpetual Benefits

The power of the ASSISTments TestBed as a collaborative research tool did not come about overnight. As a learning platform, ASSISTments has pivoted numerous times in the past decade (Heffernan & Heffernan, 2014). The steady improvements from which the TestBed were largely driven by the results of pilot studies within the system. This growth and adaptation exemplifies perpetual evolution. Essentially, a simple hypothesis acts as the seed for an expanse of research that germinates through related ideas, eventually pushing the limits of the system until infrastructure improvements must be made to accommodate further questions - a cycle depicted in Figure 1.7. As the cycle begins, researchers form novel hypotheses that compare manipulations within the platform to best (known) practices (either comparable traditional classroom practices or previous versions of the platform's material). Early results inspire collaborative idea expansion through replications and extensions of studies that serve to enhance system content and content delivery, while improving student learning and advancing the state of knowledge in the field through peer reviewed publication. New hypotheses form and grow as results are observed, naturally evolving until pushing the boundaries of the platform's infrastructure. In response, scientifically validated infrastructure improvements can be tailored to research demand, forming the final stage of this cycle. New system features, a mark of evolution, allow researchers to start the cycle anew with novel hypotheses.

Ever expanding progress is a core concept for effectively marketing commercial products, but is far less common in education. Education is a difficult rock to move, with teachers and administrators holding tight to traditional methods, and pushing back against the changes brought about by modern technologies (Bush & Mott, 2009). It is hardly surprising that most educational technologies lack collaborative research infrastructures. Administrators have not been focused on examining the effectiveness of new instructional strategies made possible by these platforms because most platforms have instead been tailored to simplify traditional teaching methods (Bush & Mott, 2009). As educators continue to grow more open to the possibilities of learning technologies, the value of collaborative research at scale will escalate. By establishing research environments like the TestBed, creators and users of educational technologies will learn of the unprecedented benefits made possible by the cycle of perpetual evolution. The following sections step through this cycle, defining exemplary research at each stage, as conducted within ASSISTments and the ASSISTments TestBed.

1.3.1 The Seed: Comparing Research Generated Content To Best (Known) Practices

In *Estimating the Effect of Web-Based Homework*, Kelly, Heffernan, Heffernan, et al. (2013) used ASSISTments to compare traditional mathematics homework (with delayed, next day feedback) to the same assignment featuring immediate correctness feedback. All students participating in this RCE used ASSISTments to complete their homework, with feedback settings differing

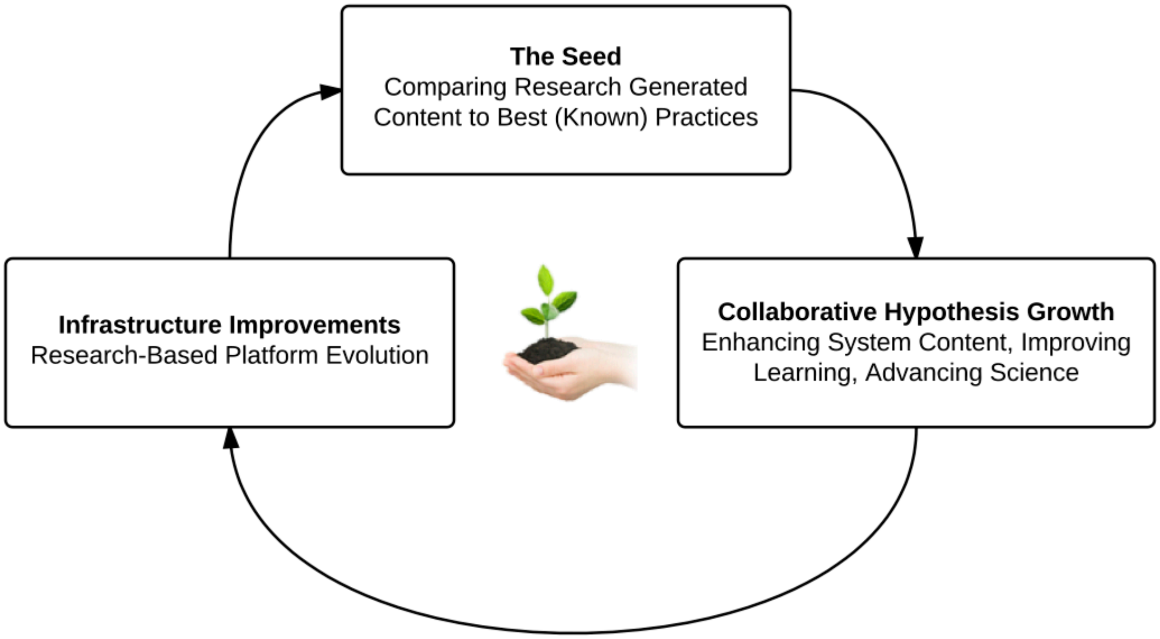
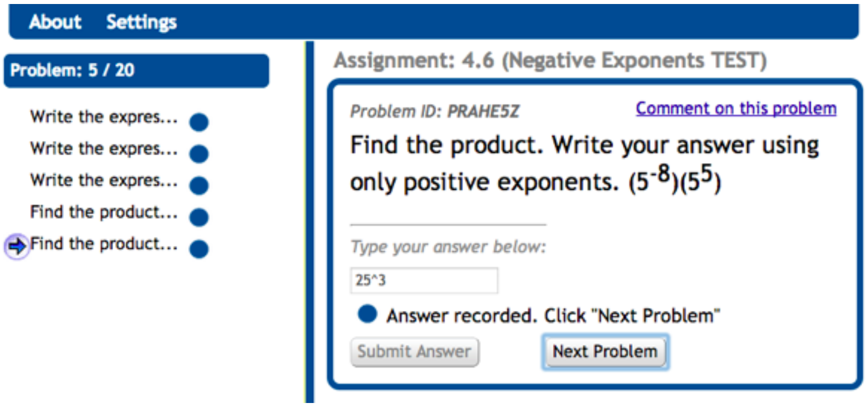

The Seed
Comparing Research Generated Content to Best (Known) Practices
Collaborative Hypothesis Growth
Enhancing System Content, Improving Learning, Advancing Science
Infrastructure Improvements
Research-Based Platform Evolution

Figure 1.7. The cycle of perpetual evolution that stems from use of an educational platform as a collaborative research tool. An initial hypothesis comparing new methods to best known practices grows into a series of ideas that improve system content while benefiting students and advancing knowledge in the field. These ideas continue to grow until limited by the platform's capabilities. Infrastructure improvements validated by previous findings and inspired by research demand can then be made to return the cycle to a fresh starting point, where new hypotheses can be formed.

between randomly assigned conditions. The research design included 20 questions delivered using skill triplets (i.e., three similar skill problems presented consecutively) to determine the effectiveness of correctness feedback. Students in the control condition did not receive feedback while completing their homework, as shown in Figure 1.8. Blue dots within the left menu show completed problems. The next day in class, the teacher reviewed the homework without using ASSISTments reports, and simply read answers aloud as students corrected their work. The teacher then worked through requested problems on the board. Students in the experimental condition received immediate correctness feedback while completing their homework, as shown in Figure 1.9. The next day in class, the teacher used data from the Item Report to determine which problems to focus on during the homework review, with an emphasis on common wrong answers shared by multiple students.

Analysis of 63 students suggested reliable improvements in student learning through the addition of correctness feedback. Students in the control group showed an average gain of 59% from pretest to posttest (an effect size of 0.52), while students in the experimental group showed an average gain of 74% (an effect size of 0.56). It should be noted that Cohen's "rule of thumb" for interpreting effect sizes has been somewhat discredited as a measure for benchmarking the



Assignment: 4.6 (Negative Exponents TEST)

Problem ID: PRAHE5Z

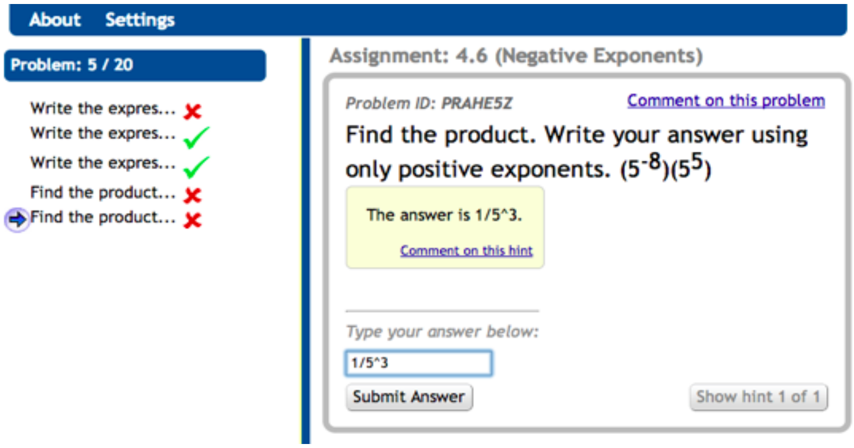
Find the product. Write your answer using only positive exponents. $(5^{-8})(5^{5})$

Type your answer below:

25^3

● Answer recorded. Click "Next Problem"

Figure 1.8. The control condition as experienced by the student (Kelly, Heffernan, Heffernan, et al., 2013). Students were not told if their answers were correct or incorrect. This approach mirrors traditional homework. This study implemented problem triplets, or sets of three questions per skill, providing multiple opportunities to display skill knowledge.



Assignment: 4.6 (Negative Exponents)

Problem ID: PRAHE5Z

Find the product. Write your answer using only positive exponents. $(5^{-8})(5^{5})$

The answer is 1/5^3.

Type your answer below:

1/5^3

Figure 1.9. The experimental condition as experienced by the student (Kelly, Heffernan, Heffernan, et al., 2013). Students were provided immediate correctness feedback as they responded to each problem. The student in this example was able to self-correct and progress through the first skill triplet but struggled with the second.

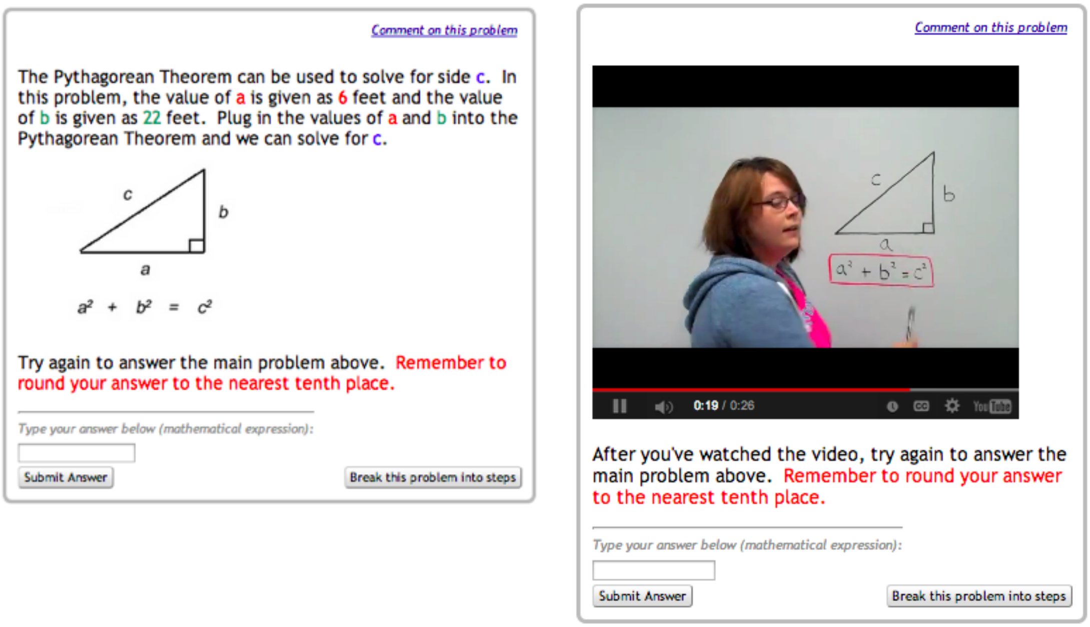
practical significance of effects, especially when working with researcher defined measures (Lipsey, et al., 2012). Instead, it is recommended that researchers compare growth attributed to an intervention to normative expectations. Comparing gains across conditions, this method suggests a reliable 15% increase in average learning gains. It is also possible to benchmark these findings against the results of similar studies, which have a mean effect size of 0.43 (Lipsey, et al., 2012), showing the clear strength of providing immediate correctness feedback as an intervention. Kehrer, Kelly & Heffernan (2013) replicated the positive effects of immediate correctness feedback observed in Kelly, Heffernan, Heffernan, et al.'s original work (2013).

Similar hypotheses examining the efficacy of feedback within ASSISTments have led to numerous publications over the past decade. Mendicino, Razzaq & Heffernan, (2009) examined the effectiveness of mathematics homework with scaffolded tutoring in comparison to traditional paper and pencil homework. Students that received adaptive scaffolding showed significant learning gains over those following traditional homework procedures. Razzaq, Heffernan, & Lindeman (2007) suggested that adaptive scaffolding led to greater learning gains than on demand hints. Researchers observed an interaction between students' proficiency levels and the effectiveness of feedback styles, with less proficient students benefiting from scaffolding and more proficient students benefiting from hints. Follow-up studies confirmed that on demand hints produced more reliable and robust learning in highly proficient students (Razzaq & Heffernan, 2010). Singh et al. (2011) then compared correctness feedback with on demand hints. Multiple trials consistently revealed that hint feedback led to significantly improved learning over correctness feedback alone. Research has also examined the content presented *within* feedback, through comparisons of worked examples and scaffolded problem solving (Shrestha, et al., 2009; Kim, et al., 2009) and investigations of motivational feedback (Kelly, Heffernan, D'Mello, et. al., 2013; Ostrow, Schultz, & Arroyo, 2014). Results suggesting the consistent benefits of feedback have allowed researchers working within ASSISTments to expand their questions from a seed - "Does immediate feedback help?" - to more detailed investigations - "What type of immediate feedback is most effective?"

1.3.2 Collaborative Hypothesis Growth: Enhancing System Content, Improving Learning, Advancing Science

In *Testing the Multimedia Principle in the Real World: A Comparison of Video vs. Text Feedback in Authentic Middle School Math*, Ostrow & Heffernan expanded on "feedback is good" hypotheses to examine the effectiveness of various feedback mediums (2014). Prior to this study, ASSISTments delivered feedback via text, altering color and typeface to draw students' attention to significant variables and themes. This RCE pushed that boundary to compare learning outcomes when identical feedback was delivered using short video snippets. Outcomes of student performance and response time were measured across six problems pertaining to the Pythagorean theorem. All students received the same six questions in mixed orders, receiving three opportunities for text feedback and three opportunities for video feedback over the course of the assignment. As shown in Figure 1.10, feedback was matched across medium; videos were comprised of a researcher working through each feedback step while referencing images on a white board. Students received feedback through scaffolds, by requesting assistance or answering a problem incorrectly. Learning gains were compared on the second question, across students who received feedback on the first question. Following the problem set, students were asked a series of survey questions to judge how they viewed the addition of video to their assignment.

Results of an analysis of 89 students that completed the assignment and were able to access video content revealed that video feedback increased the likelihood of accuracy on the next problem. Students spent significantly longer consuming video feedback but answered their next question more efficiently. Assessing self-report measures, 86% of students found the videos at



Comment on this problem

The Pythagorean Theorem can be used to solve for side c. In this problem, the value of a is given as 6 feet and the value of b is given as 22 feet. Plug in the values of a and b into the Pythagorean Theorem and we can solve for c.

$a^2 + b^2 = c^2$

Try again to answer the main problem above. Remember to round your answer to the nearest tenth place.

Type your answer below (mathematical expression):

Comment on this problem

After you've watched the video, try again to answer the main problem above. Remember to round your answer to the nearest tenth place.

Type your answer below (mathematical expression):

Figure 1.10. A comparison of text and video feedback conditions as experienced by students (Ostrow & Heffernan, 2014). Isomorphic problems featured matched content feedback across mediums.

least somewhat helpful and 83% of students wanted video in future assignments (Ostrow & Heffernan, 2014). Based on these findings, teachers and researchers have been recruited to create video feedback for Skill Builder problems to expand the amount of video content available within the system and allow for further examination into the effects of video. The ease with which teachers and researchers are able to record short video messages and upload them to the system suggests that this approach is a plausible avenue for crowdsourcing feedback (Kittur, et al., 2013; Howe 2008). Crowdsourcing and learnersourcing (Kim, 2015) feedback are future directions for the ASSISTments platform, as infrastructure improvements are required to optimally support, organize, and vet feedback collection at scale.

Many of the studies that best define 'collaborative hypothesis growth' are currently underway within the TestBed, examining the effectiveness of particular types of feedback. Numerous researchers are investigating what drives the apparent effects of video feedback by comparing various types of videos (e.g., recorded human tutoring, a "pencast" problem walkthrough with audio explanation, and peer videos with tutoring led by other students). Many of these studies are pushing ASSISTments' technological boundaries, establishing a demand for specific infrastructure improvements that will help the system and its content to evolve.

1.3.3 Infrastructure Improvements: Research-Based Platform Evolution

Research on the efficacy of feedback mediums laid the groundwork for debates about the possible impacts of allowing students to choose between mediums. Without any real capacity to provide choice, ASSISTments was reaching a tipping point for infrastructure improvement. A pilot study was conducted by taking advantage of bugs in the system to ‘mock up’ student choice (Ostrow & Heffernan, 2015). This simple RCE examined interactions between student choice and feedback medium using a 2x2 factorial design, depicted in Figure 1.11. Two versions of a problem set on simple fraction multiplication were created, one incorporating text feedback and one incorporating video feedback. Short, 15-30 second video snippets were designed to be as comparable to text feedback as possible, in order to compare delivery medium. At the start of the assignment, students were randomly assigned to either the choice condition or the control condition. Those assigned to the choice condition were asked what type of feedback they wished to receive while working on their assignment, as shown in Figure 1.12, and were routed accordingly. Those assigned to the control were immediately re-assigned to either video or text feedback.

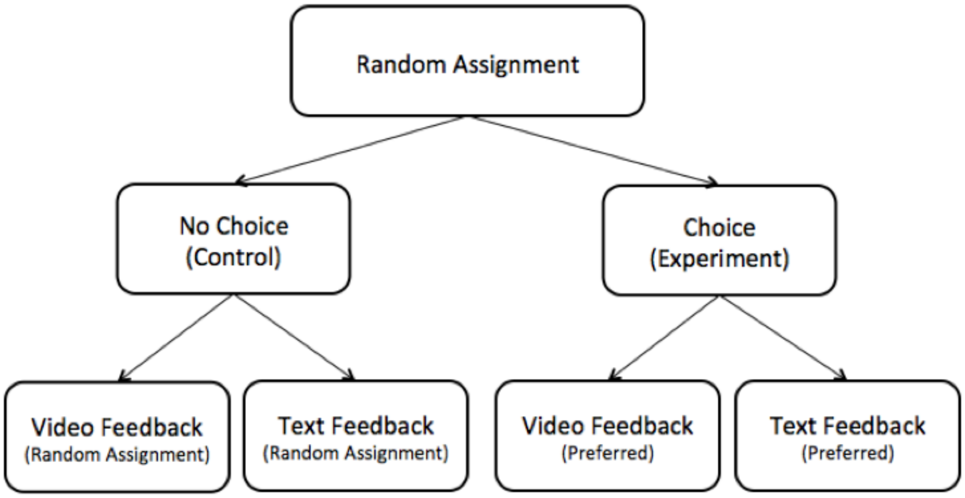
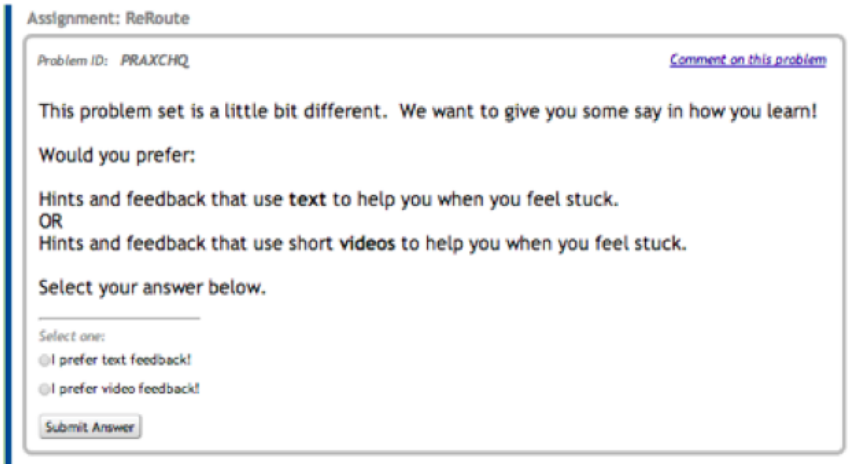

Random Assignment
No Choice (Control)
Choice (Experiment)
Video Feedback (Random Assignment)
Text Feedback (Random Assignment)
Video Feedback (Preferred)
Text Feedback (Preferred)

Figure 1.11. Experimental design used to investigate student choice as a pilot study within ASSISTments (Ostrow & Heffernan, 2015). Prior to this study, students were not able to exert control over their assignments within the platform.



Assignment: ReRoute

Problem ID: PRAXCHQ

Comment on this problem

This problem set is a little bit different. We want to give you some say in how you learn!

Would you prefer:

Hints and feedback that use **text** to help you when you feel stuck.
OR
Hints and feedback that use short **videos** to help you when you feel stuck.

Select your answer below.

Select one:

○ I prefer text feedback!

○ I prefer video feedback!

Figure 1.12. Student preference prompt guiding medium routing for those in the experimental condition (Ostrow & Heffernan, 2015).

Considering a sample of 78 middle school students that completed this pilot, results suggested that feedback medium did not have a specific impact on learning gains within this context (contrary to results presented earlier on the efficacy of video feedback, suggesting that perhaps video is not effective for all age ranges or skill domains and beginning to answer "What works best? For whom? When?"). However, students that were able to choose their feedback medium showed significant improvements over students that were randomly assigned a medium. Students with choice earned higher scores on average, used fewer hints and attempts, and persisted longer than those not provided choice. Perhaps the most interesting observation: learning gains were higher in students that were provided choice, regardless of whether or not the student actually ended up requesting feedback during the assignment (Ostrow & Heffernan, 2015). These results became the driving force for a significant infrastructure improvement within the ASSISTments platform that would allow for conditional path routing. An If-Then routing structure was developed under the SI2 NSF grant (SI2-SSE&SSI: 1440753) to extend research capabilities within ASSISTments and the ASSISTments TestBed. Hypotheses regarding student choice, and other routing-based research, can now be easily examined with greater validity and at scale.

A replication of the choice pilot by Ostrow & Heffernan (2015) was designed using the If-Then routing structure, as shown in Figure 1.13. The inclusion of conditional path routing helped to enhance the internal validity of video-based research by allowing sample populations to be refined to only include students with the technological capacity to view video content. While in hindsight this feature seems like an obvious requirement for video-based research, it was not possible within ASSISTments prior to If-Then routing. Thus, it is clear how this new feature has the potential to improve and expand research within the ASSISTments TestBed.

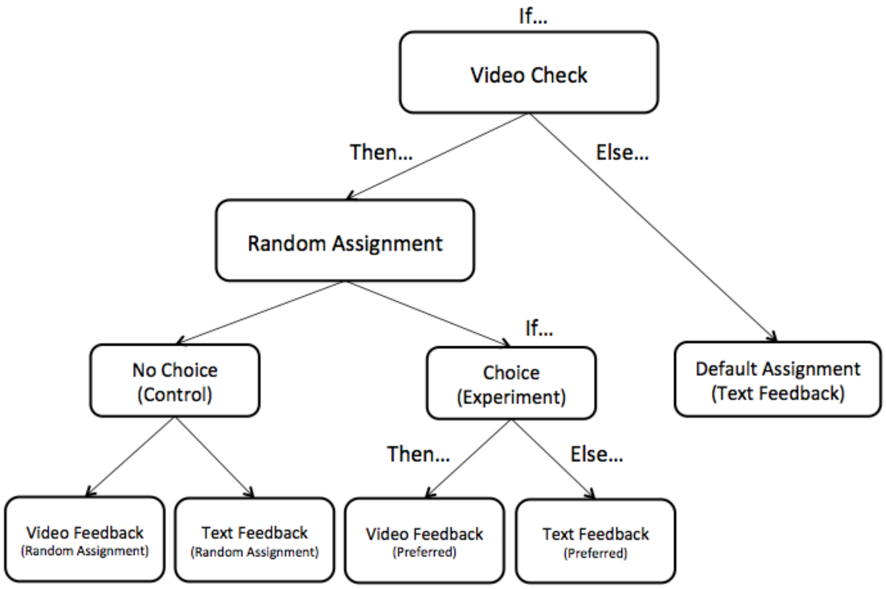
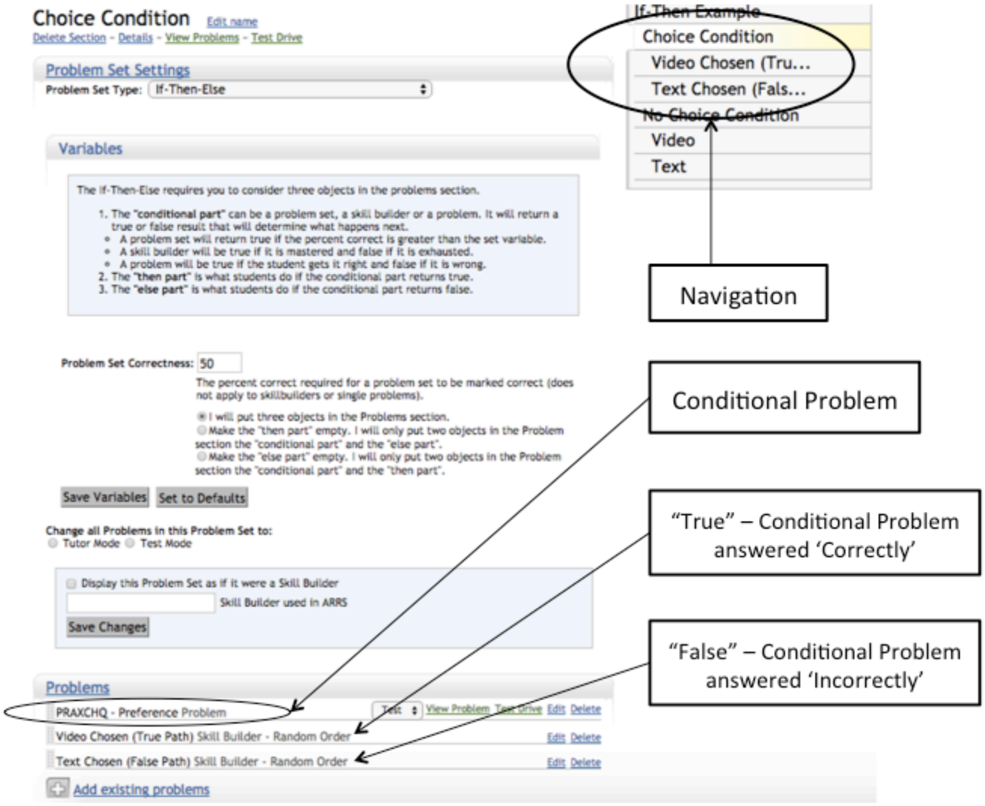

If...
Video Check
Then...
Else...
Random Assignment
If...
No Choice (Control)
Choice (Experiment)
Default Assignment (Text Feedback)
Then...
Else...
Video Feedback (Random Assignment)
Text Feedback (Random Assignment)
Video Feedback (Preferred)
Text Feedback (Preferred)

Figure 1.13. Updated choice design replicating Ostrow & Heffernan (2015) with an If-Then routing structure for greater internal validity. The initial If-Then statement assesses students' technological capacity for viewing video content, while the second If-Then controls routing in the choice condition.

An example of how a researcher might go about building an ASSISTments problem set with simple If-Then routing is shown in Figure 1.14. The building process requires three elements: a conditional statement, a true path, and a false path (Ostrow & Heffernan, 2016). The conditional statement can include a problem or problem set, with an adjustable setting that guides path routing based on student performance as measured by completion or accuracy. If performance meets this preset threshold, the student is routed into the true path, or the second section in Figure 1.14 ("Video Chosen"). If performance does not meet this preset threshold, the student is routed into the false path, or the third section in Figure 1.14 ("Text Chosen"). In this example, the conditional statement is a single preference question, much like that shown in Figure 1.12. Video feedback is set as the 'correct' answer, routing students based on the "Then" clause, while text feedback is set as the 'incorrect' answer, routing students based on the "Else" clause. Students receive this problem in test mode (i.e., without correctness feedback, showing only a blue dot for completion) therefore restricting the inner workings of the routing system from student view and removing the risk of undue penalization for a 'wrong opinion.' Numerous studies now running within the ASSISTments TestBed implement If-Then routing in some capacity (e.g., as technical validation, as adaptive performance routing, to trigger interventions for struggling students, or to buffer sampling within intent-to-treat studies seeking to help only students with low skill proficiency). This simple infrastructure improvement completes an iteration of the cycle of perpetual evolution, opening new avenues for fresh 'seed-level' hypotheses to start the cycle anew.



Choice Condition Edit name

Delete Section - Details - View Problems - Test Drive

Problem Set Settings

Problem Set Type: If-Then-Else

Variables

The If-Then-Else requires you to consider three objects in the problems section.

1. The **"conditional part"** can be a problem set, a skill builder or a problem. It will return a true or false result that will determine what happens next.
   - A problem set will return true if the percent correct is greater than the set variable.
   - A skill builder will be true if it is mastered and false if it is exhausted.
   - A problem will be true if the student gets it right and false if it is wrong.
2. The **"then part"** is what students do if the conditional part returns true.
3. The **"else part"** is what students do if the conditional part returns false.

Problem Set Correctness: 50

The percent correct required for a problem set to be marked correct (does not apply to skillbuilders or single problems).

● I will put three objects in the Problems section.
○ Make the "then part" empty. I will only put two objects in the Problem section the "conditional part" and the "else part".
○ Make the "else part" empty. I will only put two objects in the Problem section the "conditional part" and the "then part".

Save Variables | Set to Defaults

Change all Problems in this Problem Set to:
○ Tutor Mode ○ Test Mode

□ Display this Problem Set as if it were a Skill Builder

Skill Builder used in ARRS

Save Changes

Problems

PRAXCHQ - Preference Problem | Test | View Problem Test Drive Edit Delete

Video Chosen (True Path) Skill Builder - Random Order | Edit Delete

Text Chosen (False Path) Skill Builder - Random Order | Edit Delete

Add existing problems

If-Then Example
Choice Condition
Video Chosen (Tru...
Text Chosen (Fals...
No Choice Condition
Video
Text

Navigation
Conditional Problem
"True" – Conditional Problem answered 'Correctly'
"False" – Conditional Problem answered 'Incorrectly'

Figure 1.14. The researcher's view while constructing a study using If-Then routing within an ASSISTments problem set. The study design shown here mirrors that in Figure 1.13.

1.4 Future Directions of the ASSISTments Platform

It is difficult to advocate for a future consisting of research infused educational technologies without touching briefly on future goals for the ASSISTments platform. With a focus on disseminating the ASSISTments TestBed and enhancing its validity as a collaborative tool for sound science, the cycle of perpetual evolution will bring about a number of significant infrastructure improvements for ASSISTments in the near future. Perhaps the most immediate change, as suggested by the research presented herein, will be extending the platform to support teachersourced and learnersourced feedback. The platform has 25,000 vetted mathematics problems that were created by Worcester Polytechnic Institute and Carnegie Mellon University. In addition, teachers have added over 100,000 problems to the platform (many that already include some form of feedback). The first step toward crowdsourcing feedback for these problems is to allow teachers to create tutoring strategies in support of content owned by others (rather than only in support of their own content). Differing teachers will offer differing solution approaches, which may help struggling students to see a problem from a different perspective. A select group of teachers and students have already recorded video feedback for use in a set of RCEs examining the potential benefits and obstacles of crowdsourcing feedback at scale. Eventually, this approach will be scaled to allow students to 'show their work' and provide explanations for their peers through a tool called PeerASSIST (Heffernan, Ostrow, et al., 2016). A task already appreciated by most mathematics teachers, showing work will help students to solidify their understanding of the content while creating feedback to benefit other users (Kulkarni, et al., 2013). The network effects inherent to teachersourcing and learnersourcing feedback will enhance system content at an impressive scale (Bush & Mott, 2009).

The implementation of crowdsourcing will naturally give way to another goal for the future of ASSISTments: establishing an automated process to select optimal feedback using contextual k-armed bandits. This approach, rooted in the theory of sequential design (Robbins, 1952), is an algorithmic approach to the exploration/exploitation trade-off. Essentially, with a pool of content available to students (i.e., many types of feedback), it is necessary to repeatedly sample the efficacy of assigned content in order to maximize the delivery of effective content while minimizing the delivery of ineffective content. The use of k-armed bandits will minimize detriment to students while allowing for the dynamic versioning of materials and setting the stage for personalized learning (i.e., algorithmically establishing "What works best? For whom? When?"). An important feature that will grow from the implementation of k-armed bandits will be the capacity to store user variables for lasting personalization. Variables like initial performance, particular student responses, or specific student characteristics could help to optimize content and feedback delivery for each student, both within and across assignments. The ASSISTments team expects that these goals will strengthen the platform and inspire new avenues for scientific inquiry.

1.5 Infuse Educational Technologies with Collaborative Research to Promote Sound Science

Systemic change does not stem from a small number of large-scale RCEs funded by government grants, but instead from a revolution in thought surrounding the value of technology-based learning applications. As shown herein, infusing pre-existing learning technologies with the capability to support RCEs is the first step in kick starting this revolution. From there, the platform can expand as a shared scientific tool utilized by a community of researchers collaborating to better understand the efficacy of educational interventions. ASSISTments bridges practice and research by enabling researchers to work collaboratively with teachers and students, and by providing unprecedented access to authentic learning environments and actionable classroom data. The collaborative nature of the ASSISTments TestBed gives way to a cycle of perpetual evolution that inspires continuous advancements to ASSISTments content while simultaneously advancing knowledge of best practices. Insights and innovations drawn from research findings can be incorporated into the system itself as well as future research, with each successive step building upon previous contributions.

Research infused platforms have the potential to drive inquiry for a diverse community of researchers through the low-cost, rapid iteration of valid, generalizable, and noninvasive investigations within authentic learning environments. Systems like ASSISTments can provide researchers with access to an extensive and diverse subject pool, an automated fine-grained logging of educational data, validated measures of student learning and affect, and automated data reporting and analysis to tackle the high stakes nature of typical education research. With similar research-focused platforms in the field, it would also be possible for researchers to compare learning interventions across platforms to better measure the reliability and generalizability of results. These platforms offer a unique opportunity for the synergistic growth of research and policy detailing best practices in education. If these platforms grow to welcome collaborative research, educational technology will reach its long-awaited tipping point and begin to broadly impact the efficacy and validity of research across domains. Tomorrow's educational technology demands a revolution in today's approaches to research at scale: pave the way for sound collaborative science and the rest will follow.

References

Achenbach, J. (2015). Many scientific studies can't be replicated: that's a problem. The Washington Post, Speaking of Science. August 27, 2015. Retrieved 2/19/16 from https://www.washingtonpost.com/news/speaking-of-science/wp/2015/08/27/trouble-in-science-massive-effort-to-reproduce-100-experimental-results-succeeds-only-36-times/

Adjei, S.A. & Heffernan, N.T. (2015). Improving learning maps using an adaptive testing system: PLACEments. In Conati, Heffernan, Mitrovic, & Verdejo (eds), Proceedings of the 17th International Conference on Artificial Intelligence in Education. Springer International Publishing. 9112: 517-520.

Arnold, K.M. & McDermott, K.B. (2013). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*. 20 (3): 507-513.

Belenky, D.M. & Nokes-Malach, T.J. (2013). Mastery-approach goals and knowledge transfer: An investigation into the effects of task structure and framing instructions. *Learning and Individual Differences*. 25: 21-34.

Bernacki, M.L., Byrnes, J.P. & Cromley, J.G. (2012). The effects of achievement goals and self-regulated learning behaviors on reading comprehension in technology-enhanced learning environments. *Contemporary Educational Psychology*. 37 (2): 148-161.

Booth, J.L., Lange, K.E., Koedinger, K.R. & Newton, K.J. (2013). Using example problems to improve student learning in algebra: Differentiating between correct and incorrect examples. *Learning and Instruction*. 25: 24-34.

Bush, M.D. & Mott, J.D. (2009). The transformation of learning with technology: Learner-centricity, content and tool malleability, and network effects. *Educational Technology. 4*9 (2): 3-20.

Butler, A.C., & Roediger, H.L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*. 19: 514-527.

Chernyak, N., & Kushnir, T. (2013). Giving preschoolers choice increases sharing behavior. *Psychological Science*. 24 (10): 1971-1979.

Cordes, S., Williams, C.L., & Meck, W.H. (2007). Common representations of abstract quantities. *Current Directions in Psychological Science*. 16 (3): 156-161.

Ehrlinger, J. & Shain, E.A. (2014). How accuracy in students' self perceptions relates to success in learning. In. Benassi, Overson, & Hakala (eds.) Applying science of learning in education: Infusing psychological science into the curriculum. Society for the Teaching of Psychology. Retrieved from http://teachpsych.org/ebooks/asle2014/index.php.

Fyfe, E.R., Rittle-Johnson, B. & DeCaro, M.S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology*. 104: 1094-1108.

Gladwell, M. (2002). The tipping point: How little things can make a big difference. Boston: Back Bay.

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. *International Journal of Artificial Intelligence in Education*. 24 (4): 470-497.

Heffernan, N.T., Ostrow, K.S., Kelly, K., Selent, D., Van Inwegen, E.G., Xiong, X., & Williams, J.J. (2016). The future of adaptive learning: Does the crowd hold the key? *International Journal of Artificial Intelligence in Education*. Springer New York. DOI: 10.1007/s40593-016-0094-z

Howe, J. (2008). Crowdsourcing: Why the power of the crowd is driving the future of business. New York, NY: Crown Business.

Institute of Education Sciences. (2013). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. U.S. Department of Education. Retrieved 2/18/16 from http://ies.ed.gov/ncee/pubs/evidence_based/randomized.asp.

Institute of Education Sciences. (2015). Request for applications: Education research grants, CFDA Number: 84.305A. U.S. Department of Education. Retrieved 2/19/16 from https://ies.ed.gov/funding/pdf/2016_84305A.pdf.

Ioannidis J.P.A. (2005). Why most published research findings are false. *PLoS Med.* 2 (8): e124. Retrieved on 2/18/16 from http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124.

Jee, B., Uttal, D., Gentner, D., Manduca, C., Shipley, T. & Sageman, B. (2013). Finding faults: analogical comparison supports spatial concept learning in geoscience. *Cognitive Processing*. 14 (2): 175-187.

Keehner, M., Hegarty, M., Cohen, C., Khooshabeh, P. & Montello, D.R. (2008). Spatial reasoning with external visualizations: what matters is what you see, not whether you interact. *Cognitive Science*. 32: 1099-1132.

Kehrer, P., Kelly, K. & Heffernan, N. (2013). Does immediate feedback while doing homework improve learning. In Boonthum-Denecke & Youngblood (eds) Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference. AAAI Press. 542-545.

Kelly, K., Heffernan, N., D'Mello, S., Namias, J., & Strain, A. (2013). Adding teacher-created motivational video to an ITS. In Boonthum-Denecke & Youngblood (eds) Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference. AAAI Press. 503-508.

Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, G. & Soffer, D. (2013). Estimating the effect of web-based homework. In Lane, Yacef, Mostow & Pavlik (eds) Proceedings of the 16th International Conference on Artificial Intelligence in Education. Springer Berlin Heidelberg. 7926: 824-827.

Kim, J. (2015). Learnersourcing: Improving learning with collective learner activity. MIT PhD Thesis. Retrieved 2/19/16 from http://juhokim.com/files/JuhoKim-Thesis.pdf.

Kim, R, Weitz, R., Heffernan, N. & Krach, N. (2009). Tutored problem solving vs. “pure” worked examples. In Taatgen & van Rijn (eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society. 3121-3126.

Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. (2013). The future of crowd work. In Proceedings of the 2013 Conference on Computer Supported Cooperative Work. ACM, New York. 1301-1318.

Koedinger, K.R., Booth, J.L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*. 342: 935-937.

Koedinger, K., McLaughlin, E. & Heffernan, N. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research*. Baywood Publishing. 4: 489-510.

Koedinger, K.R., Baker, R.S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In Romero, Ventura, Pechenizkiy, & Baker (eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

Kulkarni, C., Wei, K.P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., & Klemmer, S.R. (2013). Peer and self assessment in massive online classes. In ACM Transaction on Computer-Human Interaction (TOCHI). New York: ACM. 20 (6): 33.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., & Busick, M.D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Mazzocco, M.M.M., Murphy, M.M., Brown, E.C., Rinne, L. & Herold, K.H. (2013). Persistent consequences of atypical early number concepts. *Frontiers in Psychology*. 4: 486.

Mendiburo, M., Sulcer, B., Biswas, G., Hasselbring, T.S. (2012). Interactive virtual representations, fractions, and formative feedback. In Cerri, Clancey, Papadourakis, & Panourgia (eds.) Proceedings of the 11th International Conference on Intelligent Tutoring Systems. Springer Berlin Heidelberg. 7315: 716-717.

Mendicino, M., Razzaq, L. & Heffernan, N.T. (2009). Improving learning from homework using intelligent tutoring systems. *Journal of Research on Technology in Education*. 41 (3): 331-346.

Miller, G., Zheng, Y., Means, B., & Van Brunt Miller, B. (2013). Next generation learning challenges wave II. Retrieved on 2/19/16 from https://docs.google.com/file/d/0B2X0QD6q79ZJUl9Kd2JuVTN0VWhTYVRhX254QV85Njdqc1Vj/edit?pli=1.

Murphy, R., Gallagher, L., Krumm, A.E., Mislevy, J. & Hafter, A. (2014). *Research on the use of Khan Academy in schools: Research brief*. Menlo Park: SRI International. Retrieved 2/18/16 from http://www.sri.com/sites/default/files/publications/2014-03-07_implementation_briefing.pdf

National Governors Association Center for Best Practices (NGACBP) & Council of Chief State School Officers (CCSSO). (2010). Common Core State Standards. Washington, DC: Authors.

National Research Council (2002). Scientific research in education. Washington, DC: The National Academies Press. Retrieved on 2/19/16 from http://www.nap.edu/catalog/10236/scientific-research-in-education.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*. 45 (3): 487-501.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*. 349 (6251). DOI: 10.1126/science.aac4716

Ostrow, K. & Heffernan, C. (2016) ASSISTments TestBed resource guide: The if-then-else section type. Retrieved on 2/18/16 from http://tiny.cc/IfThenElse.

Ostrow, K.S. & Heffernan, N.T. (2014). Testing the multimedia principle in the real world: A comparison of video vs. text feedback in authentic middle school math assignments. In Stamper, Pardos, Mavrikis, McLaren (eds.) Proceedings of the 7th International Conference on Educational Data Mining, 296-299.

Ostrow, K. & Heffernan, N. (2015). The role of student choice within adaptive tutoring. In Conati, Heffernan, Mitrovic & Verdejo (eds.) Proceedings of the 17th International Conference on Artificial Intelligence in Education. Springer International Publishing. 752-755.

Ostrow, K.S., Schultz, S.E., & Arroyo, I. (2014). Promoting growth mindset within intelligent tutoring systems. In Ritter & Fancsali (eds.) NCFPAL Workshop, In CEUR-WS (1183), Gutierrez-Santos, S., & Santos, O.C. (eds.) EDM 2014 Extended Proceedings. 88-93.

Ostrow, K., Selent, D., Wang, Y., Van Inwegen, E., Heffernan, N., & Williams, J.J. (In Press). The Assessment of Learning Infrastructure (ALI): The Theory, Practice, and Scalability of Automated Assessment. To be included in the Proceedings of the 6th International Conference on Learning Analytics and Knowledge, 2016.

Pashler, H., Rohrer, D., Cepeda, N. & Carpenter, S.K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*. 14 (2): 187-193.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In Cortes, Lawrence, Lee, Sugiyama, & Garnett (eds.) Proceedings of the 28th Conference on Neural Information Processing Systems.

Rafferty, A.N. & Griffiths, T.L. (2014). Diagnosing algebra understanding via bayesian inverse planning." In Proceedings of the 7th International Conference on Educational Data Mining. 351- 352.

Rau, M., Aleven, V., Rummel, N., & Rohrbach, S. (2012). Sense making alone doesn't do it: Fluency matters too! ITS support for robust learning with multiple representations. In Cerri, Clancey, Papadourakis, & Panourgia (eds.), Proceedings of the 11th International Conference on Intelligent Tutoring Systems. Springer Berlin Heidelberg. 7315: 174-184.

Razzaq, L. & Heffernan, N. (2010). Hints: is it better to give or wait to be asked? In Aleven, Kay, & Mostow (eds.) Proceedings of the 10th International Conference on Intelligent Tutoring Systems, Part 1. Springer Berlin Heidelberg. 6094: 349-358.

Razzaq, L., Heffernan, N.T., Lindeman, R.W. (2007). What level of tutor interaction is best? In Luckin & Koedinger (eds.) Proceedings of the 13th Conference on Artificial Intelligence in Education. IOS Press. 222-229.

Richland, L.E., Kornell, N., & Kao, L.S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*. 15 (3): 243.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society. 58 (5): 527-535.*

Roediger, H.L. & Karpicke, J.D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*. 1 (3): 181-210.

Roll, I., Holmes, N.G., Day, J., & Bonn, D. (2012). Evaluating metacognitive scaffolding in guided invention activities. *Instructional Science*. Springer Netherlands. 40 (4): 691-710.

Roschelle, J., Feng, M., Heffernan, N. & Mason, C. (2015). Preliminary findings from an efficacy study of online mathematics homework. Poster presented at a U.S. Department of Education (IES) meeting of investigators of funded projects. Available at http://tiny.cc/ASSISTmentsStudy.

Shrestha, P., Wei, X., Maharjan, A., Razzaq, L., Heffernan, N.T., & Heffernan, C., (2009). Are worked examples an effective feedback mechanism during problem solving? In Taatgen & van Rijn (eds.) Proceedings of the 31st Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society. 1294-1299.

SI2-SSE&SSI, National Science Foundation. (2014). Adding research accounts to the ASSISTments' platform: Helping researchers do randomized controlled studies with thousands of students. Award # 1440753. Abstract retrieved from: http://www.nsf.gov/awardsearch/showAward?AWD_ID=1440753.

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, Bull, Kay, & Mitrovic (eds) Proceedings of the 15th International Conference on Artificial Intelligence in Education Conference. Springer Berlin Heidelberg. 6738: 328–336.

Soffer, D., Das, V., Pellegrino, G., Goldman, S., Heffernan, N., Heffernan, C., & Dietz, K. (2014). Improving long-term retention of mathematical knowledge through automatic reassessment and relearning. American Educational Research Association Conference, Division C - Learning and Instruction / Section 1c: Mathematics.

Stamper, J., Eagle, M., Barnes, T. & Croy, M. (2013). Experimental evaluation of automatic hint generation for a logic tutor. In Biswas, Bull, Kay, & Mitrovic Proceedings of the 15th International Conference on Artificial Intelligence in Education. Springer Berlin Heidelberg. 6738: 345-352.

Storm, B.C., Bjork, E.L., Bjork, R.A., & Nestojko, J.F. (2006). Is retrieval success a necessary condition for retrieval-induced forgetting? *Psychonomic Bulletin & Review*. 13: 1023-1027.

Sweller, J., Kirschner, P.A., Clark, R.E. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. *Educational Psychologist*. 42 (2): 115-121.

Szpunar, K.K., Khan, N.Y., Schacter, D.L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. Proceedings of the National Academy of Sciences. 110 (16): 6313 - 6317.

U.S. Department of Education. (2010). Transforming American education: Learning powered by technology, the national educational technology plan. Office of Educational Technology, Washington, D.C.

Walker, E., Rummel, N., & Koedinger, K.R. (2011). Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *International Journal of Computer-Supported Collaborative Learning*. Springer US. 6 (2): 279-306.

Walkington, C. (2013). Using learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, 105 (4): 932-945.

Wang, Y. & Heffernan, N. (2014). The effect of automatic reassessment and relearning on assessing student long-term knowledge in mathematics. In Trausan-Matu, Boyer, Crosby, & Panourgia (eds.) Proceedings of the 12th International Conference on Intelligent Tutoring Systems. Spring International Publishing. 8474: 490-495.

Whorton, S. (2013). Can a computer adaptive assessment system determine, better than traditional methods, whether students know mathematics skills? Master's Thesis, Computer Science Department. Retrieved 2/19/16 from http://www.wpi.edu/Pubs/ETD/Available/etd-041913-095912/.

Williams, J.J. (2013). Improving learning in MOOCs by applying cognitive science. In Pardos & Schneider (eds.) Part 1: Workshop on Massive Open Online Courses (moocshop). In Walker & Looi (eds.) Proceedings of the Workshops at the 16th International Conference on Artificial Intelligence in Education.

Williams, J.J. (2014). Using randomized experiments as a methodological and conceptual tool for improving the design of online learning environments (Working Paper). Retrieved from the Social Science Research Network: http://ssrn.com/abstract=2535556

Williams, J.J., Ostrow, K., Xiong, X., Glassman, E., Kim, J., Maldonado, S.G., Reich, J., & Heffernan, N. (2015). Using and designing platforms for in vivo educational experiments. In Woolf, Russell, & Kiczales (eds.) Proceedings of the Second ACM Conference on Learning @ Scale. ACM, New York. 409-412.

Williams, J.J., Paunesku, D., Haley, B., & Sohl-Dickstein, J. (2013). Measurably increasing motivation in MOOCs. Talk presented at the MOOCshop Workshop, International Conference on Artificial Intelligence in Education.

Williams, J.J., Maldonado, S., Williams, B.A., Rutherford-Quach, S., & Heffernan, N. (2015). How can digital online educational resources be used to bridge experimental research and practical applications? Embedding in vivo experiments in "MOOClets." Society for Research on Educational Effectiveness Spring 2015 Conference. Washington, D.C.

Xiong, X. & Beck, J.E. (2014). A Study of Exploring Different Schedules of Spacing and Retrieval Interval on Mathematics Skills in ITS Environment. In Trausan-Matu, Boyer, Crosby, & Panourgia (eds.) Proceedings of the 12th International Conferences on Intelligent Tutoring Systems. Springer International Publishing. 8474: 504-509.

2 Data Reporting Infrastructure: ALI

This chapter was originally published as a conference paper in the Proceedings of the 6th International Conference on Learning Analytics and Knowledge (LAK). I presented this work in Edinburgh, Scotland in April 2016. It is included in Part I because it was the seminal paper describing the development of the Assessment of Learning Infrastructure (ALI), or a programmatic approach to deliver anonymized data to researchers using the ASSISTments TestBed. The paper discussed how online learning platforms can be used to conduct educational research at scale and described the evolution of ASSISTments as a research tool. The paper also focused on automated data reporting and delivery. ALI was described in theory, detailing its potential benefits for the research community, and in practice, describing the infrastructure's capabilities. ALI was then used to conduct analysis at scale by running chi-squared tests across 25 experiments from the ASSISTments TestBed to gather information about the attrition rates of experimental conditions. This paper advocated use of the TestBed and ALI by researchers around the country as a promotional tool for the NSF SI2 grant (NSF SI2-SSE&SSI: 1440753).

Proper citation of this chapter is as follows:

Ostrow, K., Selent, D., Wang, Y., Van Inwegen, E., Heffernan, N., & Williams, J. J. (2016). Assessment of Learning Infrastructure (ALI): The Theory, Practice, and Scalability of Automated Assessment. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge*, 2016. Edinburgh, Scotland, April 25-29.

2.1 Introduction

An immense community of researchers, educators, and administrators seeks to enhance the effectiveness of education through rigorously controlled experimentation. Those invested in K-12 education struggle not just to enhance pedagogy, curriculum, and student engagement, but also to harness the power of technology in ways that will optimize learning. Researchers often fall back on observational studies or turn to data mining large longitudinal datasets due to the difficulties inherent to conducting student-level randomized controlled experiments (RCEs) in authentic learning environments. However, the RCE remains the "gold standard" of determining causal relationships and was referred to when the U.S. Department of Education advocated for K-12 schools to apply basic findings from cognitive science to improve educational practices (Pashler, Rohrer, Cepeda & Carpenter, 2007). Without the assistance of scalable technologies, the call to conduct RCEs within authentic scholastic settings (Institute of Education Sciences, 2003) is difficult for many researchers to answer due to the high cost of establishing and maintaining sample populations, the complications inherent to randomization at the class level (i.e., larger samples are required in comparison to studies with student level randomization), and the curriculum restrictions necessary to establish sound controls.

When designed with flexibility and collaboration in mind, online learning platforms offer a unique and scalable approach to educational research and data analysis. Users of online learning platforms (i.e., students and teachers) create hundreds of thousands of data points on a daily basis, with databases of rich learner information growing exponentially as systems gain popularity and validity as powerful learning aids. These systems offer the opportunity to collect information including (but not limited to) behavior and affect (Beck & Gong, 2013; San Pedro, Baker, Gowda & Heffernan, 2013), learning interventions based in content and delivery (Ostrow & Heffernan, N. T., 2014; Ostrow, Heffernan, Heffernan & Peterson, 2015), and interactions within skill domains that help guide curriculum development. By allowing data scientists, psychologists, psychometricians, and K-12 educators to work collaboratively within online learning platforms, they are able to dynamically evaluate and improve the effectiveness of the platform and its content while empowering learner analytics.

Software intended for sharing educational data and instruments has driven tremendous progress in educational research and best practices. For instance, the Pittsburgh Science of Learning Center's Datashop (Koedinger, Baker, Cunningham, Skogsholm, Leber & Stamper, 2010), funded by the National Science Foundation, provides an extensive database of educational datasets for data mining and analysis. However, the pace and power of educational research would increase drastically if researchers had easier access to environments in which they could design, implement, and analyze real time data with regard to their own hypotheses. When online learning platforms offer flexibility in content design, manipulation, and delivery, researchers are able to access robust student populations that already use these systems for classwork and homework in authentic K-12 learning environments. Content can be manipulated to include parallel assignments, fashioned as conditions within RCEs, allowing researchers to quickly determine what works for whom and when. Designing these environments with the open, collaborative, and perhaps even competitive design of RCEs in mind can lead to strengthened internal validity through the sound design of content and the promotion of open source data reporting for review and replication upon publication (Open Sci Collab, 2015).

2.1.1 ASSISTments as a Research Tool

ASSISTments is a unique online learning platform that was designed with educational research as one of its primary goals (Heffernan & Heffernan, 2014). The platform has grown into a shared scientific instrument that allows researchers to conduct RCEs within authentic learning environments. The process typically involves a researcher modifying preexisting certified content to include treatment interventions and student-level random assignment. The latter feature makes the ASSISTments system unique and robust for conducting research; rather than all students within a single class experiencing the same condition, each student may receive slightly different content or feedback within the same assignment. The library of certified ASSISTments content consists primarily of middle and high school mathematics skills, with content organized and tagged by Common Core State Standard (NGACBP & CCSSO, 2010). However, this library has grown to include content in physics, chemistry, and electronics, and researchers are able to develop their own content for experimentation in other domains.

Figure 2.1 depicts a simple study design implemented within ASSISTments. This universal design could be applied to any skill within the platform. The design depicts the paths a student might take based on their ability to access video content. When a student begins the assignment, she must first answer a “Video Check,” or a standard problem that essentially serves as password protection to study participation. If the student can access video, she enters the password provided, and her ‘correct’ response serves as the “Then” in an “If-Then” routing structure. If the student enters anything other than the password, she is provided a default assignment without video content and is removed from analysis of the intervention under examination. Upon being routed into the study depicted in Figure 2.1, students are randomly assigned into one of two assignments using a “Choose Condition” routing structure. Note that two conditions are presented here for simplicity, although the system is able to compare any number of conditions. This random assignment will be discussed further in Section 2.3.1.2.

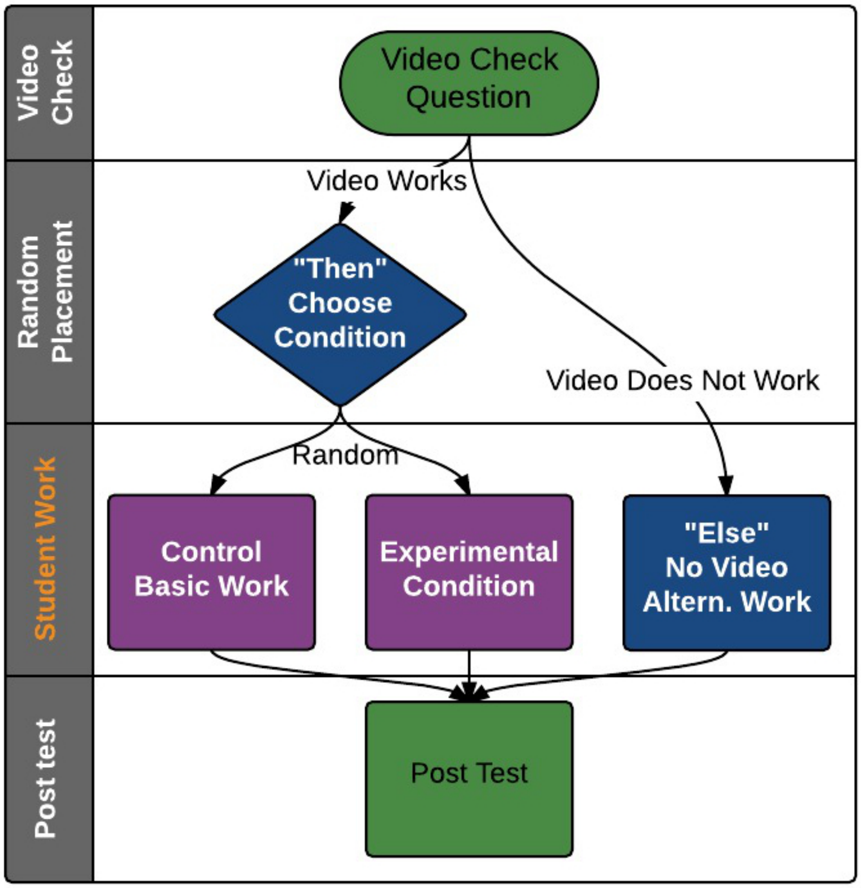


Figure 2.1. A simple research design that can easily be built within ASSISTments to compare learning interventions.

In the present example, there are three possible paths that a student may follow as she progresses through her assignment. The trace of these paths will become important in the automatic reporting and analysis of student performance presented in Section 2.3. For each student, regardless of path, ASSISTments logs substantial data detailing the student's performance as they progress through the assignment. This data includes binary measures of problem accuracy (i.e., a correct or incorrect first response), the student's first action (i.e., an attempt vs. requesting tutoring), the number of attempts per problem, the number of feedback interactions per problem (i.e., hints requested or scaffolds seen), whether or not the student saw the bottom out hint (i.e., the correct answer, provided to keep the student from getting stuck within the assignment), and start and end times for each problem. ASSISTments is also able to track a variety of user information that is ultimately helpful to researchers, including data on the student's past performance in the system, student characteristics (i.e., gender, age), and variables at the class and school levels. For researchers with a fine-toothed comb, ASSISTments can also provide logged information at the action level, detailing each step taken within a problem. All of this information is consolidated, anonymized, and provided to researchers through unified reports (depicted in Section 2.3.1.1) to enhance the ease with which researchers conduct RCEs within ASSISTments.

2.1.2 Utility of Automated Data-Preprocessing

With students accessing experiments naturally in authentic learning environments, sample populations increase as a function of time. For instance, within three months of deploying a study within ASSISTments, the study may accrue 740 participants. The data mentioned above is logged for every problem seen by every student. As external researchers are unfamiliar with the ASSISTments database and the inner workings of the platform, universal data reporting and preprocessing techniques were designed to ease the hurdle of interpreting system output. Without this preprocessing, a researcher analyzing data from the study depicted in Figure 2.1 would need to use raw data to decipher whether students should be included in analyses, what condition each student experienced, details pertaining to each student's experience within that condition (i.e., how many problems were completed, their content, and all associated data), and how each student performed at posttest. While such rich information is surely helpful in analyzing a study, providing researchers with a surplus of data necessitates larger and more complex datasets that must still meet ease of use requirements. Although different researchers focus on different information (as it applies to their particular hypotheses), an infrastructure for data preprocessing, restructuring, and reporting was necessary to bring ASSISTments to the next level as a shared scientific instrument for educational research.

In the following sections, we discuss the creation of an automated reporting and analysis system built to provide researchers with data logged from RCEs conducted within ASSISTments. The Assessment of Learning Infrastructure (ALI) has built upon existing technology to foster a symbiotic relationship beneficial to students, researchers, the platform and its content, and the science of learning. Evolving from a universal data logging and retrieval tool, ALI is quickly becoming a sophisticated system for automated analysis, providing an overview of sample populations and conducting a selection of analyses for researchers to consider when assessing their data. ALI's benefits can also be felt at scale, with analyses spanning content to drive content and platform development, and to eventually offer users with a more personalized learning experience.

2.2 ALI in Theory

ALI is an automated research assistant that, while not meant to replace the researcher, is meant to lighten the load of working with large data files output from RCEs conducted within ASSISTments. ALI alerts the researcher to new data, presents that data in a meaningful way, tentatively examines effects observed between conditions, and flags potential threats to validity. On a weekly basis, or immediately upon a request from the researcher, ALI consults all logged information pertaining to a study and conducts preliminary analyses on student participation and performance. The benefits of such automated reporting and analysis are broad, and in the next four sections we briefly discuss how ALI's success will affect ASSISTments and its users, as well as the greater learning analytics community.

2.2.1 Benefits to the User

ALI's work at scale helps to guide the development of stronger learning interventions and, eventually, to personalize the learning experience for student users. Research conducted within ASSISTments is unique in that while researchers are able to alter content to deliver versatile interventions as exemplified in Figure 2.1, the learning experience remains authentic and uninterrupted. Although student performance information is logged quite passively, it forms a surplus of data for learner analytics. A student may notice that some of their assignments include video feedback or have extra survey questions while others do not, but they are not usually informed that they are participating in an RCE. One of the goals driving the platform's ability to implement RCEs is the provision of normal instructional practice and interventions that do not compromise learning.

ALI is also beneficial to teachers, as the infrastructure is able to separate rich study information from daily assessment data. Teachers are responsible for assigning content within ASSISTments to their students. Although research designs would appear to complicate daily assessment, class and student reports were designed to provide teachers with pertinent information and omit superfluous data pertaining to the intervention under investigation. This low-profile approach to conducting research maintains a highly participatory subject pool.

2.2.2 Benefits to the Researcher

For those conducting RCEs within ASSISTments, ALI plays the role of research assistant. The infrastructure intelligently communicates with researchers when new data is available for analysis and provides an overview of the sample distribution across conditions to signify the power available from current analyses. Although researchers will undoubtedly run their own in-depth analyses, standard high-level analyses can be automated to save time and reduce monotony. For example, ALI's ability to trace a student's path through an assignment allows the infrastructure to infer the condition that the student experienced. This allows ALI to test for differential attrition rates across conditions and notify the researcher of apparent selection biases. Even this simple analysis can be a beneficial warning against analyzing posttest results due to the potential threat to internal validity. At the scale of a single study, ALI is able to give researchers insights about when and how they should approach data analysis. Combined with the preprocessing and reporting that ALI's analytics are built upon, this basic information is often enough to save the researcher from hours of repeated analysis.

2.2.3 Benefits to the Platform

When considered at scale, ALI's capabilities for data reporting and analysis also contribute to the enhancement of the ASSISTments platform by supporting practical improvements to skill content without interrupting student learning. As researchers collaborate and compete to design interventions within ASSISTments content, it will grow increasingly possible to evaluate interventions at scale, both across content and longitudinally within students. Ideally, the best iteration of content and delivery observed for a particular skill will be delivered to students as the control condition within new RCEs. Through this approach, each study will serve as a potential iteration toward the platform's improvement as experiments are launched and re-launched, capturing key features of design-based educational research methodology (Brown, 1992). These iterative improvements also benefit users through the predicted outcome of enhanced learning gains, and benefit researchers through the rapid succession and enhanced validity of positive findings.

ALI's ability to analyze at scale will also benefit the platform, as it will help the ASSISTments team to quickly isolate and remove ineffective interventions. In the near future, ALI will also be able to conduct analyses across multiple studies simultaneously while considering student, class, and school level characteristics. Roughly speaking, ALI will allow ASSISTments to personalize learning by better understanding why certain educational practices and interventions work for certain students but not for others.

2.2.4 Benefits for Learning Analytics

How can ALI and the promotion of infrastructures like ALI within other learning platforms benefit the learning analytics community? At its very core, ALI answers the general call of learning analytics, in that the infrastructure "emphasizes measurement and data collection as activities that institutions need to undertake and understand, and focuses on the analysis and reporting of the data" (U.S. DOE, 2012). A strong focus on providing universal measures of learning garnered from authentic learning environments will strengthen the validity of findings from a broad range of interventions that seek to isolate best practices in education.

Further, much attention in the broader scientific and psychological research communities has recently been focused on the general inability to replicate research findings (Ioannidis, 2005; Open Sci Collab, 2015). The same is likely true for educational research, with little emphasis placed on data accountability. Perhaps the best outlet for promoting open data, the Pittsburgh Science of Learning Center's Data Shop (Koedinger, et al., 2010) takes a number of steps in the right direction with regard to shared datasets that promote open, replicable, and sound science. ALI builds upon the PSLC's model of open data reporting by establishing stable, timestamped links to every data analysis report ever provided to a researcher throughout the duration of their work within ASSISTments. Researchers are simply asked to cite the report from which they draw data for final analyses that lead to publication (explained further in Section 2.3.1.5). These links will also drastically increase the availability of preprocessed and anonymized educational datasets to researchers wishing to mine big data without designing specific interventions.

In some ways, ALI is also an extension of industry track research focused on learning analytics, as companies like Google and Microsoft increasingly implement large-scale experimentation in online learning environments to consider reporting and analysis metrics and methods that meet practical goals rooted in scientifically sound evidence (Kohavi, Longbotham, Sommerfield & Henne, 2009). Infrastructures like ALI should be incorporated into learning platforms of all styles in order to promote the importance of learning analytics.

2.3 ALI in Practice

The Assessment of Learning Infrastructure has grown considerably over the past year. ALI began as a robust SQL query to the ASSISTments database to retrieve unified information across multiple studies and to present it to researchers in a single format. Ease of use requirements, communication considerations, and feedback from external researchers has helped ALI to grow beyond data preprocessing and reporting to a tool for learning analytics at scale. The following sections discuss how ALI has evolved and provides examples of the infrastructure's current capabilities in reporting, analyzing, and communicating data from RCEs conducted within ASSISTments.

2.3.1 ALI's Current Capabilities

2.3.1.1 Data Reporting at Scale

When a study is activated within ASSISTments, details about the study and the researcher's contact information are entered into ALI's study repository. Although researchers can request immediate data analysis reports on demand, ALI defaults to a weekly inspection of each study in the database and makes decisions regarding whether or not to process a data analysis report for the researcher. These decisions are based on measured increases in sample size. Due to common curriculum structures, certain skills are only used at specific times of year and thus, an assignment may be highly used during the Fall term but less used during the Spring term. When ALI inspects the study's logged data, if at least three new students have participated in the study since the researcher last received ALI communication, the data analysis report is triggered. If ALI detects that no additional students have participated in the study, the infrastructure refrains from communication in order to keep from spamming the researcher with multiple versions of the same data.

As teachers using ASSISTments are able to make copies of assignments and alter their content, ALI is also able to detect when teachers have assigned a study copy. ALI is sophisticated enough to recognize when a copy is identical to the original study and include the data associated with those copies in each report. If a copy of the assignment has been altered (i.e., problems were removed, or sections were changed), ALI no longer considers the content equivalent to the researcher's study and does not report data for students assigned that copy. This ensures that researchers get all data associated with their experiment without the addition of corrupt data.

Once ALI has determined that new data is available, several robust SQL queries are run on the ASSISTments database. The three major queries are used to a) retrieve student data detailing student, class, and school level characteristics for each student recorded prior to random assignment (see Table 2.1), b) retrieve problem level data (see Table 2.3), and c) detect the problem set structure (i.e., the paths depicted in Figure 2.1) for each student with logged data. These three queries provide ALI with the necessary information to establish reports and conduct automated analysis on experiments. By working closely with researchers throughout the development of ALI, four different universal data representations have been designed and are provided to all researchers in an attempt to meet dynamic research needs. Subsets of data exemplifying each type of report are provided below. Table 2.2 shows fields from the Action Level file. This file is the finest granularity of data logged by ASSISTments as a student works through an assignment. Each row provides information pertaining to a single step within a problem (i.e., when the problem is initiated, or when the student asks for a hint). A subset of the Problem Level file is depicted in Table 2.3. This file provides the same data as that found in the Action Level file, but the granularity has increased. Each row provides information pertaining to a single problem, with actions collapsed into columns. Student Level files, as depicted in Table 2.4, offer the coarsest granularity of data reporting. Within this type of file, each row provides information pertaining to the entire assignment of a single student. Problem information is presented across columns in the order in which the student experienced the assignment, with the number of columns extrapolated to the maximum number of problems experienced by any student in the file. An alternative version of Student Level data is also provided in which each student assignment is represented by a series of rows, each representing a piece of information about problems that are displayed across columns (similar to a pivot table, not displayed here). A full example of each data report is available at (Ostrow & Heffernan, 2014) for further consideration. Each data report is presented to the researcher in a single, organized communication, depicted in Figure 2.2 and discussed further in Section 2.3.1.5.

Table 2.1. A theorized subset of student historical data. Each row contains student, teacher, and school characteristics linked to a particular student, using information sourced prior to random assignment.

Student	Class ID	Grade	School ID	Guessed Gender	Birth Year	Prior HW % Completion	Prior Class HW % Completion	Z-Scored HW Mastery Speed
A	1007475	1	1	1	1	1	18.2	14.3
B	1180278	0	0	0	0	1	17	14.1
C	1180278	1	0	0	1	-	24.6	14.3
D	1322778	0	1	1	1	1	18.2	14.3

Table 2.2. A theorized subset of an action level data file. Each row represents a single action within a single problem as experienced by a student. This is the finest granularity of data reported by ALI.

Student	Problem ID	Sub-Problem ID	Order	Action Type	Timestamp	Answer	Correctness
A	PRAUVJS	806533	1	Start	08/26/15 15:25:26	--	--
A	PRAUVJS	806533	2	Hint	08/26/15 15:25:52	--	--
A	PRAUVJS	806533	3	Answer	08/26/15 15:26:40	18.2	TRUE
A	PRAUVJS	806533	4	End	08/26/15 15:26:42	--	--
A	PRAVKJX	833840	1	Start	08/26/15 15:26:43	--	--

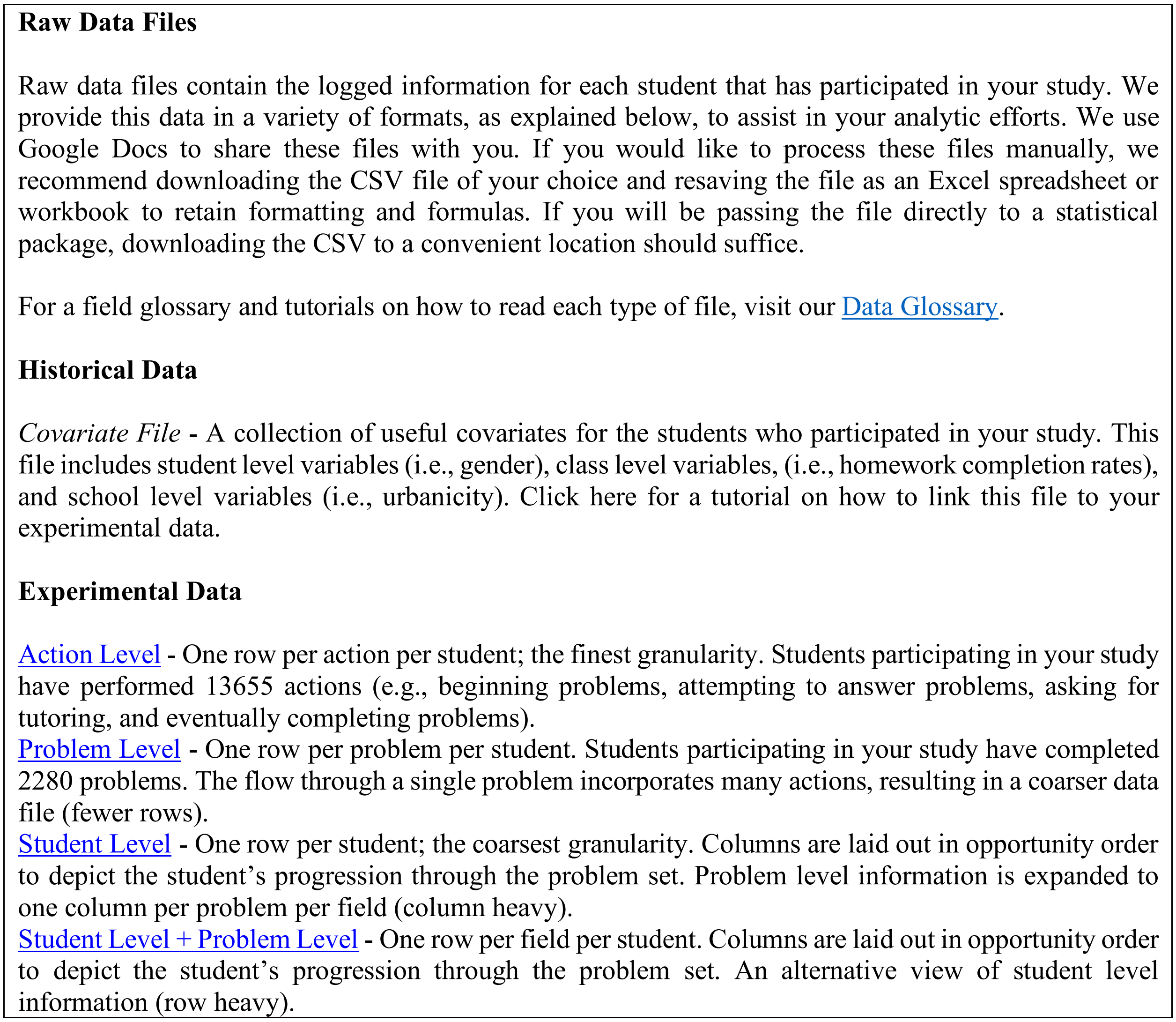
Table 2.3. A theorized subset of a problem level data file. Each row contains all the information linked to a single problem as experienced by a student. This is a popular form of data for student modeling and analytics.

Student	Assignment ID	Problem ID	Correct	Answer	Hints	Attempts	Start Time	End Time
A	1007475	PRAUVJS	1	18.2	0	1	08/26/15 15:25:26	08/26/15 15:26:42
A	1007475	PRAVKJX	1	14.3	0	1	08/26/15 15:26:43	08/26/15 15:27:45
A	1007475	PRAVKHT	1	6.4	0	1	08/26/15 15:27:50	08/26/15 15:28:47
B	1180278	PRAUVJX	0	22.8	2	3	08/26/15 17:14:22	08/26/15 17:15:42
B	1180278	PAVKGZ	0	7.2	0	2	08/26/15 17:15:43	08/26/15 17:17:31

Table 2.4. A theorized subset of a student level data file. Each row contains all information linked to a single student's experience of the problem set. Assignment information is presented across columns in the order in which the student experienced problems.

Student	Assignment ID	Late	Mastered	Correct Q1	Correct Q2	Correct Q3	Answer Q1	Answer Q2	Answer Q3
A	1007475	1	1	1	1	1	18.2	14.3	6.4
B	1180278	0	0	0	0	1	17	14.1	6.4
C	1180278	1	0	0	1	-	24.6	14.3	-
D	1322778	0	1	1	1	1	18.2	14.3	6.4

When preprocessing is complete, and these reports have been compiled, ALI sends analytic commands to Rserve, an extension to the R programming language that allows for other applications to call R functions via TCP/IP connection (Urbanek, 2003). The ASSISTments team created a client-side API to interact with Rserve, allowing ALI to send requests to R via a TCP/IP connection. Because Rserve is not multithreaded, several instances of Rserve run on separate ports on the ALI server. The server is designed to recycle existing connections, with a connection pool equal to the maximum number of threads ALI uses. This allows several data analysis reports to occur simultaneously, all using different Rserve connections. This approach lowers the turnaround time when a researcher actively requests data. It also keeps weekly reporting as efficient as possible, as all datasets in ALI's study repository are assessed weekly for potential reporting.



Raw Data Files

Raw data files contain the logged information for each student that has participated in your study. We provide this data in a variety of formats, as explained below, to assist in your analytic efforts. We use Google Docs to share these files with you. If you would like to process these files manually, we recommend downloading the CSV file of your choice and resaving the file as an Excel spreadsheet or workbook to retain formatting and formulas. If you will be passing the file directly to a statistical package, downloading the CSV to a convenient location should suffice.

For a field glossary and tutorials on how to read each type of file, visit our Data Glossary.

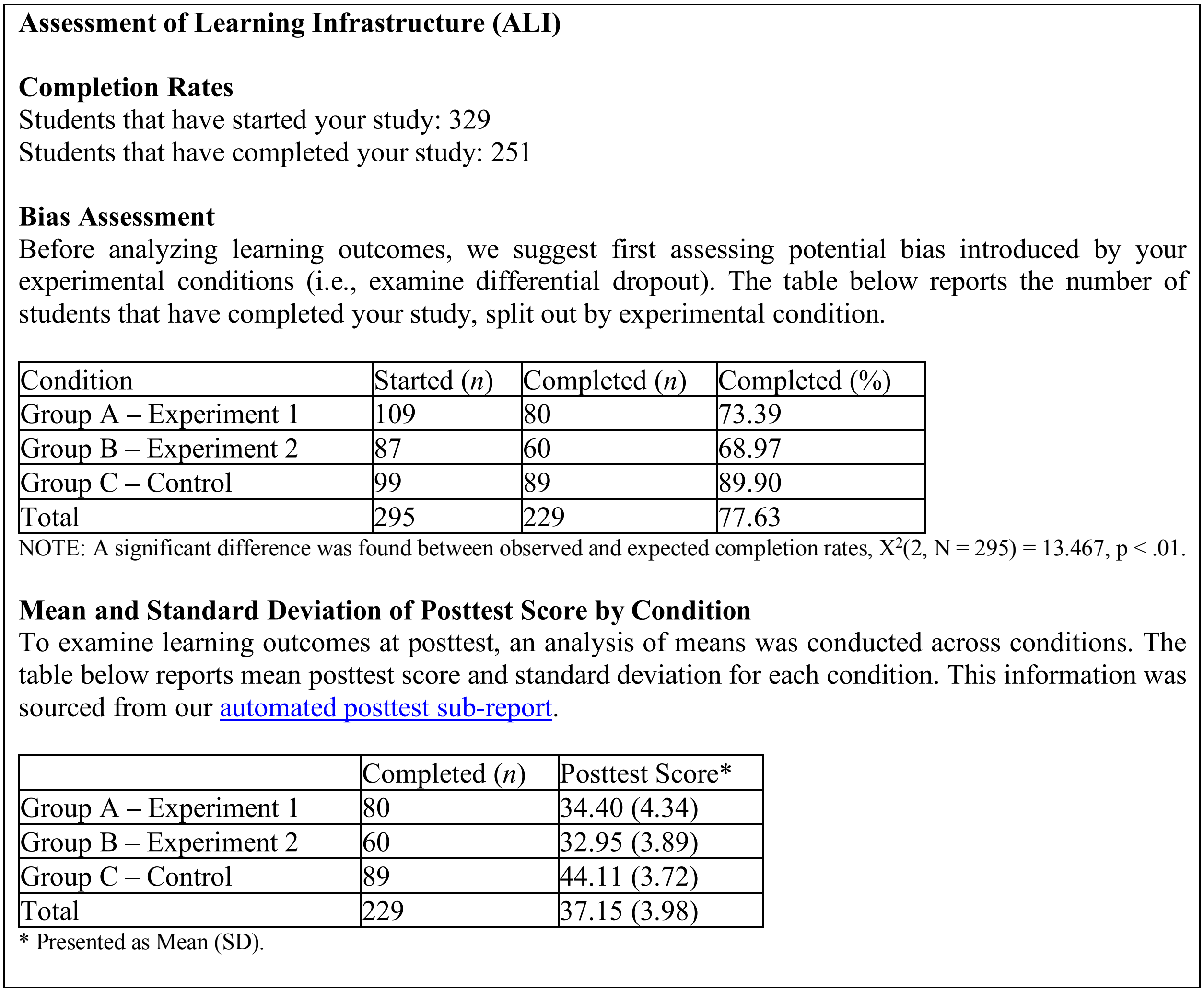
Historical Data

Covariate File - A collection of useful covariates for the students who participated in your study. This file includes student level variables (i.e., gender), class level variables, (i.e., homework completion rates), and school level variables (i.e., urbanicity). Click here for a tutorial on how to link this file to your experimental data.

Experimental Data

Action Level - One row per action per student; the finest granularity. Students participating in your study have performed 13655 actions (e.g., beginning problems, attempting to answer problems, asking for tutoring, and eventually completing problems).
Problem Level - One row per problem per student. Students participating in your study have completed 2280 problems. The flow through a single problem incorporates many actions, resulting in a coarser data file (fewer rows).
Student Level - One row per student; the coarsest granularity. Columns are laid out in opportunity order to depict the student's progression through the problem set. Problem level information is expanded to one column per problem per field (column heavy).
Student Level + Problem Level - One row per field per student. Columns are laid out in opportunity order to depict the student's progression through the problem set. An alternative view of student level information (row heavy).

Figure 2.2. A thoroughly developed universal reporting of logged data from students participating in RCEs. Each file presented here is discussed further, including depictions of file subsets, in Section 2.3.1.1.



Assessment of Learning Infrastructure (ALI)

Completion Rates
Students that have started your study: 329
Students that have completed your study: 251

Bias Assessment
Before analyzing learning outcomes, we suggest first assessing potential bias introduced by your experimental conditions (i.e., examine differential dropout). The table below reports the number of students that have completed your study, split out by experimental condition.

Condition	Started (*n*)	Completed (*n*)	Completed (%)
Group A – Experiment 1	109	80	73.39
Group B – Experiment 2	87	60	68.97
Group C – Control	99	89	89.90
Total	295	229	77.63

NOTE: A significant difference was found between observed and expected completion rates, $X^2(2, N = 295) = 13.467, p < .01$.

Mean and Standard Deviation of Posttest Score by Condition
To examine learning outcomes at posttest, an analysis of means was conducted across conditions. The table below reports mean posttest score and standard deviation for each condition. This information was sourced from our automated posttest sub-report.

	Completed (*n*)	Posttest Score*
Group A – Experiment 1	80	34.40 (4.34)
Group B – Experiment 2	60	32.95 (3.89)
Group C – Control	89	44.11 (3.72)
Total	229	37.15 (3.98)

* Presented as Mean (SD).

Figure 2.3. Current ALI analytic reporting. Available analyses include a Chi-squared test comparing the observed and expected sample distributions, simple hypothesis testing, and an analysis of means on posttest performance between conditions. Note that these analyses are currently driven by the structure of the assignment as parsed by ALI from Problem Level data. Future work includes allowing researchers to tag their study with items of interest to automate more sophisticated analysis.

2.3.1.2 Smart Structures

In order to determine exactly what to analyze, ALI must first process the structure of a study and trace each student's path through the assignment (as previously discussed in relation to Figure 2.1). As ALI parses an assignment's structure, the infrastructure is able to make intelligent decisions upon meeting certain section types within the design. This is accomplished by recursively regenerating an assignment's reported structure into tree form. Using the Problem Level data file presented in Table 2.3, ALI maps the data logged for each student onto the problem set structure. Within this file, each problem is labeled with a path, similar to that used when traversing a set of folders within an operating system. ALI steps through each problem path for each student to establish an intuitive structure of the study and to cluster students by condition.

RCEs within ASSISTments are typically designed by taking advantage of a variety of section types offered by the platform. The "If-Then" routing discussed in Section 2.1.1 was an example of a section type. When ALI observes an If-Then structure that issues a routing standard like a "Video Check," the infrastructure intelligently reports on students assigned to the study and disregards those who were routed to alternative content. Similarly, studies often employ parallel experimental and control conditions delivered using a section type referred to as a "Choose Condition." This section type is used to drive random assignment into one of multiple possible conditions. The "Choose Condition" used in the study design depicted in Figure 2.1 included two parallel conditions: an assignment with video content and a control assignment with traditional text content. If ALI reaches a "Choose Condition" when mapping an assignment's structure, the infrastructure recognizes the assignment as a research study. ALI compares the problem content within each condition and considers any problems that fall into a section immediately following the condition as a subsequent posttest (see Figure 2.1). Using this information, ALI is able to aggregate statistics and perform a selection of simple analyses across the data.

It is important to note that research designs within ASSISTments can grow far more complex than the simple structure presented herein. When assignments include nested section types and multiple "If-Then" routing standards, ALI currently has difficulty interpreting condition and isolating posttest content. In its current form, ALI is only meant to assist researchers with the analysis of common design patterns. Future work, discussed in Section 2.5, will expand ALI's ability to intelligently parse studies using rules set forth by the researcher.

2.3.1.3 Selection Bias

After establishing a study's structure and sample distribution, ALI is able to assess assignment completion rates across condition and alert researchers to potential threats to internal validity due to selection threats. ALI records the observed number of students in each condition that began the assignment, and considers logged assignment end times to consider the proportion of students that ultimately completed the assignment. The observed distribution is then compared to the expected distribution of proportional attrition in a normal sample. A Chi-squared analysis is then used to determine if the observed distribution significantly differs from the expected distribution. This allows ALI to flag conditions that have a reliably different attrition rate and alert the researcher of potential threats to internal validity. Without considering differential attrition across conditions, an analysis of posttest performance may inaccurately suggest the significant effect of a particular condition that was actually driven by the disproportionate loss of weaker students from that condition. This simple analysis, presented to researchers using the format shown in Figure 2.3, may help even the most seasoned experts to accurately assess their sample.

It is important to note that while ALI provides this warning, the infrastructure still releases all data to the researcher and never prohibits the researcher from further analysis. The goal of ALI's selection bias assessment is not to impede or prevent analysis, but rather to promote sound science.

2.3.1.4 Simple Hypothesis Testing

After conducting a selection bias assessment, ALI moves on to a set of simple hypothesis tests pertaining to posttest performance. If ALI detects a posttest section when parsing an assignment's structure, the infrastructure compares performance across conditions by referring to the previously aggregated group distributions. ALI approaches posttest analyses much like a researcher would: if only two conditions are detected within the study, ALI conducts a t-test, while if more than two conditions detected, ALI conducts an ANOVA. ALI currently has the API to support simple univariate and multivariate analyses including ANOVA, ANCOVA, MANOVA, and MANCOVA. All of these analyses are conducted using a similar approach. ALI stores all input parameters for a given statistical test in a single object. The parameters are extracted from this object and transformed into the appropriate R function calls through the Rserve API communication. Results are accumulated and presented to the researcher alongside an analysis of means, as shown in Figure 2.3, allowing the researcher to observe the direction of the reported effect.

Note that in the present work, only two conditions are considered, and the t-test results are not presented to the researcher due to ALI's detection of a potential selection bias. Further, covariates are not considered in this analysis. Future work will control for popular student, class, and school level characteristics sourced from the historical student data file (see Table 2.1) by using ANCOVA or MANCOVA approaches in an attempt to explain additional variance in learning outcomes.

2.3.1.5 Data Storage and Researcher Output

Once ALI's automated analysis is complete, ALI stores all data and analytic output on Google Drive in archival quality. This data cannot be altered but can be downloaded by anyone and will exist for as long as Google supports Google Drive. For active studies, copyright protection will be placed on new data analysis reports for one year from the run date. This means that researchers will have a full calendar year to publish on their findings before their data becomes freely available to the public.

ALI communicates to researchers via email and provides a link to a stable URL for a Google Doc housing the data analysis report. This Doc contains links to all raw data files, as shown in Figure 2.2, and provides the automated analysis depicted in Figure 2.3. The creation of this Google Doc is automated, based on an HTML template file that uses custom tagging conventions to insert variables with dynamic text or data. Using this method, the same report can be generated multiple times or across multiple assignments, with only changes only to the pertinent information. This allows for customized reporting based on the results of ALI's analysis. The Google Doc report also provides researchers with links to additional resources including a glossary explaining features of the data and video tutorials on how to understand each file type (available at Ostrow & Heffernan, C. (2014)).

When a researcher publishes their findings, a condition of working within ASSISTments requires that they include a reference in their work to the stable record from which they sourced their data files for final analysis. This approach allows reviewers and secondary researchers to gain access to their raw data, thereby encouraging replication and sound science (Open Sci Collab, 2015). In addition to the raw data, secondary researchers will be able to access ALI's analytic report, including all automated preliminary analyses.

2.4 Analysis at Scale

Considered at scale, even the relatively simple analyses that ALI is already capable of can offer substantial insights for the ASSISTments platform and in more general terms, for the learning analytics community. By examining attrition outcomes across studies simultaneously it is possible to make claims about the quality of interventions compared to their potential for effects on learning gains. As ALI's analytical capabilities increase, analysis at scale will grow even more powerful.

As a proof of concept of the potential benefits of automated analysis at scale, ALI was run across a special file including 25 studies that are currently running within ASSISTments. This file was created for another sophisticated approach to modeling student performance across multiple studies (Selent, Patikorn, Heffernan & Beck, Under Review), but serves as a perfect example of ALI's capabilities at scale. In the spirit of open data, this file is available for reference at (Ostrow, 2015). The studies in this file were selected from a group of 126 studies currently running within the ASSISTments platform based on the following criteria:

- Studies selected contained at least 50 students that completed the assignment within each condition
- Studies selected were designed within Skill Builders, a mastery learning based assignment that considers predefined thresholds for student completion (i.e. 3 right in a row).

As most of the assignments in this file were built prior to the implementation of automated path-logging (which drives ALI's ability to read in the structure of the study and infer a condition for each student) condition was manually traced and logged for each student based on his or her observed problem sequence. A number of these studies were also built before the availability of If-Then routing and checks for internal validity (i.e., the "Video Check" explained in connection to Figure 2.1) and thus it is difficult to tell if students experienced technical difficulties during their experience of particular conditions. As ALI considers these routing questions in recently designed studies, manual notation was also used to try to flag whether students experienced the condition to which they were assigned. This flagging also included whether students 'tested out' of the condition experience (i.e., if a student was assigned to a condition in which the treatment was presented through feedback but answered the first 3 consecutive problems accurately, they did not ultimately experience the treatment). Further, as only three of the studies in this file contained

Table 2.5. ALI's Bias Assessment at Scale - Observed Distributions and Chi-Squared Analyses for 25 Problem Sets

Problem Set by Condition	Started (n)	Completed (n)	Completed (%)	df	χ^2	p
Multiplying Mixed Numbers 5.NF.B.4a	775	466	60.13	1	5.30	0.021*
Control	403	258	64.02			
Experiment	372	208	55.91			
Understanding Vocabulary About Circles G-C.A.2	695	674	96.98	1	4.87	0.027*
Control	330	325	98.48			
Experiment	365	349	95.62			
Equivalent Expression 6.EE.B.4	273	240	87.91	1	0.39	0.532
Control	138	123	89.13			
Experiment	135	117	86.67			
Writing Inequalities from Situations 6.EE.B8	627	539	85.96	1	2.21	0.138
Control	338	297	87.87			
Experiment	289	242	83.74			
Dividing Mixed Numbers 6.NS.A.1	1864	1285	68.94	1	0.99	0.321
Control	943	660	69.99			
Experiment	921	625	67.86			
Finding Expected Value SS.MD.B.5	457	337	73.74	1	0.06	0.802
Control	224	164	73.21			
Experiment	233	173	74.25			
Conditional Probability SS-CP.A.3	515	366	71.07	1	0.70	0.401
Control	281	204	72.60			
Experiment	234	162	69.23			
Permutations and Combinations SS-CP.B.2	540	456	84.44	1	0.00	0.958
Control	265	224	84.53			
Experiment	275	232	84.36			
Basic Logarithm Manipulation F-BF.B.5	136	121	88.97	1	0.21	0.645
Control	62	56	90.32			
Experiment	74	65	87.84			
Properties of Exponents 8.EE.A.1	545	435	79.82	1	0.24	0.626
Control	264	213	80.68			
Experiment	281	222	79.00			
Intermediate Logarithm Manipulation F-BF.B.5	205	169	82.44	1	8.44	0.004**
Control	102	92	90.20			
Experiment	103	77	74.76			
Solving ab^ct = d LE.A.4a	147	122	82.99	1	0.01	0.914
Control	72	60	83.33			
Experiment	75	62	82.67			
Finding Inverse Functions F-BF.B.4	301	143	47.51	1	3.32	0.068†
Control	145	61	42.07			
Experiment	156	82	52.56			
Composition of Functions F-BF.A.1c	219	173	79.00	1	0.86	0.354
Control	118	96	81.36			
Experiment	101	77	76.24			
Sequences F-BF.A.2	382	241	63.09	1	0.20	0.658
Control	198	127	64.14			
Experiment	184	114	61.96			
Comparing Values - Multiplying by Fractions 5.NF.B.5a	129	121	93.80	1	1.59	0.208
Control	69	63	91.30			
Experiment	60	58	96.67			
Converting Radians to Degrees F-TF.A.1	245	226	92.24	1	0.23	0.631
Control	129	120	93.02			
Experiment	116	106	91.38			
Trigonometric Ratios G-SRT.C.8	307	266	86.64	1	0.91	0.341
Control	141	125	88.65			
Experiment	166	141	84.94			

Note. †p < .10, *p < .05, **p < .01

Table 2.5. ALI's Bias Assessment at Scale - *Continued*

Problem Set by Condition	Started (n)	Completed (n)	Completed (%)	df	χ^2	p
Prime Factorization 6.NS.B.4	1238	1058	85.46	2	0.97	0.616
Control	430	369	85.81			
Experiment 1	399	345	86.47			
Experiment 2	409	344	84.11			
Pythagorean Theorem – Finding the Hypotenuse 8.G.B.7	447	349	78.08	1	6.40	0.011*
Control	237	174	73.42			
Experiment	210	175	83.33			
Solving 1-Step Equations 7.EE.B.4a	928	818	88.15	1	0.01	0.934
Control	459	405	88.24			
Experiment	469	413	88.06			
Multiplying Simple Fractions 5.NF.B.4a	598	559	93.48	3	1.54	0.673
Group A – No Choice + Text	142	131	92.25			
Group B – Choice + Text	222	211	95.05			
Group C – Choice + Video	76	71	93.42			
Group D – No Choice + Video	158	146	92.41			
Rotations 8.G.A.3	306	186	60.78	1	0.82	0.365
Experiment 1	145	92	63.45			
Experiment 2	161	94	58.39			
Reflections 8.G.A.3	239	171	71.55	1	0.17	0.680
Experiment 1	125	88	70.40			
Experiment 2	114	83	72.81			
Order of Operations (No Exponents) 7.NS.A.3	1231	1172	95.21	2	4.50	0.105
Group A - Consistent/Neutral	597	574	96.15			
Group B - Inconsistent	300	287	95.67			
Group C - Mixed	334	311	93.11			

Note. †p < .10, *p < .05, **p < .01

posttest information, only ALI's selection bias assessments are presented for consideration at scale (see Table 2.5).

The 25 studies presented in Table 2.5 span a variety of investigations including: assessing the effect of various types of video tutoring (i.e., pencasts, teacher recorded instruction, online resources) compared to traditional text-based tutoring across multiple designs (i.e., using scaffolding, using hints, and as an intervention to wheelspinning (Beck & Gong, 2013), or provided based on student choice), investigating the manipulation of content (i.e., interspersing learning with humor through comics in content or feedback, asking students to gauge their confidence in solving problem content, and altering student mindset (Dweck, Chiu & Hong, 1995), and challenging cognitive principles (i.e., mental representations, and alterations in the consistency of math equations). The names of the assignments, as presented in Table 2.5, are tagged with the grade level and domain of the skill content as defined by Common Core State Standards (NGACBP & CCSSO, 2010). Despite differences in domain and experimentation, ALI is able to present a picture of condition quality across studies at scale.

The results of the simple chi-squared analyses in Table 2.5 may not seem significant at first but are actually quite insightful at scale. In studies with two conditions, experiment vs. control (20 comparable sets of the 25 shown in Table 2.5), the control group showed greater mastery in 15, while the experimental group showed greater mastery in only five. On its own, this comparison suggests that experimental conditions correlate with higher attrition rates. However, attrition significantly different than that of a normally distributed sample is only observed in five of the studies, with experimental conditions showing greater attrition than expected in 4/20 sets, and control conditions showing greater attrition than expected in only one set.

Even without considering posttest analyses, selection bias assessment at scale can lead to design and development guides for ASSISTments and can provide insights for the greater learning analytics community. As many of these studies were designed prior to the implementation of internal validity checks, it is possible that observed attrition was higher in experimental conditions because students were assigned to content that they had difficulty accessing. This finding would not likely be true of studies run more recently, and confirms the importance of the recent implementation of validity checks. Usability is a concern of any online learning system, and providing students with access to default assignments when they cannot access enriched content is a safe practice. Future iterations of ALI's at-scale capabilities will also be able to monitor attrition and posttest performance while controlling for particular student characteristics in order to assess the true variance established by these interventions.

It is also important to consider the percentage of students that were excluded from analysis prior to the automated assessments presented in Table 2.5. Within all sets, an average of 22.85% of students did not actually experience condition and were removed from the sample prior to analysis. Students that failed to experience interventions implemented within feedback provide valuable information regarding the raw (inflated) sample size required to achieve statistical power. Certain elements of a study's design, including the content domain chosen (i.e., some topics are easier than others and students require less feedback), and the type of feedback provided (i.e., on demand feedback requires a larger raw population than feedback provided automatically upon the student's incorrect response), can have a significant impact on the raw sample size required to attain a population of treated students powerful enough to reliably detect effects. This reduction in the sample size available for analysis also suggests that RCEs conducted within online learning platforms should consider interventions that can be implemented within problem content in order to boost sample populations.

Finally, analyzing the selection effects inherent to multiple assignments simultaneously allows ASSISTments to evolve more rapidly, providing benefits to users, researchers, and the learning analytics community. Considering that the experimental conditions shown in Table 2.5 showed approximately 1.5% greater attrition on average than control conditions, perhaps the potential benefits of these experimental interventions still outweigh the increase in attrition. Additional data mining would be necessary to determine a standard at which the potential for emphasized learning gains within an experimental condition no longer outweighed the potential for increased attrition. However, regularly conducting this type of broad scale analysis across assignments could isolate studies with conditions considered extremely detrimental, and the condition could be discontinued in order to limit the intervention's negative impact on students. ALI's automated analysis makes the process of content validation dramatically more efficient and robust. From these findings, and from future, more powerful iterations of ALI's at-scale capabilities, ASSISTments will be able to deliver rapid iterations of interventions with the goal of optimizing students' interactions with the system (by enhancing usability and by strengthening content and delivery methods).

2.5 Limitations & Future Work

ALI is constantly evolving and gaining new capabilities, and thus the version of the infrastructure presented here carries a number of limitations. As made apparent by the methods applied to consider ALI's effect at scale, ALI is currently only able to recognize studies with logged path information. The implementation of path logging occurred in March 2015, and ALI is currently only able to reliably analyze studies that were created after this date. This limitation is compounded by ALI's inference of the study design and posttest items. As studies can be designed using a number of complex, nested structures, it is possible that ALI's decisions about a study are not actually that intelligent. Thus, the infrastructure is currently only able to reliably recognize and analyze study designs with simple structures (i.e., "If-Then" routing, a single "Choose Condition," and a clear-cut posttest section that directly follows an intervention).

While these limitations seem to outweigh ALI's significance, they can easily be resolved through future work. One of the primary next steps for the team behind ALI is implementing a tagging system that will allow researchers to identify pertinent sections of a study prior to its distribution. Using unified naming structures for the design of assignment sections, researchers will essentially be able to tell ALI exactly how to approach analysis of their study. This will allow ALI to provide customized analysis and, potentially, refined data files that are preprocessed according to the researcher's distinct needs. Tagging will allow for analyses that collapse similar treatment groups, that isolate unconventional posttest items, and that assess growth models of student performance (i.e., pre-posttest gains).

Future work for the ALI team also includes defining a powerful list of student, class, and school level variables for use as covariates in statistical analyses. Variables that have already been established include measures of each student's prior performance within ASSISTments, measures of their completion rate on classwork and homework assignments, and normalized values that compare the student's performance and attrition against that of their class. Additional content will also be built into ASSISTments in order to collect self-report measures from students for use as possible covariates. Rich covariates will provide ALI with the ability to examine the effects of experimental interventions across groups while controlling for substantial variance, making automated analysis far more robust.

2.6 Contribution

The learning analytics community will benefit greatly from the effects of the Assessment of Learning Infrastructure and the promotion of similar systems built within online learning platforms. Currently, very few systems serve as tools for researchers to conduct and communicate the findings of sound scientific investigations into the subtleties of learning. By allowing researchers to conduct research within authentic learning environments through classwork and homework completed within online learning platforms, it is possible to collect rich log files that can be reported in universal formats and analyzed using automated processes. A strong focus on providing universal measures and analyses will strengthen the validity of findings from interventions that seek to isolate best practices in education. Further, the promotion of open, replicable, and sound science will benefit the future of learning analytics research.

References

Beck, J.E. & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In Lane, Yacef, Mostow & Pavlik (eds.) Proc of the 16th Int Conf on AIED. Springer-Verlag. pp. 431-440.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. J of Learning Sciences, 2(2), 141-178.

Dweck, C. S., Chiu, C., & Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A world from two perspectives. Psychological Inquiry, 6(4), 267-285.

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 of AIED. 24(4), pp.* 470-497.

Institute of Education Sciences. (2003). Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide. U.S. Dept of Ed. Washington, D.C.

Ioannidis J.P.A. (2005). Why most published research findings are false. *PLoS Med.* 2 (8): e124. Retrieved on 2/18/16 from http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124.

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of EDM*, *43*.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, *18*(1), 140-181.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). Common Core State Standards. Washington, DC: Authors.

Open Sci Collab. (2015). Estimating the reproducibility of psychological science. Science, 349 (6251).

Ostrow, K. (2015) Data for ALI Submission. Accessed from http://tiny.cc/LAK2016-ALI

Ostrow, K. & Heffernan, C. (2014). How to Create Controlled Experiments in ASSISTments. Retrieved from https://sites.google.com/site/assistmentstestbed/

Ostrow, K. S. & Heffernan, N. T. (2014). Testing the Multimedia Principle in the Real World: A Comparison of Video vs. Text Feedback in Authentic Middle School Math Assignments. In Stamper, et al. (eds.) Proc of the 7th Int Conf on EDM. pp. 296-299.

Ostrow, K., Heffernan, N., Heffernan, C., & Peterson, Z. (2015). Blocking vs. Interleaving: Examining Single-Session Effects within Middle School Math Homework. In Conati, Heffernan, Mitrovic, & Verdejo (eds.) Proc of the 17th Int Conf on AIED. Springer. pp. 388-347.

Pashler, H., Rohrer, D., Cepeda, N. & Carpenter, S.K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review. 14* (2), 187-193.

San Pedro, M., Baker, R., Gowda, S., & Heffernan, N. (2013). Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System. In Lane, Yacef, Mostow & Pavlik (eds.) Proc of the 16th Int Conf on AIED. Springer-Verlag. pp. 41-50.

Selent, D., Patikorn, T., Heffernan, N. & Beck, J. (Under Review). Unified Meta-Analysis: Using a single model to estimate treatment effects of multiple interventions. Submitted to the 3rd Annual ACM Conf on L@S.

Urbanek, S. (2003). Rserve—a fast way to provide R functionality to Applications. In Hornik, Leisch, & Zeileis, Proc of the 3rd Int Workshop on DSC, ISSN 1609-395X. http://rosuda.org/rserve.

U.S. Department of Education, Office of Educational Technology. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An Issue Brief. Washington, DC.

3 Collaborations for Iterative Refinement

Over the span of my Ph.D. candidacy I have been involved with the creation, analysis, and/or reporting of more than 90 studies using the ASSISTments TestBed. The majority of these studies were collaborations with other researchers, many external to the WPI ASSISTments Team. These opportunities for collaboration have helped guide the evolution of the ASSISTments TestBed and ALI through lessons learned and iterative refinement. In addition, I have been involved in a variety of collaborative educational data mining explorations using ASSISTments data to learn how to improve various aspects of ALI's reporting infrastructure.

This chapter includes the citations (**bold**) and abstracts (*italics*) of eight publications featuring my RCT and EDM collaborations (representing a balance of the two perspectives of my work). Each example is supplemented by an explanation of how the work influenced the evolution and/or advocacy of conducting educational research at scale using ASSISTments.

3.1 Randomized Controlled Trial Collaborations

Inventado, P. S., Scupelli, P., Ostrow, K., Heffernan, N., Ocumpaugh, J, Almeda, V., & Slater, S. (2018). Contextual factors affecting hint utility. *International Journal of STEM Education*. 5 (13): 1-13.

Background*. Interactive learning environments often provide help strategies to facilitate learning. Hints, for example, help students recall relevant concepts, identify mistakes, and make inferences. However, several studies have shown cases of ineffective help use. Findings from an initial study on the availability of hints in a mathematics problem-solving activity showed that early access to on-demand hints were linked to lack of performance improvements and longer completion times in students answering problems for summer work. The same experimental methodology was used in the present work with a different student sample population collected during the academic year to check for generalizability.*

Results*. Results from the academic year study showed that early access to on-demand-hints in an online mathematics assignment significantly improved student performance compared to students with later access to hints, which was not observed in the summer study. There were no differences in assignment completion time between conditions, which had been observed in the summer study and has been attributed to engagement in off-task activities. Although the summer and academic year studies were internally valid, there were significantly more students in the academic year study who did not complete their assignment. The sample populations differed significantly by student characteristics and external factors, possibly contributing to differences in the findings. Notable contextual factors that differed included prior knowledge, grade level, and assignment deadlines.*

Conclusion*. Contextual differences influence hint effectiveness. This work found varying results when the same experimental methodology was conducted on two separate sample populations* *engaged in different learning settings. Further work is needed, however, to better understand how on-demand hints generalize to other learning contexts. Despite its limitations, the study shows how randomized controlled trials can be used to better understand the effectiveness of instructional designs applied in online learning systems that cater to thousands of learners across diverse student populations. We hope to encourage additional research that will validate the effectiveness of instructional designs in different learning contexts, paving the way for the development of robust and generalizable designs.*

This journal article was an extension of a prior conference paper that had featured only a portion of the data. The addition of school year data (vs. summer data) revealed the significance of the data collection period in the observation and interpretation of intervention effects. Collaborators observed different results for summer work and school year samples. Essentially, students taking the Skill Builder as part of a multi-assignment "summer math packet" showed different tendencies than those completing the assignment as nightly homework with stricter deadlines, expectations, and consequences. These results signify that both temporal context of assignment completion and concurrent assignments may serve as powerful controls. This finding also revealed the risk of repeated testing through automated, weekly ALI reports. The ASSISTments Research team had worried about researchers' spot checking p-values and waiting for significance (i.e., the poor practice of cherry picking) but had not necessarily thought about ramifications for generalizability and reproducibility when considering the time frame of data collection. It is possible that an effect might be observed in school year data but washed out by summer data, posing an interesting area for future work to tease apart such effects. This finding helped the team decide to cease weekly reporting and fall back on data requests as the default protocol until further progress can be made on DIBBs template goals.

This work also helped the ASSISTments Research team realize that in-house design language and implementation is not universal. It is not easy for external researchers to create experimental flowcharts. The need for a visual environment when building is critical as a next step for the TestBed. The TestBed resource guide (www.ASSISTmentsTestBed.org) provides flowcharts and templates for common experimental designs but a disconnect remains when researchers attempt to implement their studies. The Template Tool and other developments that will be made possible by the DIBBs grant (see Chapter 4) have already been proposed to remedy this issue.

Finally, this work made it clear that experience with a system does not necessarily beget understanding. These researchers had been working together and collaborating within ASSISTments for a number of years. Despite having created and analyzed multiple studies, new questions would come up each week due to differences in data processing protocol across studies, complications in interpretation, or iterative design issues. These same issues make the concept of coding automated analyses to support ALI very difficult in an open research environment, making the DIBBs Template Tool a viable future goal.

McGuire, P., Tu, S., Logue, M. E., Mason, C., & Ostrow, K. (2017). Counterintuitive Effects of Online Feedback in Middle School Math: Results from a Randomized Controlled Trial in ASSISTments. *Educational Media International*. 54 (3): 231-244.

This study compared the effects of three different feedback formats provided to sixth grade mathematics students within a web-based online learning platform, ASSISTments. A sample of 196 students were randomly assigned to one of three conditions: (1) text-based feedback; (2) image-based feedback; and (3) correctness only feedback. Regardless of condition, students solved a set of problems pertaining to the division of fractions by fractions. This mathematics content was representative of challenging sixth grade mathematics Common Core State Standard (6.NS.A.1). Students randomly assigned to receive text-based feedback (Condition A) or image-based feedback (Condition B) outperformed those randomly assigned to the correctness only group (Condition C). However, these differences were not statistically significant (F(2,108) = 1.394, p = .25). Results of this study also demonstrated a completion-bias. Students randomly assigned to Condition B were less likely to complete the problem set than those assigned to Conditions A and C. To conclude, we discuss the counterintuitive findings observed in this study and implications related to developing and implementing feedback in online learning environments for middle school mathematics.

This journal article involved a study conducted early on in my graduate career that investigated how feedback mediums influence learning outcomes. Collaboration with this team has spanned many years and the article took multiple years to construct and publish. This project made it clear that external researchers, especially those with less experience using vast datasets, have trouble interpreting ALI files without the assistance of the ASSISTments Research team. A glossary established during the development of the Testbed was revised and made more accessible during this project. The range of covariates available in ALI files was also expanded to provide researchers with more potential controls. It remains clear that a set of tutorials or a richer manual supporting ALI data files would be helpful for external researchers seeking variable interpretation.

This study also made clear that results, and how researchers interpret and discuss them, may vary considerably based on available data. Data from this study was passed around for more than a year before the journal paper was written and interpretations changed and grew with time. This back and forth made it clear that the accuracy, robustness, and objectivity of ALI files is critical to researchers' interpretations. If researchers retrieve data that they do not fully understand, they are much more likely to run nonsensical tests or draw inaccurate conclusions, producing work that may not provide valid value to the field. This concept of data flexibility grew to form the basis of a *Journal of Learning Analytics* paper (Ostrow, Wang, & Heffernan, 2017) that examined how conclusions drawn from data can change considerably with variations in preprocessing and feature generation.

The final major lesson learned from this collaboration was that it is critical for the ASSISTments Research team to let go and place control in the hands of external researchers. Given the "expert blind spot" of those working with the system daily and understanding its intricacies, it can be difficult to guide external researchers through the process without "taking the wheel." If the idea of a shared scientific instrument is to be successful, those on the ASSISTments Research team must maintain roles as consultants, supplementing the research efforts of external researchers without guiding the story or offering full interpretation. The need to "let go" also impacts the bottleneck discussed in the DIBBs grant proposal (see Chapter 4). Members of the ASSISTments Research team currently maintain responsibilities at each step of the TestBed process, a routine that is impossible to maintain if the end goal is to properly scale the tool. By devising the infrastructure necessary to step back and maintain a consultant's perspective, members of the ASSISTments Research team will avoid driving the research endeavors of external researchers, ensuring that outcomes and interpretations remain objective.

Lang, C., Heffernan, N., Ostrow, K., & Wang, Y. (2015). The Impact of Incorporating Student Confidence Items into an Intelligent Tutor: A Randomized Controlled Trial. In Santos, Boticario, Romero, Pechenizkiy, Merceron, Mitros, Luna, Mihaescu, Moreno, Hershkovitz, Ventura, & Desmarais (Eds.) Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015). Madrid, Spain. June 26-29. 144-149.

For at least the last century researchers have advocated the use of student confidence as a form of educational assessment and the growth of online and mobile educational software has made the implementation of this measurement far easier. The following short paper discusses our first study of the dynamics of student confidence in an online math tutor. We used a randomized controlled trial to test whether asking students about their confidence while using an Intelligent Tutor altered their performance. We observe that (1) Asking students about their confidence has no statistically significant impact on any of several performance measures (2) Student confidence is more easily reduced by negative feedback (being incorrect) than increased by positive feedback (being correct) and (3) confidence accuracy may be a useful predictor of student behavior. This paper demonstrates how psychological ideas can be imported into Educational Data Mining and our findings point to the possibility of using student confidence to better predict performance and differentiate between students based on the way they approach items.

This work on student confidence was the first of its kind within ASSISTments and one of the first ventures into the study of how self-reported socio-emotional constructs influence learning outcomes within a single math assignment. The collaboration revealed that asking students about their confidence did not alter their performance and that it was easier to shatter confidence through negative feedback than to build it up with positive feedback. The study suggested that confidence estimates are also potentially helpful covariates for estimating student performance as later shown in Chapter 9. It is critical that the designers of ASSISTments, and those of similar platforms, remember that while structure and focus on teachers are important elements for recruitment, the vast majority of users are children. Their socio-emotional well-being can alter their learning outcomes, and extending beyond cognitive measures, expectations, and barriers is an important step for ASSISTments Research.

Another lesson learned from this collaboration was that ALI files should be better organized or easier to manipulate based on researchers' predefined experimental structures. In this work it was difficult to parse math content from confidence content in student-level and problem-level file structures. Collaborators brought programming skills and patience to the work, which should not necessarily be asked of all external researchers. This issue grows more complicated in the presence of posttest content following a Skill Builder as one student may begin the posttest on their 6th problem opportunity while another begins on their 10th, causing complex disorganization in resulting files. The "Problem Set Organizer" created by Anthony Botelho offers a first step in remedying this issue. The tool allows researchers to specify the structure of their research design before using ALI's problem-level and covariate files as input to provide a structured student-level output file for easier interpretation. This tool will be integrated into the DIBBs infrastructure in the coming months to provide greater ease of use to external researchers working with ALI's data files.

This work also revealed that it is difficult to cater to all researchers when the research environment is open. The files that ALI provides are robust and universal, but they are not optimally tailored to all analyses because external researchers often bring skills and methodologies to the table that members of the ASSISTments Research team did not anticipate when designing ALI's output. When this work was published in 2015, ALI offered fewer file types. This work offered one instance of the need for more versatile output. Essentially, the constraints of ALI's data, and the ASSISTments database for that matter, limit the analyses that researchers can conduct. Striving to strengthen ALI's output files and how researchers can access and manipulate them will help to broaden the capacity for sound research.

This work was also one of the first recorded instances of external researchers exploiting system features or bugs to push the limits of their experimental designs. Since this collaboration, pushing boundaries has become a common theme among external researchers hoping for compromises between what their research designs demand and what ASSISTments can achieve. The extension was relatively simple in this study: change the Skill Builder to demand four right in a row for mastery, but place confidence gauges (or control questions) with all accurate answers as every fourth problem. This pattern would allow students to complete the assignment with any pattern of three math problems right in a row, thus functioning like a traditional Skill Builder. It is important that the ASSISTments Research team be prepared for researchers to use the system in unexpected ways. When each new collaboration is thinking outside the box or moving the goal post, the only limitation becomes how fast the system itself and its output can grow (and deciding whether it always should). It is possible that this trend will continue after confining the TestBed's research design capacity through the proposed DIBBs template tool, which will be interesting to assess in due time.

Ostrow, K., Heffernan, N. T., Heffernan, C., & Peterson, Z. (2015). Blocking vs., Interleaving: Examining Single-Session Effects within Middle School Math Homework. In Conati, Heffernan, Mitrovic & Verdejo (Eds.) Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015). Springer International Publishing. Madrid, Spain, June 22-26. 388-347.

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 work was largely internal to the ASSISTments Research team (with the inclusion of a summer research intern) but is included here as collaborative work inspiring iterative improvement because of lessons learned from its unique "orchestrated" design. The work had a complex longitudinal design that required the authors to work closely with participating teachers to assign multiple problem sets, with the posttest assigned days later as a separate assignment. The findings were interesting and aligned with theoretical underpinnings of spacing and interleaving. However, it is difficult to achieve orchestrated designs with external researchers unless they wish to use a private sample population. Many researchers request delayed posttests, or similar features that would allow them to propagate assigned conditions across multiple assignments. Future goals to strengthen capacity for this type of design are included in the redesign of the ASSISTments tutor (to be released in 2018) and align well with plans for restructuring research designs as proposed by the DIBBs grant.

This work also displayed the top-down practice of considering a theoretical principle and testing it specifically, rather than the bottom-up practice of establishing a learning invention and then finding concurrent theory for its support. Both approaches have been used for research within the ASSISTments TestBed and supporting the transparency of research designs and processes would clarify and strengthen reporting. The ASSISTments Research team supports the practices of the Open Science Foundation in preregistration, transparent practices, and accessible data. As such, requiring that researchers log their designs and progress with the OSF (https://osf.io), or similar platforms, should increase the replicability and generalizability of studies and their findings.

This work was also crucial in revealing the importance of allowing students to learn during posttests or similar assessments. While it can add noise to scores or accuracy, the information obtained by considering metrics like hint and attempt count or first response time can alter results and interpretations. This work compared the observation of a null effect using univariate approaches with significant effects using multivariate approaches, giving way to the potential importance of partial credit and the consideration of robust performance measures within ASSISTments. It also helped to inspire later work on data flexibility and group differentiation (Wang, Ostrow, Beck, & Heffernan, 2016; Ostrow, Wang, & Heffernan, 2017).

Finally, this work was one of the first requiring merged ALI data files. I authored two pieces of code (A.2 and A.3 in the Appendix) to merge datasets either left-to-right (i.e., concatenating student data from a posttest or a second assignment by adding new columns appended to existing student-level data) or top-to-bottom (i.e., merging assignments into a single file by adding new rows appended to the original file). These programs were basic tools that set the stage for more robust code like the Problem Set Organizer.

Williams J. J., Ostrow, K., Xiong, X., Glassman, E., Kim, J., Maldonado, S., Li, N., Reich, J., & Heffernan, N. (2015). Using and Designing Platforms for In Vivo Educational Experiments. In Russell, D.M., Woolf, B., & Kiczales, G. (Eds.), Proceedings of the 2nd ACM Conference on Learning at Scale (L@S 2015). Vancouver, British Columbia, March 14-15. 409-412.

In contrast to typical laboratory experiments, the everyday use of online educational resources by large populations and the prevalence of software infrastructure for A/B testing leads us to consider how platforms can embed in vivo experiments that do not merely support research, but ensure practical improvements to their educational components. Examples are presented of randomized experimental comparisons conducted by subsets of the authors in three widely used online educational platforms – Khan Academy, edX, and ASSISTments. We suggest design principles for platform technology to support randomized experiments that lead to practical improvements – enabling Iterative Improvement and Collaborative Work – and explain the benefit of their implementation by WPI co-authors in the ASSISTments platform.

This work provided early evidence that ASSISTments serves a unique role for ushering digital learning environments into the research age. The paper emphasized that ASSISTments' capacity for A/B testing in authentic learning environments was special and that it provided research opportunities beyond those observed in other popular platforms. In hindsight, the question that this paper raises is whether capacity for personalization through collaborative iterative work is realistic. Ideas like PeerASSIST and TeacherASSIST that bank on the efficacy of crowdsourcing and the promise of AI driven content delivery have been discussed for three years with minimal implementation. Of course, this is an exceptionally difficult task, and it remains a focus of future work within the 'new world' of ASSISTments, the new generation of the tutor to be released in 2018. The task is so hard in fact, that many in the field are chasing their own red herrings of personalization. Given that the ASSISTments Research team is largely comprised of graduate students focused on projects lasting two to five years, it is important to consider the longevity of various ASSISTments features, and even the lasting potential of the ASSISTments TestBed and ALI. Making sure successors are in place to carry important projects will be key.

This work also highlighted the potential influences of scaled classroom research. Authentic learning environments offer less controlled but more generalizable contexts for study. These environments make it more difficult to observe effects and isolate significant interventions, but when uncovered, effects are usually more generalizable and of greater practical importance than those discovered in strictly controlled laboratory settings. The use of student-level randomization enhances these analyses by adding power and controlling for school effects that are traditionally observed in class- or school-level randomization used in the majority of educational research. In addition, student-level randomization is thought to allow for data to be collected with greater efficiency. Future work should assess assumptions surrounding the efficacy of student-level randomization by comparing results obtained from common models with more robust hierarchal models.

This article also paralleled the call of the SI2 grant, with the goal of promoting the ASSISTments TestBed as a shared scientific instrument. Considering ASSISTments' unique role in supporting educational research at scale, it is important that the ASSISTments Research team continues to advocate for transitions toward sound, accessible, and transparent research practices in online learning. Part II of this dissertation made it clear that the message of ASSISTments Research should transition from "authentic research at scale should be the new default" to "this is a powerful method for research with its own limitations." Still, advocacy should continue with caution as these tools offer a viable alternative to traditional methods.

3.2 Educational Data Mining Collaborations

Ostrow, K. S., Wang, Y., & Heffernan, N. T. (2017). How Flexible Is Your Data? A Comparative Analysis of Scoring Methodologies Across Learning Platforms in the Context of Group Differentiation. *Journal of Learning Analytics*. 4 (2): 91-112.

Data is flexible in that it is molded not only by the features and variables available to a researcher for analysis and interpretation, but also by how those features and variables are recorded and processed prior to evaluation. Big Data stemming from online learning platforms and intelligent tutoring systems is no different. The present work questions the quality and flexibility of data from two popular learning platforms, comparing the scoring method typically used to inform learner analytics (binary measures of problem-level accuracy) with partial credit, a more robust, real-world methodology. This work extends previous research in the field by examining how the manipulation of scoring methodology has the potential to alter outcomes in the context of hypothesis testing, or specifically, when looking for significant differences between groups of students. Datasets from ASSISTments and Cognitive Tutor are used to assess the implications of data availability and manipulation within twelve mathematics skills. A resampling approach is employed to determine the size of equivalent samples of high- and low- performing students

required to reliably differentiate between the two groups using each scoring methodology. Results suggest that in eleven out of twelve observed skills, partial credit offers more efficient group differentiation, increasing analytic power and reducing Type II error. Alternative applications of this approach and implications for the Learning Analytics community are discussed.

This work was an extension of previous work on using partial credit to enhance the efficiency and reliability of group differentiation (Wang, Ostrow, Beck, & Heffernan, 2016). The original work used both experimental and non-experimental data from ASSISTments to address how more robust student data offers stronger differentiation between groups of students. Partial credit, or performance data supplemented by students' hint and attempt usage, led to successful group differentiation with smaller sample sizes and greater reliability. Essentially, it improves the power.

For this work, the explanation of the effect expanded to consider the concept of 'data flexibility.' This concept integrates many of the lessons learned from past work situated in both experimental design and educational datamining. The concept of partial credit, or using student covariates to strengthen the observation of experimental effects, has been prevalent in much of the recent work conducted within ASSISTments (some examples include: Adjei, Ostrow, Erickson, & Heffernan, 2017; Wang, Ostrow, Beck, & Heffernan, 2016; Ostrow, Donnelly, Adjei, & Heffernan, 2015; Ostrow, Heffernan, Heffernan, & Peterson, 2015). By phrasing the issue as one of data availability, the argument for careful data preprocessing and feature generation became a critical talking point that was not previously well discussed. Analyses are limited by available data and outcomes are merely extrapolations of their inputs. The influx of educational data made possible by online learning platforms has fed the capacity for a wealth of analytics and inferences about best practices in learning that were not achievable from traditional classroom data. However, the creators of online learning platforms are cautioned that the type, quality, and quantity of data they make available to researchers can significantly influence findings and interpretations.

For the ASSISTments TestBed and ALI, this realization means that the team should remain cognizant of their impact in the future of these tools. By continuing to strive toward stronger feature generation techniques and a greater pool of student, class, and school level covariates, improving ALI's datasets affords researchers the capacity for stronger and more diverse methodologies that allow for more accurate and generalizable findings and interpretations. The team has often made comparisons between ASSISTments as a 'freely available scientific instrument' and popular scientific tools like the VLT in Chile that allows researchers from around the world to explore space, or supercomputers that offer shared time for the analysis of massive datasets like the human genome. However, it is not enough to open the platform as a shared scientific instrument: the team must calibrate the instrument's output to strengthen and unify scientific contributions. If the output of a telescope is blurry, or the findings from a supercomputer's analysis are off by a factor of 10, findings would lack practical importance and limit future hypothesis growth. The clarity and accuracy of ALI's output should be considered in a similar manner, and the ASSISTments Research team should continue advocating for strong reporting standards in other platforms that begin to support educational research at scale.

Adjei, S., Ostrow, K., Erickson, E., Heffernan, N. (2017). Clustering Students in ASSISTments: Exploring System- and School-Level Traits to Advance Personalization. In Proceedings of the 10th International Conference on Educational Data Mining. 340-341.

Few attempts have been made to create student models that cluster student and school level traits as a means to design personalized learning interventions. In the present work, data from ASSISTments was enriched with publicly available school level data and K-Means clustering was employed. Results revealed the importance of school locale, measures of district wealth, and system interaction patterns as potential foci for personalization. Clusters were then applied to a test set of held out data and cluster assignments were used to help predict end-of-year standardized mathematics test scores. Findings suggest that while cluster interpretations were not generalizable to held out data, clustering was generally helpful in predicting standardized test scores.

This work was an extension of investigations into partial credit and data flexibility (Wang, Ostrow, Beck, & Heffernan, 2016; Ostrow, Wang, & Heffernan, 2017) meant to assess the value of available covariates in defining groups of students within ASSISTments. The clustering task considered a set of variables typically reported in ALI files along with supplementary school-level data regarding SES, school size, and student-to-teacher ratios. Findings suggested that these variables may be important in defining student types and that the added information is helpful in predicting performance outcomes. However, the results were not generalizable, suggesting that student variation is potentially more complex than current covariates allow, or suggesting the need for hierarchical approaches.

Still, this work made it clear that the covariates provided by ALI reports alter the sensitivity of potential analyses and that it is important that the ASSISTments Research team continues its data mining endeavors to expand upon the data made available to external researchers. In cases where it is possible to support ALI files with external data (i.e., school-level information made available by the NCES) it is important to consider making these resources available to researchers. However, it is also important to maintain a balance between the amount of information provided and the privacy protection of students and teachers.

Eventually, work in this area could typify learners within and across assignments, providing student labels that could serve as higher level covariates. ALI could include output from this type of clustering analysis and ‘student type’ variables might serve as mediators or moderators, ideally uncovering the interaction effects and mechanisms underlying effective interventions.

Ostrow, K., Donnelly, C., Adjei, S., & Heffernan, N. (2015). Improving Student Modeling Through Partial Credit and Problem Difficulty. In Russell, D.M., Woolf, B., & Kiczales, G. (Eds), Proceedings of the 2nd ACM Conference on Learning at Scale (L@S 2015). Vancouver, British Columbia, March 14-15. 11-20.

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.*

This work, while internal to the ASSISTments Research team, was important to the establishment of partial credit within the tutor. Prior to this work, problems could only be scored using first response binary correctness (0 = *Incorrect*, 1 = *Correct*). If students requested a hint, they would be marked incorrect and could easily lose motivation for completing the problem. This work used a simple tabling approach to reveal that partial credit scoring could strengthen student performance models while serving as a potentially motivating intervention for students. This work was conducted in the same era as the blocking vs. interleaving RCT (Ostrow, Heffernan, Heffernan, & Peterson, 2015) that showed partial credit approaches could also be influential in assessing the efficacy of learning interventions and isolating group differences. Partial credit was implemented as a teacher-level setting, allowing teachers a more lenient grading option for their students' assignments and promoting student use of system features like hint and scaffolding support.

This work also supported allowing students to learn during posttests or similar assessments. This finding was originally revealed by RCT work (Ostrow, Heffernan, Heffernan, & Peterson, 2015). Essentially, predictive capacities of partial credit scores measured during an assignment suggest value gains similar to those observed by including students' hint and attempt usage within assessments. As partial credit is a familiar practice for most teachers, the scoring methodology also helps to enhance the authenticity of the learning environment and to ensure that research practices are minimally invasive.

In the sense that more information about student performance allows for greater accuracy in student modeling, this work was also influential in establishing an influx of covariates into ALI's data files. Students' historical data (collected from their system usage prior to their participation in a randomized controlled trial) was separated into a clear and concise covariate report. This file can be merged into any other data granularity offered by ALI to enhance analyses and provide descriptive statistics about the sample population. Still, it is important to remember that the use of too many covariates would constitute overfitting, leaving models lacking generalizability.

References

Adjei, S., Ostrow, K., Erickson, E., Heffernan, N. (2017). Clustering Students in ASSISTments: Exploring System- and School-Level Traits to Advance Personalization. In Proceedings of the 10th International Conference on Educational Data Mining. 340-341.

Inventado, P.S., Scupelli, P., Ostrow, K., Heffernan, N., Ocumpaugh, J, Almeda, V., & Slater, S. (2018). Contextual factors affecting hint utility. *International Journal of STEM Education*. 5 (13): 1-13.

Lang, C., Heffernan, N., Ostrow, K., & Wang, Y. (2015). The Impact of Incorporating Student Confidence Items into an Intelligent Tutor: A Randomized Controlled Trial. In Santos, Boticario, Romero, Pechenizkiy, Merceron, Mitros, Luna, Mihaescu, Moreno, Hershkovitz, Ventura, & Desmarais (eds.) Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015). Madrid, Spain. June 26-29. 144-149.

McGuire, P., Tu, S., Logue, M.E., Mason, C., & *Ostrow, K.* (2017). Counterintuitive Effects of Online Feedback in Middle School Math: Results from a Randomized Controlled Trial in ASSISTments. Educational Media International. 1-14.

Ostrow, K., Donnelly, C., Adjei, S., & Heffernan, N. (2015). Improving Student Modeling Through Partial Credit and Problem Difficulty. In Russell, D.M., Woolf, B., & Kiczales, G. (eds), Proceedings of the 2nd ACM Conference on Learning at Scale (L@S 2015). Vancouver, British Columbia, March 14-15. 11-20.

Ostrow, K., Heffernan, N.T., Heffernan, C., & Peterson, Z. (2015). Blocking vs., Interleaving: Examining Single-Session Effects within Middle School Math Homework. In Conati, Heffernan, Mitrovic & Verdejo (eds.) Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015). Springer International Publishing. Madrid, Spain, June 22-26. 388-347.

Ostrow, K.S., Wang, Y., & Heffernan, N.T. (2017). How Flexible Is Your Data? A Comparative Analysis of Scoring Methodologies Across Learning Platforms in the Context of Group Differentiation. *Journal of Learning Analytics*.

Wang, Y., Ostrow, K., Beck, J., & Heffernan, N. (2016). Enhancing the Efficiency and Reliability of Group Differentiation through Partial Credit. In the Proceedings of the 6th International Conference on Learning Analytics & Knowledge, 454-458. ACM.

Williams J.J., Ostrow, K., Xiong, X., Glassman, E., Kim, J., Maldonado, S., Li, N., Reich, J., & Heffernan, N. (2015). Using and Designing Platforms for In Vivo Educational Experiments. In Russell, D.M., Woolf, B., & Kiczales, G. (eds.), Proceedings of the 2nd ACM Conference on Learning at Scale (L@S 2015). Vancouver, British Columbia, March 14-15. 409-412.

4 Looking Forward: NSF Funded DIBBs Grant

This chapter provides the grant proposal for the NSF funded Data Infrastructure Building Blocks (DIBBs) grant (1724889 - CIF21 DIBBs PD: Enhancing and Personalizing Educational Resources through Tools for Experimentation, $494,644) that was awarded to PI Neil Heffernan. The grant is projected to span three years (2017-2020) and the work encompassed will guide the future of ASSISTments Research through drastic improvements to the protocol underlying the ASSISTments TestBed and the Assessment of Learning Infrastructure.

This grant proposal is included in Part I because I led the effort to secure funding and I hope to join the ASSISTments team upon completion of my Ph.D. as a Research Scientist to serve as a Co-PI on the grant and drive the immediate future of ASSISTments Research in other capacities through continued advocacy, research publication, and evolution. Thus, it serves to close Part I by framing my vision for how the tools that comprise the foundation of ASSISTments Research will continue to grow and their potential influence on the future of the Learning Sciences.

Proper citation of this chapter is as follows:

CIF21 DIBBs PD, National Science Foundation. (2017). Enhancing and Personalizing Educational Resources through Tools for Experimentation. Award # 1724889. Abstract available at: http://www.nsf.gov/awardsearch/showAward?AWD_ID=1724889.

4.1 The Problem: Obstacles in Understanding Best Practices in Education

4.1.1 The Gold Standard in Causality is Currently Difficult to Achieve in Authentic Learning Environments

Educational psychologists, researchers, and practitioners have grown accustomed to the complex and time-consuming nature of studying effective classroom practices. When studying learning interventions, the *gold standard* should be used in determining causality: the randomized controlled experiment (RCE). Yet despite recent calls encouraging the use of RCEs within authentic learning environments in the K12 arena, to apply findings from cognitive science to improve educational practices (IES, 2013; Pashler et al., 2007), and despite the nearly infinite array of complexities to be examined within the context of instruction (Koedinger, Booth, & Klahr, 2013) RCE's are not often used. RCEs can be difficult to conduct in real-world classrooms (National Research Council, 2002). Common complications include Institutional Review Board (IRB) restrictions, difficulties maintaining relationships with schools in order to conduct studies with lengthy and invasive Pre- and Posttests, and large sample populations required to detect significantly reliable results.

4.1.2 Education Research at Scale is Time Consuming and Costly

When quantifying the time required to conduct large-scale RCEs, consider that some of the best educational researchers only generate one or two experiments per year. Time is invested up-front through the design of an experiment, creation or adjustment of educational materials, and thoroughly vetting the design for flaws (lest the experiment fall apart in the classroom, or the data retrieved be inadequate). When the researcher considers the experiment ready to run, additional time is required to work through an IRB review. Slow reviews from an IRB can delay a curriculum-linked intervention by a full school year. Meanwhile, researchers pour time into repeated school visits, developing and maintaining relationships with teachers and balancing a schedule for conducting their research with the teachers' curriculum goals. When an experiment finally comes together and is actively run within classrooms, researchers often step in to deliver their content and remove the potential for teacher bias. The researcher must organize random assignment of students or classes into experimental conditions, and often, track that students remain in condition across multiple assignments or tasks. Implementing robust measures of student knowledge, like delayed reassessments, require the orchestration of additional site visits and data collection. Even speaking in such broad terms, the time investment required to conduct a large-scale educational experiment is obvious and massive.

Attempts to quantify the costs required to conduct large-scale RCEs paint a similar picture. In their book addressing the scarcity of RCEs in education, Mosteller and Barouch (2002) highlight the considerations education researchers make that tend to drive up the cost of research. They note that researchers care deeply about whole-group instruction and classroom management, calling for teacher or class to serve as the unit of analysis. While this design is considerate to teachers and their curricula, it poses high costs to achieve adequate statistical power (i.e., establishing a robust sample of many teachers and classrooms) for analysis of effects. To actually apply a dollar amount to some of this work, consider that the U.S. Department of Education funds Efficacy Trials that examine promising interventions at the school level using a sample of about 50 schools and costing approximately $3 million. At a higher echelon, Effectiveness Trials are more rigorous and are valued at a median cost of $6 million. One would hope that with the government pouring millions of dollars into this work, researchers would be learning a great deal about best practices in education. Unfortunately, of 22 Efficacy Trials and five Effectiveness Trials funded in the realms of math and science, few show promising results. The U.S. Department of Education set up the What Works Clearinghouse to apply rigorous standards about what works, and thus far, for mathematics, only five interventions have qualified with "reliably positive effects." If we can make testing ideas faster and less expensive, we may end up funding more studies that work.

4.1.3 Crisis of Faith in the Scientific Method: Results Fall Short of Reproducibility

The obstacles continue beyond conducting the research, with issues that muddy the waters of reporting findings including publication bias and the cherry picking of results that can lead to the inability for replication, contributing to a growing crisis of faith in RCEs spanning numerous scientific fields (Ioannidis, 2005; Achenbach, 2015; Open Science Collaboration, 2015). Ioannidis' (2005) provocatively named paper, "Why Most Published Research Findings Are False," set off a firestorm of debate. He argues persuasively that things like publication bias (null results tend not to be reported) as well as researchers well-meaning selective picking of samples to analyze, can cause havoc in the library of published results. This resulted in a great deal of recent press in the NY Times, Economist, NPR and Wall Street Journal (Johnson, 2014 Jan; Johnson, 2014 Mar; Broad, 2014 Mar; Lombrozo, 2014 June; The Economist, 2013, Oct). Amgen Inc. tried to replicate the findings of 53 "landmark" cancer studies, but only succeeded in replicating six (Begley & Ellis, 2012). Bayer Inc. made a similar attempt with drug studies; of the 67 results deemed important enough for replication, it could only be achieved in one quarter of studies (Prinz, Schlange, & Asadullah, 2011). Failure to replicate a finding does not mean it is false, but with large numbers, one should have the power to detect such results. This is what makes a recent attempt in Psychology to replicate 13 landmark studies so unique. Dozens of labs replicated the same experiments, creating tremendous statistical power, and in conclusion, showed that two of the 13 seminal findings were simply false (Yong, 2013, November). But with hundreds of RCEs published each year, the cost of multi-lab replication is too high to combat errors of commission and omission in research.

The replication crisis is not the only obstacle currently plaguing scientific fields. Researchers are also not as open about their data and materials as we would like them to be. Ioannidis and colleagues (Alsheikh-Ali et al., 2011) report that in the top 50 leading journals with data-sharing policies, only 143 of the 351 papers he sampled complied with those policies (Alsheikh-Ali et al., 2011).

4.2 The Opportunity: What We Can Leverage to Address These Problems

In the previous section, we discussed a set of problems, while in this section we will discuss a set of opportunities that make the present a good time to fund our proposed project. Only after discussing these opportunities will we lay out our solution that will leverage these opportunities to solve the above-listed problems.

4.2.1 The Growth of Educational Technologies

The U.S. Department of Education's National Educational Technology Plan (2010) suggests that technology is going to be a key driver in delivering interventions to students, and that those interventions will personalize the learning experience. Interestingly, the plan does not mention how technology can also be a key driver in reforming the scientific process we use to study learning. As more and more students use technology at scale, data collected from these platforms can be leveraged in dynamic ways that may reveal revolutionary insights about learning. Entire fields of research are growing alongside educational technologies in hopes of better understanding how these tools and their data can be used to improve education (e.g., Learning Analytics, Educational Data Mining). However, despite significant growth in researcher interest, few technologies, that are currently available to teachers and students, allow real-time hypothesis testing. There is an opportunity to extend "Big Data" to "Big Experimentation" allowing researchers to begin answering three questions that truly drive personalized education: "What works best? For whom? And under what circumstances?"

We intend to support research communities that do not normally use technology to begin to benefit greatly from harnessing the power of educational technologies to simplify the process of conducting RCEs within authentic learning environments, making research at scale more feasible, accessible, and accurate. The education community does not currently have shared scientific instruments, such as the Green Bank Radio Telescope or the Hubble Telescope, to be used for the benefit of efficient and collaborative research. The modest investment of a DIBBs **Pilot Demonstrations** grant would make possible the promotion of such an instrument that harnesses a popular piece of educational technology to be used to run RCEs that do not interfere with classroom assignments but allow for researchers to learn what works, for whom, and when.

4.2.2 ASSISTments: A Brief Overview

The specific opportunity at hand, and the basis of this proposal, leverages ASSISTments (www.ASSISTments.org), a popular online learning platform that was designed with the flexibility to run behind-the-scenes RCEs. This has supported numerous research publications from within the ASSISTments team since its inception in 2002. A brief overview of ASSISTments as a teaching tool helps to guide understanding of how the platform is leveraged for student learning and for research. ASSISTments is used by hundreds of teachers spanning more than 500 schools who assign classwork and homework to over 50,000 students, generating over 10 million problems solved in the 2014 – 2015 school year. Features in national press, such as NPR and the NY Times (Heffernan, 2014), have helped teachers learn about ASSISTments. There are over 25,000 problems in ASSISTments written at Worcester Polytechnic Institute (WPI) that include substantial tutoring, and over 100,000 problems that have been written by teachers. Nearly all of the current NSF-funded content is for middle-school math, with teachers from domains outside of math building their own content for their courses. The platform, offered as a free service of WPI, is a tool that provides students with instructional *ASSIST*ance while offering teachers assess*MENT* reports*,* establishing "ASSISTments" as our moniker. Teachers can select content from a problem set library or build their own.

In ASSISTments, content is built at the problem level. The problem builder enables teachers and researchers to design *"problems,"* a combination of questions and support using a simple interface for the inclusion of text, graphics, and hypermedia elements. Problems can then be combined to form *"problem sets"* for assignment to students. Teachers are then responsible for assigning the problem sets and monitoring student completion. Use of ASSISTments has been shown to reliably improve students' learning in comparison to traditional paper and pencil approaches (Mendicino, Razzaq, Heffernan, 2009; Koedinger, McLaughlin & Heffernan, 2010; Singh et al., 2011; Kelly, Heffernan, Heffernan, et al., 2013; Miller, et al., 2013; Soffer, et al., 2014). Most recently, SRI International reported results of an Efficacy Trial of ASSISTments, showing that use of the platform for homework over the course of a year resulted in large, reliable, learning gains on standardized assessments (Roschelle, et al., 2016).

There are two primary types of problem sets within ASSISTments. A *Complete All* problem set has a predetermined number of problems and the assignment is considered complete when the student has finished all problems, whether or not the answers are accurate. Alternatively, in a *Skill Builder* problem set, students must solve problems selected at random from a question pool until reaching a predetermined threshold of mastery, which normally involves answering three consecutive questions accurately on first attempts. Mastery can be redefined to include any number of consecutive accurate problems.

4.2.3 Results from Prior NSF Support: in Creating ASSISTments

ASSISTments was created while conducting research for NSF awards. Heffernan's NSF CAREER award ("Learning about Learning" award #0448319, $646,075, 2006 – 2013) is the most relevant grant that helped create ASSISTments. The **intellectual merit** includes more than four dozen peer-reviewed publications in learner analytics (see the separate section in references for the publications). Other NSF grants have also supported ASSISTments, including grant #0742503 whose **intellectual merit** includes the 24 published RCEs that compared different ways to provide feedback to students (see the separate section in the references for these 24). Both of these grants had the **broader impact** of helping 50,000 students receive better instruction in mathematics.

4.2.4 Results from Prior NSF Support: A NSF SI2 pilot testing the TestBed Idea

The goals presented in this proposal intend to build upon a concept established by a third NSF grant (#1440753) that helped to launch a formal infrastructure supporting external researchers to leverage ASSISTments as a shared scientific tool. The ASSISTments TestBed (www.ASSISTmentsTestBed.org) allows researchers to follow the five steps (outlined in the System Architecture Diagram in the Supplementary Documents) to conduct and report on their research. In Steps 1 and 2, researchers design minimally invasive RCEs within the ASSISTments Skill Builders. Then their study is run within the population of students using ASSISTments. In Step 4, they receive organized reports detailing student performance that streamline the analysis of learning interventions. The experiments that researchers conduct using the TestBed are low cost, high power, noninvasive to learning, and feature student-level randomization through access to ASSISTments' user population. Part of the **broader impact** of the SI2 grant was that researchers from Southern Methodist University, Northwestern, Indiana University, Stanford, Columbia Teacher's College, Harvard, University of Wisconsin, Boston College, Carnegie Mellon University, and University of Illinois Chicago have all participated in TestBed research.

A major goal of the SI2 grant was to build an infrastructure to leverage ASSISTments as a shared scientific instrument. One measure of the success of this goal was the degree to which researchers were publishing work not co-authored by PI Heffernan or another member of the WPI Team. Two such publications resulted from RCEs conducted in the TestBed (Koedinger & McLaughlin, 2016; Fyfe, 2016). These two, as well as the dozen other papers that use publicly released data sets are listed in the reference section and constitute the **intellectual merit** of that prior NSF funding. Finally, the complete system was described in Ostrow, Selent, Wang, Van Inwegen, Heffernan, & Williams (2016).

4.2.5 DIBBs Funding Impacting Multiple Disciplines

DIBBs funding will help streamline the process of conducting educational experiments, which will advance the science of learning through short-cycle evaluations. Simultaneously, the impact on the science of learning will be broad, covering multiple disciplines. Researchers recruited will hail from disciplines including educational psychology, cognitive psychology, mathematics education, and behavioral economics (see letter in the Supplementary Documents from interested researchers). Research topics of interest spanning these disciplines are broad, including **Types of Feedback** (Fyfe, Rittle-Johnson & DeCaro, 2012; Stamper, et al., 2013; Sweller, Kirschner & Clark, 2007; Keehner, et al., 2008), **Comparison** (Jee, et al., 2013; Stamper, et al., 2013; Sweller, Kirschner & Clark, 2007; Keehner, et al., 2008), **Sequencing and Spacing** (Roediger & Karpicke, 2006; Richland, Kornell & Kao, 2009; Pashler, Roher, et al., 2007; Butler & Roediger, 2007), **Self-Regulated Learning & Metacognition** (Ehrlinger & Shain, 2014; Belenky & Nokes-Malach, 2013; Roll, et al., 2012), **Social Context & Interaction** (Walkington, 2013; Walker, Rummel & Koedinger, 2011; Mazzocco, et al., 2013), **Assessment** (Rafferty & Griffiths, 2014; Ocumpaugh, et al., 2014; Storm, et al., 2006), **Motivation** (Kelly, Heffernan, D'Mello, et al., 2013; Williams, et al., 2013; Bernacki, Byrnes & Cromley, 2012; Chernyak & Kushnir, 2013), and **Mathematics Education** (Rau, et al., 2012). This list provides an example of the breadth of topics that can be investigated using the ASSISTments TestBed but is by no means complete. Many authors of the work listed above have already started working with the TestBed through the NSF SI2 grant, and DIBBs funding would be critical in enhancing the efficiency and validity of their work.

4.3 Current Workflow and Flaws of the ASSISTments TestBed

The work done with outside researchers for the SI2 grant has shown that there is now a TestBed within ASSISTments where researchers can use ASSISTments and the existing subject pool to create and run studies. While the current workflow underlying researchers' use of the ASSISTments TestBed has grown from its pilot form under the SI2 grant, it remains bulky, with bottlenecks requiring the WPI Team when the researcher has to build their study and when the researcher has to handle the large amount of resulting data. Due to the need for such support, the team at WPI has to intervene often. The following sections outline the current workflow in two phases (Building Studies and Data Reporting/Analytics) and pinpoints potential areas for improvement that could be targeted through DIBBs funding. To embody the process, the true stories of two researchers using the TestBed are highlighted.

4.3.1 Building Studies

4.3.1.1 Current Process

The process of conducting an RCE within ASSISTmentsTestBed.org typically involves researchers using the preexisting Skill Builder content as a control and a modified version as the treatment. All studies are done with student-level random assignment, which means that students in the same class will be randomly assigned to different conditions.

Researchers first sign up for an ASSISTments account from which they can access our Skill Builder library. In order to access the ASSISTments subject pool, the current workflow requires researchers to submit a structured abstract of their work to the WPI Team by emailing assistments-research@wpi.edu. At www.assistmentstestbed.org we provide the researchers with forms to help us gather information about their intent for the experiment. The researchers are also expected to sign the form approved by the WPI-IRB regarding terms of use (see www.assistmentstestbed.org/home/terms-of-use) where they pledge not to attempt to identify students. Researchers have to propose experiments that qualify under WPI's IRB-approved exemption for studies "comparing normal instructional strategies." In addition, researchers are required to attain approval for their work from their own IRB (and send that approval to WPI), which is typically a smooth process through the exemption of working with de-identified data.

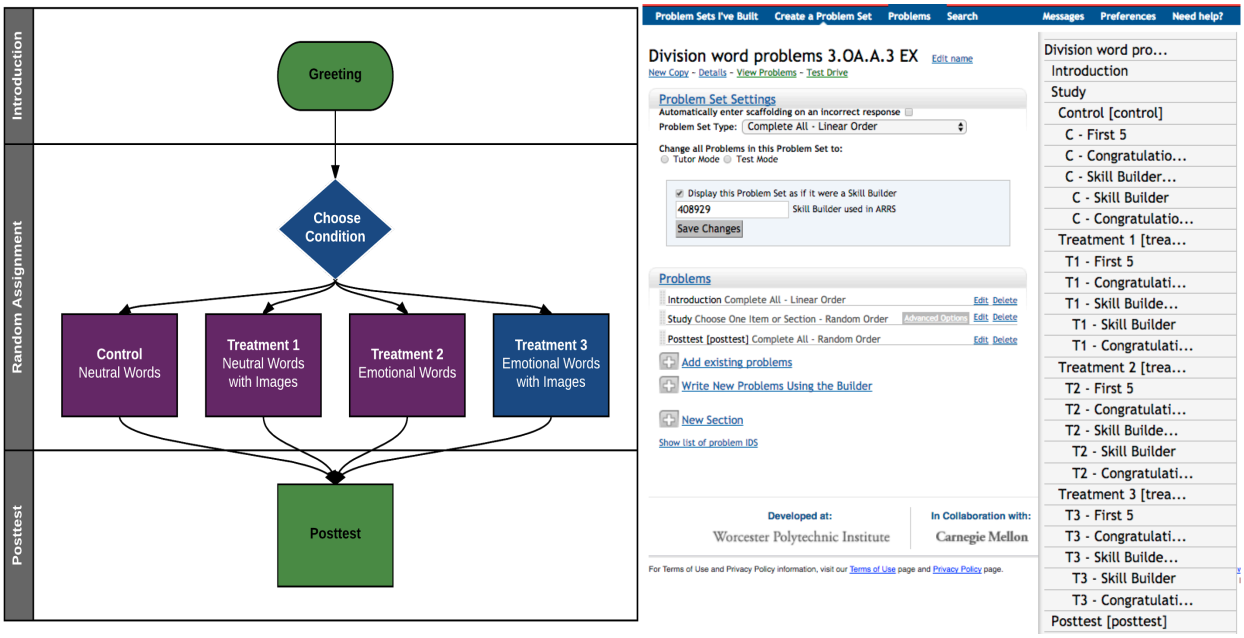
The research idea is reviewed by the WPI Team to ensure the work can be considered a 'normal instructional strategy,' and that the design can be implemented using the TestBed (some studies may be too complex for the capabilities allowed by our infrastructure). A conference call is then scheduled with the researcher to begin the process of building the experiment. Researchers are asked to pre-register their research through avenues such as the Open Science Foundation (https://osf.io/). In order for a manipulated Skill Builder to reach the ASSISTments subject pool, the experiment must be reviewed a final time by the WPI Team for functionality and content accuracy before it can be released to the teachers and students through our "ASSISTments Certified" folder.

4.3.1.2 Areas for Improvement Highlighted in a Case Study

The current workflow for researchers to build RCEs within the TestBed is heavily dependent on contact with the WPI Team. From the initial vetting of the proposed design, through to final reviews of content and functionality, the WPI Team oversees study creation and integration with ASSISTments content, serving as a distinct bottleneck for the process when considering scaling this design. To highlight specific weaknesses in our current protocol, we present the recent work of Karina Hamamouche, a graduate student working in the Infant and Child Cognition Lab with Sara Cordes at Boston College. Cordes conducted a study using the TestBed as part of the NSF SI2 cohort of researchers and introduced her students to the tool. Karina's story is a perfect example of how the TestBed workflow can be improved with regard to helping researchers build studies.

Karina initially created a researcher account and presented a study proposal on October 17, 2016. Her proposed work sought to examine how altering problems to include emotional content might impact students' learning. The work was built around previous findings that mixing emotion into counting content could enhance student performance (Young & Cordes, 2013). She proposed taking a Skill Builder focused on third-grade mathematics word problems and altering it to include emotionally valenced statements (e.g., "Carla *was so excited when her* mom gave her $41 to go to the pet store. Then, Carla's grandparents gave her $67. How many dollars does Carla have to buy a puppy?") in comparison to the original problems (e.g., "Carla's mom gave her $41 to go to the pet store. Then, Carla's grandparents gave her $67. How many dollars does Carla have to buy a puppy?").

Karina worked with Cristina Heffernan, co-founder of ASSISTments, via numerous emails and two one-hour conference calls to select appropriate Skill Builders in which to embed her experimental content. She was pointed to www.assistmentstestbed.org for information on how to build her experiment within ASSISTments. However, the materials did not suffice and communications between Karina and the WPI Team proceeded through a chain of approximately 30 emails spanning the course of two months. The building process within ASSISTments was particularly complex because Karina hoped to front-load her experimental content into the first five problems that a student would see. This is a common approach used in the TestBed to keep researchers from having to design content for 100+ skill problems, but the implementation of the design within ASSISTments can be tricky. Back and forth communication helped to answer questions, spurred a graduate student on the WPI Team to build the experimental structure for Karina to abate confusion, and helped to settle final review requirements before the Skill Builder could be released to teachers. The process also saw revisions to the original experimental design following Karina's meeting with her advisor, and a new set of reviews was required of the content and new experimental conditions. Ultimately, her work resulted in the comparison of four conditions that students would be randomly assigned to through a "Choose Condition" within the tutor. This was a design that had been used before, it could have been a template for her to fill out but instead the team had to scramble to support her. Her design, shown in Figure 4.1 (left), grew much more complicated in the build within ASSISTments, shown in Figure 4.1 (right).



Division word problems 3.OA.A.3 EX Edit name

New Copy - Details - View Problems - Test Drive

Problem Set Settings

Automatically enter scaffolding on an incorrect response □

Problem Set Type: Complete All - Linear Order

Change all Problems in this Problem Set to:
○ Tutor Mode ○ Test Mode

☑ Display this Problem Set as if it were a Skill Builder

408929 Skill Builder used in ARRS

Save Changes

Problems

Introduction Complete All - Linear Order Edit Delete

Study Choose One Item or Section - Random Order Advanced Options Edit Delete

Posttest [posttest] Complete All - Random Order Edit Delete

Add existing problems

Write New Problems Using the Builder

New Section

Show list of problem IDS

Developed at: Worcester Polytechnic Institute

In Collaboration with: Carnegie Mellon

For Terms of Use and Privacy Policy information, visit our Terms of Use page and Privacy Policy page.

- Division word pro...
  - Introduction
  - Study
    - Control [control]
      - C - First 5
      - C - Congratulatio...
      - C - Skill Builder...
        - C - Skill Builder
        - C - Congratulatio...
    - Treatment 1 [trea...
      - T1 - First 5
      - T1 - Congratulati...
      - T1 - Skill Builde...
        - T1 - Skill Builder
        - T1 - Congratulati...
    - Treatment 2 [trea...
      - T2 - First 5
      - T2 - Congratulati...
      - T2 - Skill Builde...
        - T2 - Skill Builder
        - T2 - Congratulati...
    - Treatment 3 [trea...
      - T3 - First 5
      - T3 - Congratulati...
      - T3 - Skill Builde...
        - T3 - Skill Builder
        - T3 - Congratulati...
  - Posttest [posttest]

Figure 4.1. Karina's design, stated simply through a flow chart (left) and in a more complex manner in the current format of the TestBed builder (right).

4.3.2 Data Reporting and Analytics

4.3.2.1 Current Process

The current process that assists researchers in analyzing their data grew out of the SI2 grant and employs the Assessment of Learning Infrastructure (ALI). ALI is meant to lighten the load of working with large data-files output from RCEs conducted within the ASSISTments TestBed.

For each student participating in a study, ASSISTments logs substantial data detailing performance as the student progresses through the Skill Builder. This data includes measures of problem accuracy (e.g., a correct or incorrect first response, or a partial credit score), the student's first action (e.g., an attempt vs. requesting tutoring), the number of attempts per problem, the number of feedback interactions per problem (e.g., hints requested or scaffolds seen), whether or not the student saw the bottom-out hint (e.g., the correct answer, provided to keep the student from getting stuck within the assignment), and start and end times for each problem. Researchers can also access clickstream data at the action level, detailing each step taken by a student within problems. ASSISTments is also able to provide data on the student's performance in the system prior to their inclusion in a study, student characteristics (e.g., gender, age), and other helpful variables at the class and school levels (e.g., accuracy on problems completed in the past.) Through use of the TestBed, this information is consolidated, anonymized, and provided to researchers through a series of reports.

ALI has grown considerably over the past year. It began as a robust SQL query to the ASSISTments database to retrieve unified information across any study and to present it to researchers in a single format. When a study is activated within the TestBed, details about the study and the researcher's contact information are entered into ALI's study repository. Once ALI has determined that new data is available, several robust SQL queries are run on the ASSISTments database. By working closely with researchers throughout the development of ALI, four different universal data representations have been designed and are provided to all researchers in an attempt to meet dynamic research needs. Examples of each type of data report are available at (Ostrow & Heffernan, 2014) for further consideration. Each data report is presented to researchers in a single, organized communication, stored and delivered using Google Docs, as discussed at (Ostrow & Heffernan, 2014).

When researchers publish their findings, a condition of working within ASSISTments requires that they include a reference in their work to the stable ALI record from which they sourced their data files for final analysis. This approach allows reviewers and secondary researchers to gain access to their raw data, deterring researchers from cherry-picking significant data reports and encouraging replication (Open Science Collaboration, 2015; Ioannidis, 2005).

4.3.2.2 Areas for Improvement Highlighted in a Case Study

The current workflow for researchers to access data from their RCE is already largely automated thanks to the NSF SI2 grant. However, there are still areas for improvement including a continued focus on meeting researchers' needs through pre-processing, feature generation, and data delivery. To highlight specific weaknesses in our current protocol, we discuss the recent work of David Uttal, a Professor of Psychology and Education at Northwestern University. David's story is unique in showing how researchers need more help in examining the data from their RCEs.

As a professor with a full research laboratory of graduate and postdoctoral students, Uttal came to the TestBed with a set of easily testable hypotheses and a strong application. His goal was to examine the relation between spatial ability and math learning at scale, in order to better design problem and hint content to improve students' learning. The expectation was that the manner in which a Geometry problem was presented (visuospatial vs. analytic) could interact with the student's spatial ability and problem-solving strategies. Uttal's graduate student built three versions (for three Geometry topics: rotations, reflections and transformations) of his study within Skill Builders, following an organizational conference call with Cristina Heffernan and limited additional support.

David recently reached out to us after receiving his data files from ALI and finding them overwhelming. There were 2,213 students who had started at least one of his three experiments. The granularity of data varies within each report. One spreadsheet offers each row in the dataset as representing a student's progression through the experiment, with regard to all the features logged by the tutor. Another offers finer-grained analysis, with each row in the dataset representing a problem and all of its features as experienced by a student progressing through the experiment. The reports can be even more fine-grained by considering students' clickstream data (or "action-level" file where each row represents a keystroke or click made by a student during problem solving). It is easy to understand how educational psychologists experienced with smaller datasets may find ALI's data reports overwhelming. It can be difficult to determine which granularity of the data would be most helpful in answering particular research questions, and the scope of the data can be overwhelming.

David's work is a prime example of how researchers need assistance to examine their data. This month, PI Heffernan is scheduled to meet with David to help interpret his data. Although Heffernan knows well how to analyze this raw data, we underestimated the handholding that would be required to help researchers learn these methods. At the moment, we have to work with each researcher. However, as researchers are using only a small number of research designs (in the next section we will call these "templates") we think we can provide assistance by organizing the data for the researchers, and thus make the experience less overwhelming. If this grant is funded, we will create *templates* for the standard experimental designs we are able to support that will also provide standard analytic reports to make the researcher's job of analyzing their data much simpler.

Furthermore, while researchers like Uttal are great experimental psychologists, they do not have the skills to leverage advances in educational data mining (EDM) like "deep knowledge tracing" (Piech, et al. 2015). Deep Learning is a type of machine learning resulting in, for example, Google researchers' software beating the human world champion in the game of Go (Gibney, 2016). Researchers know that the more they account for additional sources of variance (e.g., students who have performed well in the past tend to score higher on Post Tests; researchers measure the effect of experimental condition *after* factoring out as many other sources of variance as possible), they will be better able to measure the effect of their interventions. An exciting development in EDM has been applying concepts such as deep knowledge tracing to try to better model student behavior. This grant will attempt to help education researchers by giving them access to the running of these models to determine if they can capture other sources of variance, and better analyze their own experiments.

4.4 The Solution

The momentum derived and the lessons learned from the NSF SI2 grant have the potential to propel great success from DIBBs funding. The SI2 was successful, with more than a dozen studies submitted, and many still running. Two papers fully independent of the WPI Team were published within two years of pilot work (Fyfe, 2016; Koedinger & McLaughlin, 2016). The NSF SI2 grant helped to show that this work was possible and that researchers were interested in the capacities of a shared scientific tool for educational research. However, the bottleneck established by the lengthy time investment required of the WPI Team motivates the creation of user-friendly tools for both experiment creation and data analysis proposed in this grant.

Given that the TestBed is “open for business,” with researchers still actively submitting study proposals, taking the necessary steps to automate the TestBed’s workflow using a series of experimental design templates will greatly influence the success of the TestBed as a shared scientific tool. Our proposed plan is to use DIBBs funding to restrict the breadth of experimental designs that can be implemented within the TestBed to a number of preconceived templates for Skill Builder research (e.g., 25 experimental designs that will cover the needs of most researchers proposing study ideas in the TestBed). This will allow for the design of an intelligent user interface, which we will call the ASSISTments RCE Template Tool, to help researchers as they build their experiments.

The Template Tool will walk researchers through the process of selecting the proper template to match their experimental design, and then assist researchers in completing the template with their experimental content and compare it to the existing content in the ASSISTments Skill Builder. While the Template will limit the freedom that researchers have with their designs, it will drastically increase ease of use and researchers’ independence from the WPI Team while using the TestBed, thereby boosting the efficiency and further reducing the cost of “at-scale” research. As such, we request DIBBs funding to build upon the already-existing ASSISTments TestBed infrastructure, enhancing the accessibility of the tool and promoting researchers’ success.

4.4.1 How DIBBs Funding Will Improve The TestBed Workflow

The goal of our proposed project is to build upon the initial pilot set forth by the NSF SI2 grant in order to achieve a product that is more openly and readily available to researchers across disciplines. A modest investment from DIBBs will serve to extend the cyberinfrastructure underlying the TestBed, build upon that infrastructure to redesign a workflow promoting research independent of the WPI Team, expand the knowledge base detailing how to conduct RCEs in the TestBed through an intelligent user interface, and continue to disseminate literature and promote the TestBed as a shared scientific tool that can benefit multiple disciplines.

Across the three-year project, 21 researchers in total, seven each year, will be invited to attend a one-day meeting at WPI meant to initiate the relationship with the researchers and introduce them to ASSISTments. This will prepare them to participate in think-aloud protocols throughout the year. A portion of funding earmarked for activities within participant support costs will fund these meetings. These meetings will allow researchers flexibility in determining and implementing their research questions through formal introductions to Skill Builder problem sets in ASSISTments. The Template Tool is intended to be self-explanatory, so it will only be introduced during the online think-aloud protocols. The venue will also allow researchers to meet local ASSISTments teachers through short school visits. Portions of the meeting will be recorded and may later serve as instructional material for other researchers building RCEs within the TestBed.

Each year following the meeting, the seven researchers will be expected to maintain contact with the WPI Team, participating in think-aloud protocols to iteratively improve the Template Tool and ALI's data-reporting strategies. To do this, the participating researchers will use the Template Tool and see their ALI data for the first time during recorded online meeting sessions (we will use Google's Hangout on Air). These think-aloud sessions will support iterative improvements of the Template Tool and the ALI connection. Researchers will participate in a think aloud when they initially build their RCE problem set. Researchers will also be asked to participate in a think aloud upon receiving their data from ALI, in order to help the WPI Team better facilitate researchers' needs with regard to the organization, presentation, and functionality of data reports.

The following sections explain more specifically how funding will be used to impact the two primary facets of the TestBed workflow: Building Studies and Data Reporting/Analytics.

4.4.2 Building Studies

4.4.2.1 Intended Impacts from DIBBs Funding

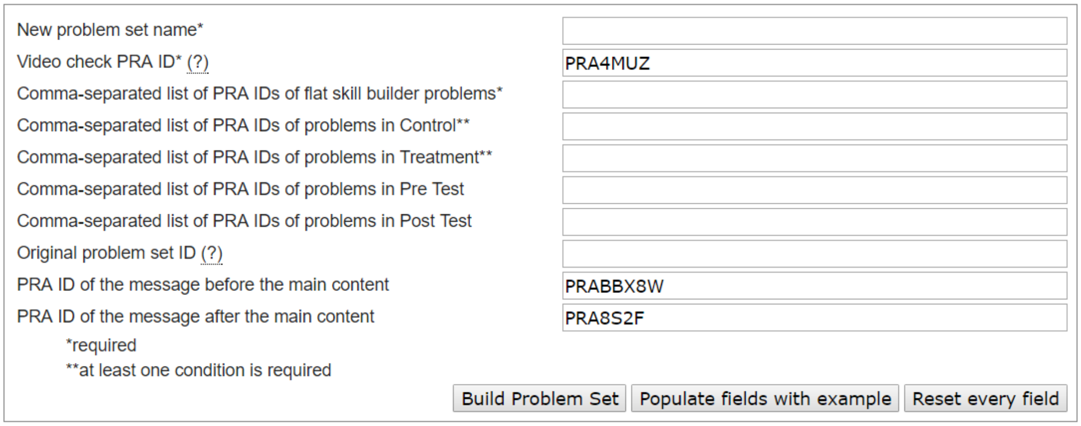
ASSISTments provides a problem set builder that is feature-rich, allowing researchers to build a wide range of experimental designs to answer a variety of research questions. As such, as explained by Karina's story, a simple experimental design can grow drastically complex when implemented using the problem set builder. It grows very difficult, especially for researchers less familiar with ASSISTments, to build a structure that adequately fulfills their intended experimental design. It is also extremely time consuming for the WPI personnel to debug problem sets created by novice builders and make sure it is ready for use by real students.

From our experience running RCEs within the TestBed, and assisting researchers in implementing their experimental designs, we discovered that most experimental ideas fall into a set of design patterns that are complicated to build. For instance, if the experiment contains videos, the researcher will need to add an If-Then-Else section at the beginning to act as a "Video Check" question; this would play a YouTube video and ask the students to type a word given in the video or to announce they could not see the video. The System Architecture Diagram found in the supplementary documents of this grant shows an example template depicted as a flowchart. The fact that it is a flowchart reminds researchers that a problem set is in fact a domain-specific language (with Conditionals like "If-Then-Else") that is compiled at run time and shipped to the client to be executed.

In addition to the video-check section, there are other common problems that have emerged as researchers bring their designs to us. For instance, a skill builder typically contains 100 items, and the system chooses items randomly from within a list. However, it becomes impractical for researchers to create 100 different responses for all problems in the Skill Builder. The ability to randomly give items to students, however, reduces instances of cheating off their friends. As a solution, we have devised compromises that require the researchers to create content for a small number of problems, perhaps only five for example, and we ensure that those five items will be randomly assigned early in the problem set (still reducing the likelihood of cheating).

The true issue emerges as these common problems, along with others not described here, are often bound together when building an experiment. That makes the resulting problem set structure feel more like a computer program (as it technically is), adding complexity and requiring building tasks beyond the skill set of most of the researchers wanting to create an RCE. What is needed to solve this problem is a simple tool to help researchers construct these experiments, similar to programs like TurboTax that guide individuals through complex tax forms by asking a series of questions and programmatically determining the least expensive way to fill out their tax forms.

By creating this tool, referred to as the *"Template Tool,"* we will 1) help researchers create their experiments with ease, and 2) reduce the work WPI has to do in checking and revising experiments. Finally, we will also be able to write analytic code for these common template designs, providing the researchers with quick, focused answers to the questions they have specified rather than providing large data files without any inferences.



New problem set name*	
Video check PRA ID* (?)	PRA4MUZ
Comma-separated list of PRA IDs of flat skill builder problems*	
Comma-separated list of PRA IDs of problems in Control**	
Comma-separated list of PRA IDs of problems in Treatment**	
Comma-separated list of PRA IDs of problems in Pre Test	
Comma-separated list of PRA IDs of problems in Post Test	
Original problem set ID (?)	
PRA ID of the message before the main content	PRABBX8W
PRA ID of the message after the main content	PRA8S2F

*required
**at least one condition is required

Figure 4.2. The form used by researchers in the pilot Template Tool to populate an experimental design Template with their content.

We have begun some pilot work toward this tool, where Heffernan's graduate student, Thanaporn Patikorn, created a simple application to help five users create their own experiments. The pilot Template Tool, shown in Figure 4.2**,** allows users to choose from five templates, asking them to simply fill in the chosen template with the problems they wish to use in each condition. As we talked with users who used the tool, they indicated that though it was useful, and though it saved them time by building a lot of the structure they would have had to manually create, it was too simple and did not allow for the nesting of templates. Thus, we believe that this tool could be useful if it were expanded to deal with a broader range of concerns. Additionally, it will link directly to a visual flowchart-like interface that will allow researchers to 'step through' their experiment naturally. The formal Template Tool, funded by a DIBBs grant, will allow researchers to better "program" their experiments though a dialog-based interview that speaks the language of the researcher. For example, it would ask, "Does your experiment use video?" or "How many conditions are you implementing?" DIBBs funding will be crucial in establishing usability of the Template Tool as an infrastructure for researchers to efficiently build RCEs.

4.4.2.2 Planned Yearly Progress

In Year 1, DIBBs funding will support one graduate research student in coding the formal Template Tool. Additionally, Cohort A of seven researchers will become part of the testing team. As researchers build their RCEs working with the pilot Template Tool, they will participate in online think alouds to help the WPI Team isolate areas in the cyberinfrastructure for improvement. Cohort A will also be asked to examine mockups of the visual interface that will become the formal Template Tool, in order to gauge user feedback and guide progression of the prototype. The amount of time the WPI Team spends assisting researchers in building RCEs will be logged for later quantitative analysis.

In Year 2, we will formally release the Template Tool with its full graphical user interface. An additional seven researchers will form Cohort B. Think alouds will be conducted with researchers in Cohorts A and B to assess the Template Tool's ease of use and to isolate areas for improvement. By Year 2, the Template Tool will be fully functional, and feedback from researchers will be crucial in helping to make final revisions to the Tool. The amount of time required of the WPI Team is expected to fall drastically in Year 2, but we anticipate continued communication with our research cohorts.

In Year 3, Cohort C, the final cohort of seven researchers will be established, and a third meeting will be held. We anticipate that researchers in Cohort C will be able to work in the TestBed using the refined Template Tool with little to no support from WPI and we will confirm this by once again conducting think-aloud sessions with them online.

The tangible outcome from DIBBs funding will be the Template Tool itself, as a piece of cyberinfrastructure meant to help researchers. The success of the Template Tool will be measured by how it alters the time investment (and thereby cost) of the WPI Team as we work with researchers using the TestBed. Success will also be measured qualitatively through the think alouds held with researchers from each cohort and survey measures collected in the final year to produce a summative assessment of the progress made via DIBBs funding. The WPI Team will log the amount of time spent each year working with the Cohorts, in order to show quantitatively that time and cost investments are significantly lowered by Year 3 as the number of publications increases.

4.4.3 Data Reporting and Analytics

4.4.3.1 Intended Impact from DIBBs Funding

When conducting an RCE within the TestBed, researchers access their data through the existing ALI cyberinfrastructure. ALI's reports are currently run using Google Scripts that communicate with a Java application implementing a Google API that programmatically creates Google Docs and Google Sheets; these Google Sheets include multiple, preprocessed forms of the same data (action-level, problem-level, and student-level) for running different types of analyses. The reports also provide a few basic analyses of the distribution of the dataset to help researchers begin their data exploration.

By implementing the Template Tool and the use of a restricted number of experimental designs through Templates, it will be possible to refine ALI to include powerful analytics in each data report, unique to the experiment at hand. ALI already contains the programming required to conduct an assessment of selection bias across conditions (e.g., a Chi-square analysis) and to conduct a series of simple hypothesis tests isolating Post Test performance as the dependent variable. Currently, ALI approaches Post Test analyses much like a researcher would: if only two conditions exist within a study, ALI conducts a t-test, while if more than two conditions exist, ALI conducts an ANOVA. ALI currently has the Python-based API to support simple univariate analyses including ANOVA and ANCOVA, as well as Deep Learning EDM methods. Results are accumulated and presented to researchers in their ALI Reports.

Through DIBBs funding, one graduate student will work to link these analyses to experimental design Templates and ensure proper analytic approaches are automated. The graduate student will process researcher's think alouds pertaining to ALI reports and use their feedback to automate analysis and improve the reporting infrastructure.

Researchers are always looking for features that help capture variance in their dependent variable (e.g., in an ANCOVA) and, as such, we will work with researchers to help them find features in a student's prior work that might be useful. We already provide features such as a student's accuracy level on problems solved before the start of an experiment. Additionally, feature generation will be inspired by feedback from researchers through data analysis think alouds. These additional features will be added to template-specific analyses in order to better control for student, class, and school-level characteristics that researchers care about by using covariate approaches to analysis.

Automating analysis might raise questions around the WPI Team's data preprocessing and cleaning practices and how our choices as a team impact automated analyses so these points will be discussed with researchers, after each think-aloud protocol, in detail to get feedback on our techniques. Additionally, researchers will always have access to the raw data from their RCE in order to conduct traditional analyses, and they will be encouraged to do so. Automatic reporting and analytics provided through ALI will always be presented as a stepping-stone to pique researchers' interest rather than results that should be published without further investigation.

4.4.3.2 Planned Yearly Progress

Near the end of Year 1, the WPI Team will conduct think alouds with the Cohort A researchers as their studies begin to have enough data to be analyzed. This process will provide feedback on the report, the data files provided, and the types of analyses that researchers are hoping to find given the Template they have employed using the Template Tool. Qualitative data gathered during Year 1 will, along with understanding of statistical methods, guide linking experimental design Templates with proper analyses for later automation.

In Year 2, automatic analyses will be embedded into ALI reports and driven by the Template selected by the researcher. Think-aloud protocols will be conducted with researchers from Cohorts A and B gathering two types of qualitative data: feedback on analytic implementation from Cohort A (i.e., those who ran their first studies without automatic analysis), and insight with regard to our functionality or usability of ALI's automated reports that could be further improved from Cohort B (i.e., those new to the process). This information will be used to make adjustments to the infrastructure's capacities.

In Year 3, researchers will be expected to use the Template Tool and ALI reports with little assistance from the WPI Team but think alouds will again be collected in order to conduct additional refinement of these tools.

The tangible outcome for the ALI report from DIBBs funding will be the addition of automated analyses linked to the Template Tool. The impact of the addition of these analyses will be measured by feedback from researchers in Cohorts A, B and C, and by the decrease in time required of WPI Team members working with researchers to understand and analyze their reports. Additionally, survey measures taken at the end of Year 3 will help to form a summative assessment of the growth of ALI's reporting through the life of DIBBs grant funding.

4.4.4 RCE Pre-Registration through the Open Science Framework

The WPI Team has always valued transparency in research and the quest for open, accessible, shared data. Since the beginning of the NSF SI2 grant, and with the creation of the ALI infrastructure, the WPI Team requested that researchers cite the data report used for analysis in their published work, allowing for reader access to the data and thereby the potential for replication.

During the last year we have required researchers preregister their study at a place like Open Science Framework (OSF; https://osf.io/). This will allow researchers to pre-register their RCEs, logging hypotheses and intended analyses prior to accessing their data in order to strengthen the validity of results observed using the TestBed. The OSF will also serve as an outlet for data preservation and transparency, logging any changes that researchers make to their dataset or their analytic efforts, and helping to restrict the cherry picking of results and to reduce the file drawer problem.

4.4.5 Potential for Interdisciplinary Work & Extending Capabilities to Other Research Communities

As part of the DIBBs grant, the WPI Team will focus on integrating researchers recruited from fields such as educational psychology with experts in the educational data mining community. These pairings will help to expand the perspective on big data and the value of short-cycle RCEs conducted at scale, serving to benefit both communities. As displayed by David Uttal's case study, researchers from Ed Psych backgrounds bring strong theory-based hypotheses to the table, while those in the EDM community bring carefully crafted big-data analytics. The pairing of these disciplines will also generate stronger findings in the science of learning.

4.4.6 Potential to Look Across Studies

Considering the wealth of studies that will be made possible by the addition of the Template Tool to the TestBed, it will be possible to identify high-impact learning interventions by looking across studies. Throughout the project, researchers' studies will be logged in detail and categorized based on various characteristics such as targeted skill, content area, format of content presentation (e.g., image, text, interactive object), type of feedback/assistance/scaffolding (e.g., text, video), etc. An initial review of these findings will identify enhancements for which evidence of effectiveness was found, helping to inform future research. Analyzing numerous studies simultaneously has the potential to open new doors for the data mining community by pushing researchers to build models that better predict experimental effect sizes with the ultimate goal of personalizing learning. Personalized learning is currently a hot topic in education, and broad datasets spanning versatile experimentation may help to reveal details about subgroups of student populations that would otherwise be overlooked. PI Heffernan's graduate students have examined the benefits of simultaneously analyzing 20 of the experiments running in the ASSISTments Testbed (Selent, Patikorn, & Heffernan, 2016). We believe that following similar practices in the future, we will better learn to estimate treatment effects by using data from across multiple experiments. We also believe that this process will lead to better fitting covariates to account for extraneous sources of variance in experimental models.

4.5 Outcomes and Metrics

To quantify the impact on the science of learning, we plan to recruit and support seven researchers each year. This will result in three cohorts of researchers, totaling 21 working in the system by Year 3. We expect that, at a minimum, each researcher will conduct one study per year using the ASSISTments Open TestBed. With this plan, there will be at least 7 + 14 + 21 = 42 studies conducted. We expect 50% will be written up and accepted for publication for a total of 21 publications. These publications will not feature PI Heffernan or other members of the WPI Team as co-authors, but will require a notice of acknowledgement of the researcher's use of the ASSISTments TestBed. All qualitative metrics, in the form of think aloud interviews and surveys, will help to guide the resulting Template Tool and ALI reporting analytics. In Year 3, we will focus on sharing what we have learned and focus on moving to a sustainable model for ongoing use of the TestBed by the educational research community, as well as to engage in a summative evaluation of the effectiveness of the project relative to original project goals.

4.6 Management Plan and Sustainability Plan

Our three-year project plan is focused around the creation, support, and ongoing expansion of a small network of researchers capable of conducting rigorous studies on student learning using the Template Tool built upon the already existing capacities of the ASSISTments TestBed. DIBBs funding will allow us to expand the infrastructure to establish a sustainable process for researchers to conduct RCEs and publish their results. Having managed over 24 grants worth over $15 million, Dr. Heffernan has the experience to manage this project that combines his experience in 1) creating randomized controlled trials, and 2) using educational data mining techniques to better detect heterogeneous treatment effects, finding out which students benefits from which interventions.

Our long-term sustainability plan follows that of other shared scientific tools, such as the Green Bank Observatory, with initial instrument construction and small maintenance funded by the NSF. In the case of the Observatory, the telescope is free to use (dubbed the "Open Skies Policy"). Our governance structure for this infrastructure seeks to be like that of the Green Bank Observatory, that is, having a scientific review panel that selects the projects that are deemed most important. The WPI Team seeks to run the TestBed in a similarly open and free manner, with the strongest idea proposals receiving priority access to Skill Builder content for experimentation. Also, as with the Observatory, all resulting data will be open and accessible to any researcher.

WPI is committed to hosting ASSISTments as a free public service, and as such, the TestBed service to support non-profit university-based scientific research. PI Heffernan envisions his role as similar to that of Jimmy Wales, creator of Wikipedia, but instead of curating encyclopedia entries, ASSISTments curates optimized educational practices. As such, we are open to whatever good ideas work. If this grant succeeds we will be able to show this is a worthy investment to help researchers do their work.

Systems like ASSISTments can provide researchers with access to an extensive and diverse subject pool, an automated fine-grained logging of educational data, validated measures of student learning and affect, and automated data reporting and analysis to tackle the problems or conducting RCE in authentic learning environments (e.g., 50,000 students' nightly homework). The **broader impact** of this work will be to enhance the learning of 50,000 students using ASSISTments, by enhancing the content those students receive.

References

Achenbach, J. (2015). Many scientific studies can't be replicated: that's a problem. The Washington Post, Speaking of Science. August 27, 2015. Retrieved 2/19/16 from https://www.washingtonpost.com/news/speaking-of-science/wp/2015/08/27/trouble-in-science-massive-effort-to-reproduce-100-experimental-results-succeeds-only-36-times/

Alsheikh-Ali, A., Qureshi, W., l-Mallah, M., & Ioannidis, J. (2011). Public Availability of Published Research Data in High-Impact Journals. Plus-One, September 2011, 6(9), e24357. Retrieved March 10, 2014, from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0024357.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. Nature, 483, 531-533.

Belenky, D.M. & Nokes-Malach, T.J. (2013). Mastery-approach goals and knowledge transfer: An investigation into the effects of task structure and framing instructions. Learning and Individual Differences. 25, 21-34.

Bernacki, M.L., Byrnes, J.P. & Cromley, J.G. (2012). The effects of achievement goals and self-regulated learning behaviors on reading comprehension in technology-enhanced learning environments. Contemporary Educational Psychology. 37(2), 148-161.

Broad, W.J. (2014, Mar 15). Billionaires With Big Ideas Are Privatizing American Science, Retrieved on June 3, 2014, from New York Times Science http//www.nytimes.com/2014/03/16/science/billionaires-with-big-ideas-are-privatizing-american-science.html.

Butler, A.C., & Roediger, H.L., III. (2007). Testing improves long-term retention in a simulated classroom setting. European Journal of Cognitive Psychology. 19, 514-527.

Chernyak, N., & Kushnir, T. (2013). Giving Preschoolers Choice Increases Sharing Behavior. Psychological Science. 24(10), 1971-1979.

Ehrlinger, J. & Shain, E.A. (2014). How accuracy in students' self perceptions relates to success in learning. In V.A. Benassi, C.E. Overson, & C.M. Hakala (Eds.). Applying science of learning in education: Infusing psychological science into the curriculum. Retrieved January, 14, 2015, from http://teachpsych.org/ebooks/asle2014/index.php.

Fyfe, E. (2016) Providing feedback on computer-based algebra homework in middle-school classrooms. Computers in Human Behavior, 63. 568-574.

Fyfe, E.R., Rittle-Johnson, B. & DeCaro, M.S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. Journal of Educational Psychology. 104,1094-1108.

Gibney, E. (2016). Google AI algorithm masters ancient game of Go. Nature. 529 (7587): January 2016.

Heffernan, N. T. (2014). Press. Retrieved on March 10, 2014, from http://www.aboutus.assistments.org/press.php.

Institute of Education Sciences. (2013). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. U.S. Department of Education. Retrieved 2/18/16 from http://ies.ed.gov/ncee/pubs/evidence_based/randomized.asp.

Ioannidis J. P.A. (2005). Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124. Retrieved on March 17, 2014, from http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0020124

Jee, B., Uttal, D., Gentner, D., Manduca, C., Shipley, T. & Sageman, B. (2013). Finding faults: analogical comparison supports spatial concept learning in geoscience. Cognitive Processing, 14(2), 175-187.

Johnson, G. (2014, Jan 20). New Truths That Only One Can See. New York Times Science, Retrieved on June 3, 2014, from http://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html.

Johnson, G. (2014, Mar 7). When Studies Go Wrong: A Coda. New York Times Science, Retrieved on June 3, 2014, from http://www.nytimes.com/2014/03/07/science/when-studies-are-wrong-a-coda.html.

Keehner, M., Hegarty, M., Cohen, C., Khooshabeh, P. & Montello, D.R. (2008). Spatial reasoning with external visualizations: what matters interact. Cognitive Science. 32, 1099-1132.

Kelly, K., Heffernan, N., D'Mello, S., Namias, J., & Strain, A. (2013). Adding Teacher-Created Motivational Video to an ITS. In Boonthum-Denecke, Youngblood (Eds) Proceedings of the Twenty- Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2013, St. Pete Beach, Florida. May 22-24, 2013. AAAI Press 2013. pp. 503-508. Retrieved on June 2, 2014, from http://goo.gl/tbmdgw.

Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, G. & Soffer, D. (2013). Estimating the effect of web-based homework. In Lane, Yacef, Mostow & Pavlik (eds) Proceedings of the 16th International Conference on Artificial Intelligence in Education. Springer Berlin Heidelberg. 7926: 824-827.

Koedinger, K.R., Booth, J.L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. Science. 342: 935-937.

Koedinger, K., McLaughlin, E. & Heffernan, N. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. Journal of Educational Computing Research. Baywood Publishing. 4: 489-510.

Koedinger, K. & McLaughlin, E. (2016). Closing the Loop with Quantitative Cognitive Task Analysis. In Barnes, Chi & Feng (eds.) The 9th International Conference on Educational Data Mining. 412-417.

Lombrozo, T. (2014, June 2). Science, Trust And Psychology In Crisis. National Public Radio Topics. June 3, 2014, from http://www.npr.org/blogs/13.7/2014/06/02/318212713/science-trust-and-psychology-in-crisis.

Mazzocco, M.M.M., Murphy, M.M., Brown, E.C., Rinne, L. & Herold, K.H. (2013). Persistent consequences of atypical early number concepts. Frontiers in Psychology. 4, 486.

Mendicino, M., Razzaq, L. & Heffernan, N.T. (2009). Improving learning from homework using intelligent tutoring systems. Journal of Research on Technology in Education. 41 (3): 331-346.

Miller, G., Zheng, Y., Means, B., & Van Brunt Miller, B. (2013). Next generation learning challenges wave II. Retrieved on 2/19/16 from https://docs.google.com/file/d/0B2X0QD6q79ZJUl9Kd2JuVTN0VWhTYVRhX254QV85Njdqc1Vj/edit?pli=1.

Mosteller, F. & Boruch, R. F. (2002). Evidence Matters: Randomized Trials in Education Research. Washington, DC; Brookings Institution Press. Retrieved on August 5, 2014 from http://www.ncspe.org/publications_files/Rev3.pdf

National Research Council (2002). Scientific research in education. Washington, DC: The National Academies Press. Retrieved on 2/19/16 from http://www.nap.edu/catalog/10236/scientific-research-in-education.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., Heffernan, C. (2014). Population validity for Educational Data Mining models: A case study in affect detection. British Journal of Educational Technology, 45 (3), 487-501 Retrieved on August 5, 2014 from http://goo.gl/V6AJbY

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science. 349 (6251). DOI: 10.1126/science.aac4716

Ostrow, K. & Heffernan, C. (2014). ASSISTments TestBed Resource Guide: Analyze Data. Retrieved from https://sites.google.com/site/assistmentstestbed/the-data

Ostrow, K.S., Selent, D., Wang, Y., Van Inwegen, E.G., Heffernan, N.T., & Williams, J.J. (2016). The Assessment of Learning Infrastructure (ALI): The theory, practice, and scalability of automated assessment. In Proceedings of the 6th International Conference on Learning Analytics & Knowledge. ACM. 279-288.

Pashler, H., Rohrer, D., Cepeda, N. & Carpenter, S.K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. Psychonomic Bulletin & Review. 14(2), 187-193.

Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. & Sohl-Dickstein, J. (2015) Deep Knowledge Tracing. Neural Information Processing Systems (NIPS) 2015 Retrieved from http://arxiv.org/pdf/1506.05908.pdf

Prinz, F., Schlange, T. & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? Nature Rev. Drug Discov. 10(9), 712.

Rafferty, A.N. and Griffiths, T. L. (2014). Diagnosing Algebra Understanding via Bayesian Inverse Planning. Proceedings of the 7th International Conference on Educational Data Mining (pp. 351-352).

Rau, M., Aleven, V., Rummel, N., & Rohrbach, S. (2012). Sense Making Alone Doesn't Do It: Fluency Matters Too! ITS Support for Robust Learning with Multiple Representations. In S. Cerri, W. Clancey, G. Papadourakis & K. Panourgia (Eds.), Intelligent Tutoring Systems (Vol. 7315, pp. 174-184). Berlin / Heidelberg: Springer.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? Journal of Experimental Psychology: Applied, 15(3), 243.

Roediger, H. L. & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. Perspectives on Psychological Science, 1, 181-210.

Roll, I., Holmes, N. G., Day, J., & Bonn, D. (2012). Evaluating metacognitive scaffolding in guided invention activities. Instructional Science, 40, 691-710. doi:10.1007/s11251-012-9208-7

Roschelle, J., Feng, M., Murphy, R. & Mason, C. (2016). Online Mathematics Homework Increases Student Achievement. AERA OPEN. October-December 2016, Vol. 2, No. 4, pp. 1–12. DOI: 10.1177/2332858416673968

Selent, D., Patikorn, T. & Heffernan, N. T. (2016) ASSISTments Dataset from Multiple Randomized Controlled Experiments. "Work in Progress" presented at Learning at Scale 2016. At ACM Digital Library. 181-184.

SI2-SSE&SSI, National Science Foundation. (2014). Adding research accounts to the ASSISTments' platform: Helping researchers do randomized controlled studies with thousands of students. Award # 1440753. Abstract retrieved from: http://www.nsf.gov/awardsearch/showAward?AWD_ID=1440753.

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, Bull, Kay, & Mitrovic (eds) Proceedings of the 15th International Conference on Artificial Intelligence in Education Conference. Springer Berlin Heidelberg. 6738: 328–336.

Soffer, D., Das, V., Pellegrino, G., Goldman, S., Heffernan, N., Heffernan, C., & Dietz, K. (2014). Improving long-term retention of mathematical knowledge through automatic reassessment and relearning. American Educational Research Association Conference, Division C - Learning and Instruction / Section 1c: Mathematics.

Stamper, J., Eagle, M., Barnes, T. & Croy, M. (2013). Experimental Evaluation of Automatic Hint Generation for a Logic Tutor. Intl. Journal on AI in Education (IJAIED), 22(1-2), 3-17.

Storm, B.C., Bjork, E.L., Bjork, R.A., & Nestojko, J.F. (2006). Is retrieval success a necessary condition for retrieval-induced forgetting? Psychonomic Bulletin & Review. 13, 1023-1027.

Sweller, J., Kirschner, P.A., Clark, R.E. (2007). Why Minimally Guided Teaching Techniques Do Not Work: A Reply to Commentaries. Educational Psychologist. 42(2), 115-121.

The Economist. (2013, Oct 19). Unreliable Researcher: Trouble in the Lab. The Economist. Retrieved on June 3, 2014, from http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.

U.S. Department of Education. (2010). Transforming American education: Learning powered by technology, the national educational technology plan. Office of Educational Technology, Washington, D.C.

Walker, E., Rummel, N., & Koedinger, K. R. (2011). Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. International Journal of Computer-Supported Collaborative Learning, 6(2),279-306.

Walkington, C. (2013). Using learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. Journal of Educational Psychology, 105(4), 932-945.

Williams, J. J., Paunesku, D., Haley, B., & Sohl-Dickstein, J. (2013, July). Measurably Increasing Motivation in MOOCs. In Proceedings of the 1st Workshop on Massive Open Online Courses at the 16th Annual Conference on Artificial Intelligence in Education (2013). Memphis, TN. Accesses on March 10, 2014, from http://www.moocshop.org

Yong, E. (2013, November). Psychologists strike a blow for reproducibility. Nature: News. Retrieved on 2014 17, 2014, from http://goo.gl/J3XptM

Young, L. N. & Cordes, S. (2013). Fewer things, lasting longer: The effects of emotional stimuli on quantity judgments. Psychological Science, 24(6), 1057-1059.

PART II: APPLICATION

5 Self-Determination Theory

This chapter gathers pertinent background information on Self-Determination Theory and related theories of motivation to preface Part II with a broad context and framing. This brief overview supplements the conference and journal submissions that form Chapters 6-9, in order to unify the work. The information provided here also supports the Synthesis, which summarizes the evaluation and application of ASSISTments Research and relays broader contributions to the Learning Sciences community.

5.1 Background

5.1.1 Self-Determination Theory

Established through decades of work by Ryan & Deci, Self-Determination Theory (SDT) integrates organismic and dialectical backgrounds to form a single framework of motivation inspired by personality growth and development (Ryan & Deci, 2002). Basic psychological needs form the basis of SDT, defined as constructs required for well-being that exist across cultures and developmental periods (Ryan & Deci, 2000; Ryan & Deci, 2002). The basic needs are threefold and work together synergistically to regulate motivation: autonomy, relatedness (or belonging), and competence. With regard to educational tasks, autonomy describes having an internal locus of causality for participation (Reeve, 2002), valuing or finding interest in the task at hand, and feeling a sense of personal initiative that aligns with external influences (Reeve & Jang, 2006). Autonomy is considered the most central need to well-being and growth (Koestner & Losier, 2002). Belonging describes a feeling of connection to others and to one's community (Ryan & Deci, 2002), and competence describes feeling capable and effective (Ryan & Deci, 2002). Together, these needs are thought to nurture regulatory and motivational drives that promote success in self-determined behaviors.

5.1.2 Four Mini-Theories Supporting SDT

SDT grew from previous work by Ryan & Deci, with four mini theories driving the development of a more connected and congruent framework. The first of these mini-theories, Cognitive Evaluation Theory (Deci, 1975; Deci & Ryan, 1980) expanded on DeCharms' (1968) locus of causality to emphasize how social contexts can influence intrinsic motivation. Early work in this area established that tangible rewards (Deci, Koestner, & Ryan, 1999), punishment (Deci & Cascio, 1972), negative feedback (Vallerand & Reid, 1984), competition (Deci, Betley, Kahle, Abrams, & Porac, 1981), deadlines (Amabile, DeJong, & Lepper, 1976), surveillance (Lepper & Greene, 1975; Plant & Ryan, 1985), pressuring climates (Ryan, 1982), imposed goals (Mossholder, 1980), and evaluation (Smith, 1975; Ryan, 1982) could all undermine intrinsic motivation. In contrast, positive feedback in the form of praise and verbal rewards, especially when coupled with autonomy and competence, have been shown to strengthen intrinsic motivation (Deci, Kostner, & Ryan, 1999). As such, this mini-theory taps into the basic needs of autonomy and competence, and interventions meant to internalize one's locus of causality or emphasize their perceived competence should strengthen their intrinsic motivation.

As their work progressed, Deci & Ryan (1985) transitioned to the promotion of Organismic Integration Theory, which delineated between the internalization and integration of values to define subtle and complex dynamics observed within extrinsic motivation. This mini-theory integrated Bandura's concept of internalization (1996) into a continuum including amotivation, extrinsic motivation, and intrinsic motivation. When amotivated, people accept a state of learned helplessness (Seligman, 1975), depleted competence (Bandura, 1977), or empty value (Ryan, 1995) and act without regulation or drive. On the opposite end of the continuum, when intrinsically motivated, people act out of personal interest, value, or satisfaction. Although the rest of the continuum focuses on extrinsic motivation, the theory posits that the concept is more complicated than just "felt motivation with an external locus of causality." Deci & Ryan (1985) theorized that extrinsic motivation could differ drastically based on various regulatory sources ranging from external to internal. When regulation is external, people complete tasks to satisfy outward social contingencies (i.e., to earn grades). Slightly more internal, introjected regulation suggests that tasks are partially internalized but are not 'owned' and are instead performed to avoid guilt or to inflate one's ego. Even more internal, identified regulation demands conscious valuing of the task as personally significant. Finally, integrated regulation occurs when tasks are fully congruent with one's goals and valued as part of the self (Ryan & Deci, 2002; Hodgins & Knee, 2002).

It is thought that capacity for assimilation along this continuum increases with the stages of cognitive development, suggesting that older children and adults have a greater capacity for identified and integrated extrinsic motivation (Katz & Assor, 2007; Piaget, 1971). Further, the theory posits that external regulation can be achieved by supporting perceived competence, while introjection requires the addition of supports for belonging, and integration requires the addition of supports for both belonging and autonomy (Ryan & Deci, 2002; Hodgins & Knee, 2002; Katz & Assor, 2007). Integration also has more recent links to the practice of mindfulness, in that autonomous regulation begets existing in the present (Hodgins & Knee, 2002; Langer, 1989; Gunaratana, 1992; Kabat-Zinn, 1994).

Simultaneous to the development of Organismic Integration Theory, work on Causality Orientations Theory (Deci & Ryan, 1985a) isolated individual differences in how people respond to social contexts with various levels of autonomy support. Three orientations were identified as typifying responses: autonomy orientation, in which behaviors are regulated based on the self; controlled orientation, in which behaviors are regulated by external controls; and impersonal orientation, in which behaviors are regulated by ineffectance (Ryan & Deci, 2002).

The final mini-theory, Basic Needs Theory (Ryan & Deci, 2000) was established to explain associations between motivation and well-being. This mini-theory suggested that goals can be distracting and can undermine intrinsic motivation if they stray from basic need satisfaction (Ryan & Deci, 2002). In effect, this suggests that educational pursuits can have negative consequences for learners in the absence of supports for autonomy, belonging, and competence. Parents also play a critical role in the development of students' early values, and Ryan & Deci (2002) note that parenting styles that fail to satisfy children's basic needs and instead promote extrinsic goals (e.g., wealth or fame) may establish "need substitutes" that are chased throughout life to minimal benefit (Deci, 1980).

5.1.3 Continuum of Motivation and Regulation

The continuum of motivation supported by Organismic Integration Theory grew to be fully endorsed by SDT. The continuum was adapted (Ryan & Deci, 2000) for the present work in Figure 5.1. The collective works of Ryan & Deci suggest that motivation can be altered along this continuum by intervening on locus of causality and adjusting regulatory processes. Chapters 7-9 of Part II of this work present randomized controlled trials targeting each of the basic needs, with the goal of supporting progress along the continuum toward integrated regulation.

It is important to note that one need not reach intrinsic motivation for success and well-being to be attained. Given that children are less capable of higher levels of self-regulation, and considering the uninteresting nature of many educational tasks, it is not necessarily shocking that research has shown stronger outcomes amongst students exhibiting internalization and identification than those with intrinsic motivation (Koestner & Losier, 2002). While intrinsic motivation is associated with greater short-term progress and stronger immediate emotions, identification has been shown to afford stronger long-term outcomes (Koestner & Losier, 2002). Identification has also been linked to greater engagement with educational activities, less drop out, and more effective educational transitional periods (Koestner, Losier, et al., 1996). As such, the interventions assessed in Part II are intended to move students along the motivational continuum, toward more intrinsically regulated forms of extrinsic motivation. Progress in this direction is thought to be achieved by support for autonomy and is associated with greater enjoyment of educational tasks, higher self-esteem (Deci, Schwartz, Sheinman, & Ryan, 1981), greater school satisfaction (Vallerand et al., 1989), greater school performance, and higher conceptual understanding (Deci, Vallerand, Pelletier & Ryan, 1991). When supports for autonomy are removed and regulation grows more external and controlling, consequences can include higher anxiety and maladjustment (Ryan &

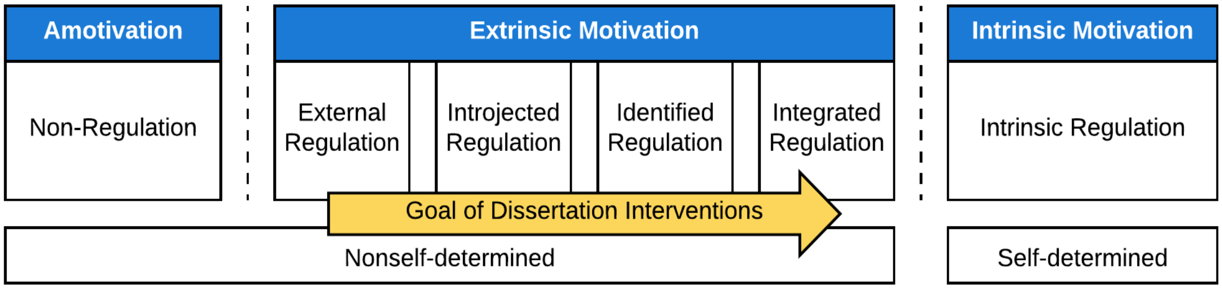


Figure 5.1. Framework of Self-Determination Theory, adapted from (Ryan & Deci, 2000).

Connell, 1989). Thus, by supporting autonomy and competence within students' mathematics assignments, it may be possible to affect learning in positive ways.

It is also important to note that in pursuit of helping students to internalize their external regulation, not all goals and motives promote well-being and happiness. Goals that undermine basic psychological needs can suppress well-being, as suggested by Basic Needs Theory (Ryan & Deci, 2002). Oftentimes, educational tasks are externally regulated, and students lack interest, value, or a sense of autonomy for their completion. For instance, standardized testing presents a high stress environment lacking supports for autonomy and competence, leading students to exhibit poorer conceptual learning and lower long-term retention (Grolnick & Ryan, 1987). External and controlling regulations lead to compliance or defiance rather than self-determined behavior (Deci, Vallerand, Pelletier, & Ryan, 1991). Moving academic environments away from testing and toward autonomy and competence supportive tasks has been shown to lead to stronger learning and memory outcomes (Grolnick & Ryan, 1987; Grolnick, Ryan, & Deci, 1991). Deci, Vallerand, Pelletier, & Ryan (1991), leaders in the area of SDT, noted that "the highest quality of conceptual learning seems to occur under the same motivational conditions that promote personal growth and adjustment" (p. 326). The randomized controlled trials presented in Chapters 7-9 aim to provide evidence for this reasoning, promoting motivation through interventions to support students' autonomy, belonging, and competence.

5.2 Motivational Subscales Under Consideration

5.2.1 Intrinsic Motivation Inventory

The Intrinsic Motivation Inventory (IMI) is a multidimensional scale intended to measure participants' subjective experiences of tasks, most often applied in laboratory settings (Intrinsic Motivation Inventory, 2018; Ryan, 1982). Iterations of the IMI have proven valid and reliable across tasks, conditions, and settings (Intrinsic Motivation Inventory, 2018). With seven subscales that can be used independently from one another and with slight modification without obvious detriment to reliability or validity, the IMI allows for self-report of interest/enjoyment, perceived competence, effort, value/usefulness, pressure and tension, perceived choice, and perceived relatedness or belonging. The perceived relatedness subscale is a more recent addition to the IMI and does not have well-established evidence of reliability (Intrinsic Motivation Inventory, 2018).

Reliability of the IMI subscales is assessed within the context of ASSISTments in Chapter 6. Scales mirroring the basic psychological needs defined by SDT are assessed and items are reduced where possible to increase reliability and reduce scale obtrusiveness. The Interest/Enjoyment subscale, considered the primary measure of intrinsic motivation, is also considered in this work. Findings from Chapter 6 are then applied to Chapters 7, 8, and 9. Students receive IMI subscale components in Chapter 7 to examine how an autonomy intervention influences performance, interest/enjoyment, and perceived autonomy, and in Chapter 9 to examine how support for competence, and thereby intrinsic motivation, is moderated by prior knowledge level and students' confidence estimates. Scale items for each chapter were modified slightly to reflect an academic task or setting - changes thought to be inconsequential to reliability outcomes (Intrinsic Motivation Inventory, 2018). Across all subscales, each scale item asked students to indicate how true the statement was for them using a Likert scale (1 = *Not at All* True, 7 = *Very True*). The full inventory in its original form can be accessed from http://selfdeterminationtheory.org/questionnaires/, and evidence for the reduced subscales as applied within ASSISTments can be found in Ostrow & Heffernan (In Press).

5.2.2 Psychological Sense of School Belonging

Given that the perceived relatedness subscale of the IMI has not been proven as a valid measure in past work (Intrinsic Motivation Inventory, 2018) and lacks evidence of validation within ASSISTments (see Chapter 6 - Ostrow & Heffernan, In Press), the belonging study instead implements the Psychological Sense of School Membership Scale (PSSM). The scale includes 18 Likert scaled items (1= "*Not at all true*" to 5 = "*Completely true*") that have been shown to have high internal consistency ($\alpha = 0.77$ to 0.88) and it is well documented as a valid measurement tool amongst middle school students (grades 6-8) (Goodenow, 1993). The scale has five reversed items, which after recoding are averaged with remaining items to form a single metric. Higher values indicate a stronger sense of belonging.

Past work in the scale's development showed the PSSM to have high construct validity, with significant correlations to students' self-reported motivation, grades and GPA, and teacher-rated student effort (Goodenow, 1993). Past work also suggested that scale items form three clusters representing caring relationships (i.e., Item 4, "Other students in this school take my opinions seriously"), acceptance (i.e., Item 13, "I can really be myself at this school"), and rejection (i.e., Item 6, "Sometimes I feel as if I don't belong here.") (You, et al., 2011). This measure is considerably longer than the IMI subscales, and efforts were not taken to reduce the scale prior to its implementation. However, scale reliability was assessed as the first research question within Chapter 8 and high reliability was achieved. The full scale in its original form can be accessed from Goodenow's original publication (1993) and the items are described as applied in Chapter 8.

5.3 Present Application

Part II of this dissertation explores the three basic needs of SDT through a series of randomized controlled trials within ASSISTments. The studies serve as applications of the tools that form ASSISTments Research. Chapter 6 presents a traditional psychometric approach to assessing scale reliability for the IMI within ASSISTments. The work offers a formal approach to instrument validation that is not often observed in the learning analytics and educational data mining communities but that is commonplace in broader psychological study. Chapters 7-9 then focus on learning interventions targeting autonomy, belonging, and competence with the goal of improving students' academic performance, motivation, and enjoyment. The studies present various levels of iterative work, some building on past research within ASSISTments and some featuring novel hypotheses based on the SDT framework. Following Part II, the Synthesis presents commentary on both the evolution and application of the ASSISTments TestBed and ALI, including lessons learned from the SDT studies, guidance for the future of ASSISTments Research, and a summary of the promising contributions of this work to the Learning Sciences community.

References

Amabile, T. M. , DeJong, W., & Lepper, M. (1976). Effects of externally imposed deadlines on subsequent intrinsic motivation. *Journal of Personality and Social Psyc*hology, 34, 92-98.

Bandura, A. (1977) Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 191-215.

Bandura, A. (1996). *Self-efficacy: The exercise of control*. New York: Freeman.

DeCharms, R. (1968). *Personal causation*. New York: Academic Press.

Deci, E. L. (1975). *Intrinsic motivation*. New York: Plenum.

Deci, E. L. (1980). *The psychology of self-determination*. Lexington, MA: D. C. Heath.

Deci, E. L., Betley, G., Kahle, J., Abrams, L., & Porac, J. (1981). When trying to win: Competition and intrinsic motivation. *Personality and Social Psychology Bulletin*, 7, 79-83.

Deci, E. L. & Cascio, W. F. (1972). Changes in intrinsic motivation as a function of negative feedback and threats. Presented at the Eastern Psychological Association. Boston.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin, 125, 627-668.

Deci, E.L. & Ryan, R.M. (1980). Self-determination Theory: When Mind Mediates Behavior. The Journal of Mind and Behavior. 1(1): 33-43.

Deci, E. L. & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Deci, E. L. & Ryan, R. M. (1985a). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, 19, 109-134.

Deci, E. L., Schwartz, A. J., Sheinman, L., & Ryan, R. M. (1981). An instrument to assess adults' orientation toward control versus autonomy with children: Reflections on intrinsic motivation and perceived competence. Journal of Educational Psychology, 73, 642-650.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*, 325–346.

Goodenow, C. (1993). The Psychological Sense of School Membership Among Adolescents: Scale Development and Educational Correlates. Psychology in the Schools. 30:79-90.

Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology, 52*, 890–898.

Grolnick, W. S., Ryan, R. M., & Deci, E. L. (1991). Inner Resources for School Achievement: Motivational Mediators of Children's Perceptions of Their Parents. Journal of Educational Psychology. 83(4): 508-517.

Gunaratana, H. (1992). *Mindfulness in plain English*. Boston: Wisdom Publications.

Hodgins, H. S. & Knee, C. R. (2002). The Integrating Self and Conscious Experience. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 87-100). Rochester, NY: University of Rochester Press.

Intrinsic Motivation Inventory, http://selfdeterminationtheory.org/questionnaires/, last accessed 2018/02/07.

Kabat-Zinn, J. (1994). *Where you go there you are*. New York: Hyperion.

Katz, I. & Assor, A. (2007). When Choice Motivates and When It Does Not. *Educational Psychology Review*. 19 (4): 429-442.

Koestner, R. & Losier, G. F. (2002). Distinguising Three Ways of Being Internally Motivated: A Closer Look at Introjection, Identification, and Intrinsic Motivation. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 101-121). Rochester, NY: University of Rochester Press.

Koestner, R., Losier, G. F., Vallerand, R. J., & Carducci, D. (1996). Identified and introjected forms of political internalization: Extending self-determination theory. *Journal of Personality and Social Psychology*, 70, 1025-1036.

Langer, E. J. (1989). Minding matters: The consequences of mindlessness-mindfulness. In L. Berkowitz (Ed.), Advances in Experimental Social Psychology: Vol. 22, (pp. 137-173). San Diego: Academic Press, Inc.

Lepper, M. R. & Green, D. (1975). Turning play into work: Effects of adult surveillance and extrinsic rewards on children's intrinsic motivation. *Journal of Personality and Social Ps*ychology, 31, 479-486.

Mossholder, K. W. (1980). Effects of externally mediated goal setting on intrinsic motivation: A laboratory experiment. *Journal of Applied Psychology*, 65, 202-210.

Ostrow, K. S. & Heffernan, N. T. (In Press). Testing the Validity and Reliability of Intrinsic Motivation Inventory Subscales within ASSISTments. To be included in the Proceedings of the 19th International Conference on Artificial Intelligence in Education (AIED 2018). London, England, June 27-30.

Piaget, J. (1971). *Biology and knowledge*. Chicago: University of Chicago Press.

Plant, R., & Ryan, R. M. (1985). Intrinsic motivation and the effects of self-consciousness, self-awareness, and ego-involvement: An investigation of internally controlling styles. *Journal of Personality*, 53, 435-449.

Reeve, J. (2002). Self-determination theory applied to educational settings. In E. L. Deci & R. M. Ryan (Eds.), Handbook of self-determination research (pp. 184-203). Rochester: University of Rochester Press.

Reeve, J., & Jang, H. (2006). What teachers say and do to support students' autonomy during a learning activity. *Journal of Educational Psychology, 98*, 209–218.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*, 450–461.

Ryan, R. M. (1995). Psychological needs and the facilitation of integrative processes. *Journal of Personality*, 63, 397-427.

Ryan, R. M. & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. Journal of Personality and Social Psychology, 57, 749-761.

Ryan, R.M. & Deci, E.L. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. American Psychologist. 55(1):68-78.

Ryan, R. M., & Deci, E. L. (2002). An overview of self-determination theory. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 3-33). Rochester, NY: University of Rochester Press.

Seligman, M. E. P. (1975). *Helplessness*. San Francisco: Freeman.

Smith, W. E. (1975). *The effect of anticipated vs. unanticipated social reward on subsequent intrinsic motivation*. Unpublished dissertation, Cornell University.

Vallerand, R. J., Blais, M. R., Brière, N. M. & Pelletier, L. G. (1989). Construction and validation of the Academic Motivation Scale. Canadian Journal of Behavioral Sciences, 21, 323-349.

Vallerand, R. J. & Reid, G. (1984). On the causal effects of perceived competence on intrinsic motivation: A test of cognitive evaluation theory. *Journal of Sport Psychology*, 6, 94-102.

You, S., Ritchey, K, Furlong, M., Shochet, I.M., & Boman, P. (2011). Examination of the latent structure of the psychological sense of school membership scale. Journal of Psychoeducational Assessment. 29(3): 225-237.

6 Validation of Intrinsic Motivation Inventory Subscales

This chapter describes a scale validation exercise intended to assess the reliability of IMI constructs within ASSISTments and to reduce IMI subscales for more efficient future use. The work applies four IMI subscales (Interest/Enjoyment, Perceived Autonomy, Perceived Relatedness (Belonging), and Perceived Competence) in support of the randomized controlled trials aligned to the basic psychological needs defined by Self-Determination Theory that are presented in Chapters 7-9.

Proper citation of this chapter is as follows:

Ostrow, K. S. & Heffernan, N. T. (In Press). Testing the Validity and Reliability of Intrinsic Motivation Inventory Subscales within ASSISTments. To appear as a full paper in the Proceedings of the 19th International Conference on Artificial Intelligence in Education (AIED, 2018).

6.1 Introduction

6.1.1 Psychometric Research in Online Learning: Value for AIED

Online learning environments allow for the implementation of psychometric scales on diverse samples of students participating in authentic learning tasks. Scales measuring personality traits, values, beliefs, motivation, and other self-reported psychological characteristics, have supported educational research for many years. However, it seems that recent opportunities for data collection at scale, made possible by omnipresent technology, have led many researchers to overlook the procedures necessary to ensure valid measurement.

Although this may not seem like an issue of particular interest to the AIED community, it should be of critical concern. Validating a measure in a learning environment before its formal use strengthens the validity and reliability of resulting claims. AIED researchers commonly focus on advancing models of student learning or affect [4, 5]. Models featuring data collected from a clickstream or sensors can be supplemented by student self-reports from psychometric scales to explain additional variance or reduce error. A recent article in the *Journal of Learning Analytics* highlighted psychometric variables relevant to academic performance including measures of cognitive ability, temperament, personality, motivation, and learning strategies [6]. Although researchers tend to cite published reliability statistics before implementing popular psychometric scales, few employ the exploratory or confirmatory factor analyses (or similar methods) necessary to validate use of the scale in their specific domain, population, and/or learning environment. This is not to say that these techniques are completely foreign to researchers in the field; one positive example observed during a review of related literature established and validated a measure of learners' perceptions of pedagogical agents prior to its use in further research [7]. In contrast, it is common practice in psychology to cite an initial publication as proof of a scale's validation prior to its formal use. As such, the AIED community may benefit from stronger approaches to psychometric application.

A concrete example of the importance of scale validation stems from recent focus in the AIED community toward personalization [8]. Researchers tackle this problem by using learner analytics, data mining techniques, and randomized controlled trials to isolate the most effective learning interventions for each student based on a set of predefined characteristics. In such contexts, self-report measures from psychometric scales can provide an opportunity to explain additional variance between students. Scale scores can be used as dependent measures for the purpose of prediction (i.e., "students with *high prior knowledge* are more likely to report feeling *competent*") or as independent variables for exploring interactions or mechanism (i.e., "students with *high perceptions of autonomy* outperformed those with *low perceptions of autonomy* differentially by treatment condition"). Thus, it is critical to strengthen these metrics by taking steps to validate psychometric scales within specific domains, populations, and/or learning environments.

6.1.2 Self-Determination Theory and the Intrinsic Motivation Inventory

The Intrinsic Motivation Inventory (IMI) [9] is a multidimensional scale developed in support of Self-Determination Theory (SDT) [1, 2, 3], a strongly validated theory claiming that motivation and regulation are guided by three innate needs: autonomy, belonging, and competence. As applied to education, this theory posits that students who perceive volition in a task, those who report stronger connections with peers and teachers, and those who perceive themselves as competent in the task at hand are more likely to internalize the task and excel. It has been shown that promotion of these needs in educational environments can lead to higher quality learning, as well as greater conceptual understanding, personal growth, and positive adjustment [10]. If validated in an online learning environment, the IMI could potentially be used to inform stakeholders of students' motivational and regulatory styles, alerting them to pertinent implications for learning outcomes and appropriate interventions.

In their landmark review outlining the growth of SDT, Ryan & Deci [2] cite applications of the theory across research domains including education, health care, religion, health and exercise, political activity, environmental activism, and intimate relationships. The IMI has also been applied broadly, with past work validating versions of its subscales in contexts including sports and competition [11, 12], reading [13], mathematics [14], language learning [14], psychiatry [15], medicine [16], puzzle completion [17], computer tasks [18], and teacher training [19]. Past work has also shown IMI subscales to have strong temporal reliability [20]. While examples of IMI application have clearly varied by domain, task, and sample population, it is important to note that they have also varied by scale and item inclusion, scale and item order, and data collection environment. As such, the developers of the IMI encourage researchers to validate the scale within their specific domains, populations, and/or environments of interest [9].

6.1.3 The Present Work

The present work provides an example of scale validation in an online learning environment using iterative exploratory factor analysis and item reduction techniques. ASSISTments (www.assistments.org), an online learning environment known for its embrace of educational research at scale [21, 22], is currently hosting a series of randomized controlled trials examining learning interventions that target the innate needs defined by SDT with the goal of promoting integrated learning and thereby improving student performance. In support of this research, the present work attempts to validate four subscales of the IMI measuring students' perceptions of autonomy, belonging (or relatedness), competence, and interest/enjoyment within ASSISTments. Validation of IMI subscales within ASSISTments is valuable because students' perceptions can serve as powerful independent or dependent measures when isolating effective learning interventions. Goals of the present work are to achieve convergent, discriminant, and face validity for each subscale, to achieve high reliability for each subscale, and to reduce the number of items within each subscale for future implementation. The latter goal will make future data collection less obtrusive (by requiring fewer items), thereby allowing survey efforts to more easily scale to the broader ASSISTments user population (approximately 50,000 users).

6.2 Methods

6.2.1 Sample

Five teachers who regularly work with ASSISTments were contacted with the request that their students participate in a 28-item Likert scale survey. Teachers were notified that the survey would immediately follow a brief assignment (of their choice) used for classwork or homework, and that it would add 5-10 minutes to the assignment based on students' reading levels. Four teachers chose to participate and provided assignments that were modified by the primary author to include two additional items, one introducing the survey as a data collection tool to strengthen students' experiences within ASSISTments and one providing access to the IMI subscales.

Participating teachers and their students were representative of different subpopulations and sampling styles. A total of 226 students participated in at least one of the four subscales. Students of Teacher 1 (n = 73) and Teacher 2 (n = 54) were enrolled in 7th grade math classes at two schools in two different suburban/rural locations in Massachusetts. Teacher 1 chose to embed the IMI subscales after an 8-question homework assignment. Teacher 2 split delivery of the subscales, enrolling her students in a randomized controlled trial including two scales (Interest/Enjoyment and Competence) for homework, and choosing to embed the remaining two scales (Autonomy and Belonging) following an 8-question classwork assignment. These two assignments were strongly conceptually linked and split scale delivery was embraced to examine the potential consequences for reliability and score interpretation within teachers. Students of Teacher 3 (n = 46) were enrolled in high school level math courses in an urban location in Massachusetts and were highly representative of ESL and low SES populations. Teacher 3 chose to embed the subscales after a "class opener" with two multiple choice questions. Students of Teacher 4 (n = 53) were enrolled in high school level engineering courses in an urban location in Massachusetts and represented accelerated learners. Teacher 4 chose to embed the survey following a 20-question assignment on velocity.

All students were familiar with ASSISTments and used the system regularly for classwork and homework in the courses in which they were surveyed. In general, students were not allowed to opt out of survey participation up front but were allowed to skip scale responses and progress to the end of their assignments at any time during their participation. For the RCT-bound scales delivered by Teacher 2, students were prompted to opt-in to survey participation causing unbalanced scale responses between subscales within Teacher 2. This caused average overall missingness (%) to vary by scale across teachers: Interest/Enjoyment (M = 13.60, SD = 0.33), Autonomy (M = 2.16, SD = 0.46), Belonging (M = 3.70, SD = 0.35), and Competence (M = 14.38, SD = 0.37). The analytic sample was reduced based on missing data using listwise deletion. This approach is appropriate for factor analysis because unbalanced items can sway factor loadings [23]. As such, results are based on samples with complete response patterns for modeled scale items.

6.2.2 Intrinsic Motivation Inventory

The IMI is a multidimensional scale intended to measure the subjective experiences of participants following task participation [9, 1]. Various iterations of the IMI have been in use for more than 30 years, with well-established validity and subscale reliability across tasks, conditions, and settings [9]. The scale has six primary subscales that can be mixed and matched to suit research needs: interest/enjoyment, perceived competence, effort, value/usefulness, felt pressure and tension, and perceived choice. A seventh subscale intended to measure perceived relatedness or belonging was added in recent years and has not yet been established as valid or reliable.

Subscales. All scale items were modified slightly to reflect an academic task or setting; such modifications are thought to be inconsequential to outcomes [9]. Students were asked to indicate how true each statement was for them using a Likert scale (1 = *Not at All True*, 7 = *Very True*). Past work has suggested that order effects of scale and item delivery are negligible and that subscales can be included or excluded as necessary [24; 9]. The four subscales considered in the present work align with the basic psychological needs defined by Self-Determination Theory, as detailed in the subsections below.

Interest/Enjoyment. This subscale is the primary measure of intrinsic motivation. It includes seven items regarding intrinsic motivation (i.e., "I enjoyed doing this assignment very much"), with two items reverse scored (i.e., "This assignment did not hold my attention at all").

Autonomy. This subscale is the primary measure of perceived autonomy, also known as choice, volition, or task-based locus of control. Scores on this scale have previously been shown to predict Interest/Enjoyment scores [9]. This subscale includes seven items regarding perceived autonomy (i.e., "I did this assignment because I wanted to"), with five items reverse scored (i.e., "I did this assignment because I had to").

Belonging. This subscale is the primary measure of perceived relatedness or belonging. This scale was added to the IMI in recent years and does not have well established validity or reliability. In addition, modifications to items in this subscale to capture how well students felt they related to their classmates may have been more significant than modifications to other scales because the effect was extrapolated to a collective group (i.e., changing "task" or "activity" to "assignment" does not extrapolate to *many tasks*). This subscale includes eight items (i.e., "I'd like a chance to interact with my classmates more often") with four items reverse scored (i.e., "I don't feel like I could really trust my classmates").

Competence. This subscale is the primary measure of perceived competence or feeling capable and confident. Scores on this scale have previously been shown to predict Interest/Enjoyment scores [9]. This subscale includes six items (i.e., "I am satisfied with my performance on this assignment") with one item reverse scored (i.e., "This was an assignment that I couldn't do very well").

6.2.3 Procedure

Data was retrieved by integrating Qualtrics, a readily accessible survey infrastructure, with ASSISTments using the ASSISTments Survey System available through the ASSISTments TestBed [25]. This system uses an iframe to establish a connection between the two platforms, resulting in the ability to link survey data to ASSISTments performance through anonymized student and assignment identification numbers. Two items were added to the end of each participating teacher's assignment: a verification item introducing the survey as a data collection tool to strengthen students' experiences within ASSISTments, and an item with an embedded iframe that connected students to the survey content in Qualtrics while they worked in ASSISTments.

IMI items were delivered through Qualtrics using subscale alignment. Except for those of Teacher 2, all students were asked to respond to all items pertaining to Interest/Enjoyment in a single page view. When finished, or if opting not to answer, they could select "Next" to move on to the next page and subscale. Students cycled through the Autonomy, Belonging, and Competence subscales in this fashion until ultimately completing the survey. Due to Teacher 2's split delivery, her students opted-in to the Interest/Enjoyment and Competence subscales after a homework assignment and received the Autonomy and Belonging subscales after a subsequent class assignment using the same protocol and infrastructure noted above.

The data collection period lasted one week. Data was retrieved from Qualtrics and ASSISTments, compiled, and preprocessed for IBM SPSS Statistics. Variables were cleaned, and missing data was labelled for proper exclusion from analysis. Redundancies were removed while merging data (24 students accessed the survey multiple times; in these cases, only first responses were retained). The resulting data file contained responses from 226 students. Items were reverse scored as necessary - a step not required for factor analysis, but critical for calculating reliability using Cronbach's α [23]. Higher score values indicated higher levels of scale sentiment across all scales (i.e., greater enjoyment). De-identified survey data is available at [26] for additional reference.

Following the guidelines set forth by Field [23] iterative scale reduction was conducted using principal axis factor analyses. Given the likelihood of correlations between subscales, oblique rotation (i.e., direct oblimin) offered a more appropriate approach than orthogonal rotation (i.e., varimax). Factors were established using traditional methods: factors with eigenvalues greater than 1.0 were considered valid for inclusion using Kaiser's criterion, and scree plots were developed to confirm factor count by estimating the point of inflexion. Items were removed as part of the iterative process to establish stronger validity and reliability. Inter-item correlations and subscale reliability measures were consulted for scale reduction. Where items were removed, factor analysis was repeated to assess potential changes to factors and loadings and to optimize the model.

6.3 Results

6.3.1 Iterative Scale Reduction

28-Item Factor Analysis. A principal axis factor analysis was conducted on all 28 items from the IMI subscales using direct oblimin oblique rotation. After listwise deletion of missing responses the analytic sample consisted of 180 students. The Kaiser-Meyer-Olkin measure verified the sample was large enough for analysis, KMO = .81, and Bartlett's Test of Sphericity was significant, χ^2 (378) = 2,818.75, p < .001 (as desired). In addition, the diagonals of the anti-image correlation matrix were above .5 (as desired). Thus, 180 students provided an adequate sample size for analysis.

This analysis was conducted to assess initial model structure and examine the potential for item reduction. The model had poor structure, as suggested by a determinant of 5.61E-008 (denoting issues of multicollinearity), numerous after extraction communalities below 0.70, and an average communality of 0.58 denoting that Kaiser's criterion was not necessarily an appropriate threshold for factor inclusion. The model resolved to six factors using Kaiser's criterion and four or six factors based on interpretation of the scree plot's point of inflexion. Four factors, as desired based on the initial subscales, accounted for 56% of the variance in the model, with the remaining two factors accounting for an additional 11%.

The correlation matrix attained from this analysis (Table 1) was helpful in reducing scale items to establish a stronger model. Field [23] suggests beginning the reduction process by assessing the correlation matrix for multiple inter-item correlations over .90 or under .30. As none of the correlations exceeded .90, items were slated for removal from the model if 50% or more of the within-scale inter-item correlations were .30 or less (i.e., item E4 had 3/6 correlations of .30 or smaller; suppressed correlations are all less than .30). Using this approach, eight items were removed from the model: Interest/Enjoyment Item 4 (50%), Autonomy Item 6 (50%), Belonging Items 2 (57%), 3 (57%), 4 (86%), 5 (57%), and 7 (57%), and Competence Item 6 (100%).

20-Item Factor Analysis. Following removal of these eight items, a second principal axis factor analysis was conducted on the remaining 20 items, again using direct oblimin oblique rotation. The analytic sample again consisted of an adequate sample size of 180 students: the Kaiser-Meyer-

Olkin measure verified the sample was large enough for the analysis, KMO = .84, and Bartlett's Test of Sphericity was significant, χ^2 (190) = 2,066.79, p < .001.

Model structure was still not ideal. The expected structure was resolved, with four factors retained (using Kaiser's criterion), explaining 66% of the variance and with all items loading on their expected subscales. However, although the determinant increased to 5.84E-006, it remained lower than the desired minimum of 1.00E-005, suggesting that a multicollinearity issue remained and that additional items should be considered for removal.

Having addressed the issue through the correlation matrix, no additional reductions were suggested using this approach. Thus, reliability analyses were conducted on each subscale to determine candidates for removal. Cronbach's α was used with listwise deletion of missing values by scale. Analysis suggested that reliability of the six remaining items in the Interest/Enjoyment subscale was high, α = .91 (n = 191), but could be increased by removing Item 3. Reliability for other subscales was mixed, as shown in Table 2, but no other items met qualifications for removal using this approach. Therefore, Interest/Enjoyment Item 3 was removed from the model, leaving 19 items.

19-Item Factor Analysis. Following item removal, a final principal axis factor analysis was conducted on the remaining 19 items, again using direct oblimin oblique rotation. Listwise deletion of missing data left an analytic sample of 181 students. The Kaiser-Meyer-Olkin measure verified the sample was large enough for the analysis, KMO = .83, and Bartlett's Test of Sphericity was significant, χ^2 (171) = 1,988.72, p < .001.

This final iteration established adequate model structure. Four factors were retained using Kaiser's criterion, explaining 68% of the variance, with all items loading on their expected subscales. Four factors were also suggested via scree plot interpretation, as shown in Figure 1. The determinant increased to 1.01E-005, just surpassing the threshold of 1.00E-005 and suggesting that multicollinearity had been sufficiently resolved. Average communality increased to 0.595, bordering the 0.60 threshold for adequacy of the Kaiser criterion. Additionally, only 12% of the residuals in the reproduced matrix had absolute values greater than 0.05. Reliability results were unchanged from those presented in Table 2, with the noted increase in reliability for the Interest/Enjoyment subscale following item removal. Further, all corrected item-total correlations were above 0.30, suggesting that each item correlated well with its respective overall scale score. Values were lowest for items in the Belonging subscale, suggesting issues with the validity of using this subscale alongside others in the IMI. Overall reliability of all 19 items in the model was moderate, α = .78.

6.3.2 Resulting Subscales in the 19-Item Model

The four resulting factors aligned with expected subscales. Table 3 provides each scale item with factor loadings from the pattern matrix after rotation. The pattern matrix was used for interpretation

Table 6.1. Inter-item correlations in the 28-item model (n = 180).

	Interest/Enjoyment							*Autonomy*							*Belonging*								*Competence*					
Scale/Item	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	1	2	3	4	5	6
Enjoyment																												
E1	--																											
E2	**.83**	--																										
E3	**.52**	**.53**	--																									
E4	**.32**	**.33**	**.62**	--																								
E5	**.69**	**.67**	**.43**	**.19**	--																							
E6	**.75**	**.77**	**.47**	**.30**	**.74**	--																						
E7	**.62**	**.64**	**.34**		**.70**	**.65**	--																					
Autonomy																												
A1	**.42**	**.37**	**.36**	**.21**	**.41**	**.35**	**.36**	--																				
A2	.15		**.24**	**.25**	.18		.14	**.39**	--																			
A3	.17	.17	**.32**	**.24**	**.21**	.17		**.43**	**.53**	--																		
A4	.13	.15	**.20**	.17	**.24**	.15	**.18**	**.31**	**.39**	**.53**	--																	
A5	**.24**	**.20**	**.30**	**.25**	**.30**	**.31**	**.23**	**.51**	**.42**	**.57**	**.48**	--																
A6	**.47**	**.52**	**.37**	**.21**	**.56**	**.43**	**.44**	**.45**	**.24**	**.31**	**.21**	**.45**	--															
A7	**.18**		**.23**	.17	**.26**	**.22**	**.22**	**.39**	**.36**	**.40**	**.58**	**.59**	**.24**	--														
Belonging																												
B1	-.13				-.14	-.14								.15	--													
B2													-.13	.14	**.67**	--												
B3							.15								**.23**	**.18**	--											
B4																	**.25**	--										
B5					**-.23**	-.14	-.16	**-.23**					**-.23**		**.45**	**.39**	**.26**		--									
B6										.15					**.43**	**.35**	**.38**		**.47**	--								
B7														**-.18**	.16		**.40**	**.44**	**.26**		--							
B8					.13		**.19**								**.35**	**.24**	**.64**	**.18**	**.27**	**.31**	**.47**	--						
Competence																												
C1									**-.24**	**-.18**	**-.19**	-.14			**.24**	.13	**.19**		**.18**			.16	--					
C2									**-.21**	-.16	**-.23**	**-.25**		**-.21**	**.18**	.13	**.27**				.14	**.19**	**.61**	--				
C3	.13	.17			.13		.15		**-.18**	-.13	-.15		.16	-.15	.17		**.33**				**.21**	**.26**	**.62**	**.56**	--			
C4									-.16	-.13	-.15				**.34**	**.22**	**.21**		.17		**.20**	.17	**.77**	**.61**	**.66**	--		
C5			-.14					-.12	**-.29**	**-.25**	**-.18**	**-.22**			**.28**	.15	**.29**		**.18**		**.23**	**.25**	**.78**	**.70**	**.68**	**.83**	--	
C6	**-.25**	-.15	**-.21**		**-.26**	-.15	**-.18**	**-.33**	**-.27**	-.16		**-.23**	**-.27**		**.22**	**.31**	.15		**.29**	**.24**			.15	.14	.17	**.19**	**.29**	--

Note. Bold correlations, p < .01; all others, p < .05. Suppressed correlations were not significant at p < .05.

Table 6.2. Reliability of subscales for the 20-item model and candidate items for removal.

	n	Scale Items	α	α if item removed
Interest/Enjoyment	191	6	91	.92 (Item 3)
Autonomy	220	6	.83	--
Belonging	217	3	.56*	--
Competence	192	5	.92	--

* The Belonging subscale does not have well-established validity or reliability; low reliability of this scale was not of immediate concern.

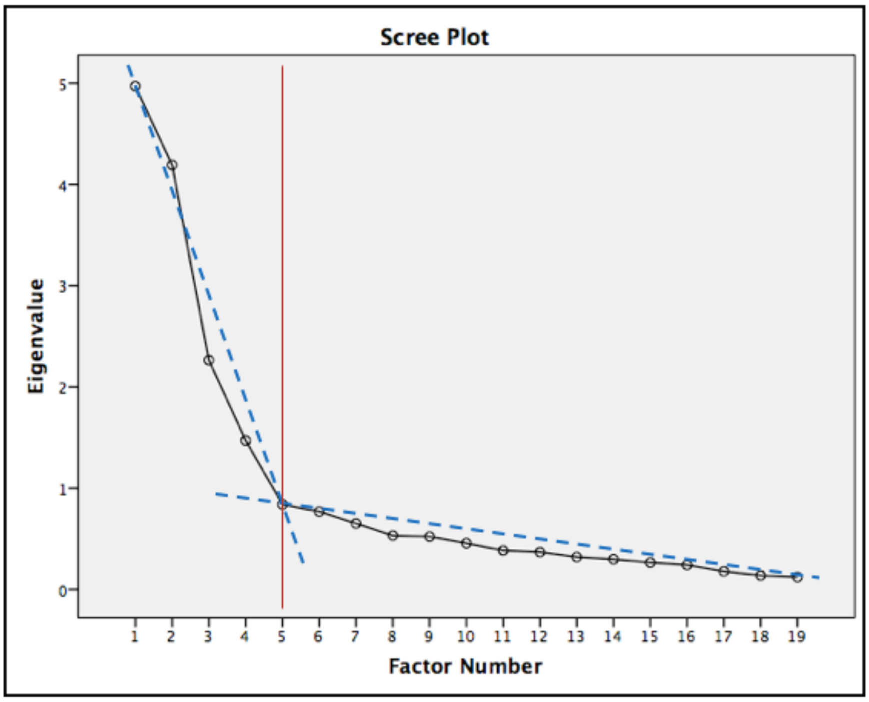


Figure 6.1. Scree plot with point of inflexion mapped. The red vertical line at factor 5 designates the non-inclusive cutoff factor, or the 'elbow' of the graph.

because it ignores shared variance and shows the unique contribution of items to factors [23]. Factor 1 aligned with the Interest/Enjoyment subscale, explaining 26.14% of model variance. Similarly, Factor 2 aligned with the Competence subscale (22.07%), Factor 3 aligned with the Autonomy subscale (11.92%), and Factor 4 aligned with the Belonging subscale (7.74%). The Interest/Enjoyment, Autonomy, and Competence subscales showed substantial convergent validity and high reliability, and all scales displayed high discriminant and face validity.

6.3.3 Class Variations

A brief investigation was conducted to examine how scale scores using the 19-item model varied across participating teachers. Split-file analysis was used to assess the reliability of each subscale using Teacher as the grouping variable. Results are shown in Table 4. Subscales showed similar patterns of reliability regardless of teacher, with some variation in magnitude

Table 6.3. Summary of exploratory factor analysis results in the 19-item model (n = 181).

Scale/Item	*Interest/Enjoyment*	*Competence*	*Autonomy*	*Belonging*
Interest/Enjoyment				
IE2 – This assignment was fun to do.	.879			
IE1 – I enjoyed doing this assignment very much.	.858			
IE6 – I thought this assignment was quite enjoyable.	.857			
IE5 – I would describe this assignment as very interesting.	.810			
IE7 – While I was doing this assignment, I was thinking about how much I enjoyed it.	.767			
Competence				
C4 – I am satisfied with my performance on this assignment.		.935		
C5 – I was pretty skilled at this assignment.		.915		
C1 – I think I am pretty good at this assignment.		.868		
C3 – After working at this assignment for a while, I felt pretty competent.		.709		
C2 – I think I did pretty well at this assignment, compared to other students.		.664		
Autonomy				
A5 – I did this assignment because I had no choice. (R)			.780	
A3 – I didn't really have a choice about doing this assignment. (R)			.720	
A7 – I did this assignment because I had to. (R)			.704	
A4 – I felt like I had to do this assignment. (R)			.663	
A2 – I felt like it was not my own choice to do this assignment. (R)			.581	
A1 – I believe I had some choice about doing this assignment.			.518	
Belonging				
B6 – I'd really prefer not to interact with my classmates in the future. (R)				.665
B1 – I feel really distant to my classmates. (R)				.643
B8 – It is likely that my classmates and I could become friends if we interacted a lot.				.492
Eigenvalues	4.97	4.19	2.27	1.47
% of variance	26.16	22.07	11.92	7.74
α	.92	.92	.83	.56

Table 6.4. Subscale reliability by teacher.

	Teacher 1		Teacher 2		Teacher 3		Teacher 4	
Scale	n	α	n	α	n	α	n	α
Interest/Enjoyment	73	.93	27	.88	38	.92	53	.93
Autonomy	71	.81	54	.77	41	.78	53	.88
Belonging	71	.71	54	.29	39	.23	53	.75
Competence	70	.93	29	.82	40	.90	53	.82

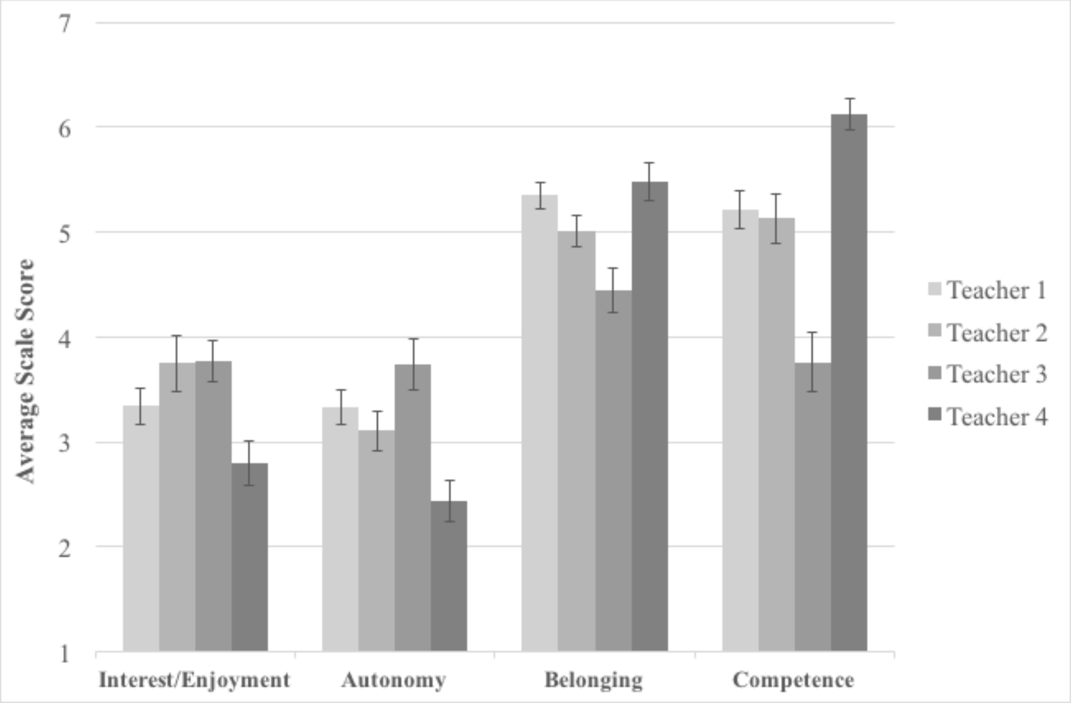


Figure 6.2. Resulting scale scores by teacher.

across teachers. Of note, Teacher 2 and Teacher 3 exhibited the lowest reliability on the Belonging subscale, while Teacher 2 and Teacher 4 exhibited lower than anticipated reliability on the Competence subscale.

Scale scores were defined for each student by averaging factor items and then aggregated by teacher for final comparison (see Figure 2). Such aggregates offer an example of how this psychometric scale could be used to establish useful variables or covariates for future research. ANOVAs revealed significant differences between teachers on all subscales: Interest/Enjoyment, F (3, 190) = 3.71, $p < .05$; Autonomy, F (3, 218) = 6.98, $p < .001$; Belonging, F (3, 216) = 6.76, $p < .001$; and Competence, F (3, 191) = 21.08, $p < .001$. Given numerous confounds in present survey collection (e.g., teacher, assignment, skill level, age range) further assessment was not considered. However, future work could control for potential sources of variance to better define the mechanisms underlying these significant differences.

6.4 Contributions & Limitations

The goal of the present work was to validate subscales of the IMI within ASSISTments to support a series of randomized controlled trials assessing the efficacy of learning interventions that target students' perceptions of autonomy, belonging, and competence. Employing an iterative factor analysis and item reduction approach with an analytic sample of 181 students established substantial convergent validity and high reliability for three reduced IMI subscales (Interest/Enjoyment, Autonomy, and Competence). Issues were observed with the reliability of the Belonging subscale, due in part to the high proportion of items removed to optimize the model. This subscale is not well-established as valid or reliable [9] and it did not perform well within ASSISTments. As such, the Belonging subscale will not be used in future data collection efforts and students' aggregate scores on this subscale will not be used in future analyses. Results also suggested that all four subscales exhibited high discriminant and face validity.

Limitations of this work include potential bias introduced by varied delivery protocol of subscales (i.e., Teacher 2's split delivery), delivery of items aligned within subscales (although previously addressed as inconsequential for psychometric scales [9, 24]), and the potential for reduced external validity due to item reduction. The randomized controlled trials supported by this work will use the subscales established by the 19-item model when compiling aggregate scores for use in future analyses. The reduction of 9 items from these subscales will also make future data collection using the IMI within ASSISTments less obtrusive, thereby allowing survey efforts to scale. With hope, these results also serve as a valuable reminder for the AIED community that contextually validating a psychometric scale prior to its formal use strengthens the validity and reliability of resulting claims.

References

ASSISTments TestBed Resource Guide. (2018). Retrieved: https://www.assistmentstestbed.org.

Brusilovsky, P., Millan, E. (2014). User Models for Adaptive Hypermedia and Adaptive Educational Systems. In Brusilovski, Kobsa, & Nejdl (eds.) The Adaptive Web: Methods and Strategies of Web Personalization. LNCS 4321, Springer-Verlag Berlin Heidelberg.

Calvo, R. A., D'Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transaction on Affective Computing.* 1 (1): 18-37.

Choi, J., Mogami, T., Medalia, A. (2010). Intrinsic Motivation Inventory: An Adapted Measure for Schizophrenia Research. Schizophrenia Bulletin. 36 (5): 966-976.

Deci, E., Eghrari, H., Patrick, B., Leone, D. (1994). Facilitating internalization: The self-determination theory perspective. *Journal of Personality*. 62: 119-142.

Deci, E., Vallerand, R., Pelletier, L., Ryan, R. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist*. 26: 325-346.

Field, A. (2013). Discovering statistics using IBM SPSS Statistics. 4th edn. SAGE Publications, London.

Filak, V., Sheldon, K. (2003). Student psychological need satisfaction and college teacher-course evaluation. *Educational Psychology*. 23: 235-247.

Gray, G., McGuinness, C., Owende, P., Carthy, A. (2014). A Review of Psychometric Data Analysis and Application in Modelling of Academic Achievement in Tertiary Education. *Journal of Learning Analytics.* 1 (1): 75-106.

Grolnick, W., Ryan, R. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology*. 52: 890-898.

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. *International Journal of Artificial Intelligence in Education.* 24 (4): 470-497.

Intrinsic Motivation Inventory. (2018). From: http://selfdeterminationtheory.org/questionnaires/.

McAuley, E., Duncan, T., Tammen, V. V. (1987). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport.* 60: 48-58.

McAuley, E., Tammen, V. V. (1989). The effects of subjective and objective competitive outcomes on intrinsic motivation. *Journal of Sport and Exercise Psychology.* 11: 84-93.

Monteiro, V., Mata, L., Peixoto, F. (2015). Intrinsic Motivation Inventory: psychometric properties in the context of first language and mathematics learning. *Psicologia: Reflexao e Critica.* 28 (3): 434-443.

Ostrow, K. S., Heffernan, N. T., Williams, J. J. (2017). Tomorrow's edtech today: establishing a learning platform as a collaborative research tool for sound science. *Teachers College Record.* 119 (3): 1-36.

Ryan, R. (1982). Control and information in the intrapersonal sphere an extension of cognitive evaluation theory. *Journal of Personality and Social Psychology.* 43: 450-461.

Ryan, R. M., Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist.* 55: 68-78.

Ryan, R. M., & Deci, E. L. (2002). An overview of self-determination theory. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research*. Rochester, NY: University of Rochester Press.

Ryan, R., Mims, V., Koestner, R. (1983). Relation of reward and interpersonal context to intrinsic motivation: A review and test using Cognitive Evaluation Theory. *Journal of Personality and Social Psychology*. 45: 736-750.

Ryu, J., Baylor, A. L. (2005). The Psychometric Structure of Pedagogical Agent Persona. *Tech. Inst. Cognition and Learning.* 22: 291-314.

Santos, O. C., Kravcik, M., Boticario, J. G. (2016). Preface to Special Issue on User Modelling to Support Personalization in Enhanced Educational Settings. *International Journal of Artificial Intelligence in Education.* 26 (3): 809-820.

Schell, K. L., Oswald, F. L. (2013). Item grouping and item randomization in personality measurement. *Personality and Individual Differences*. 55: 317-321.

Submission Data. (2018). Retrieved: https://osf.io/4gwuc/

Tsigilis, N., Theodosiou, A. (2003). Temporal Stability of the Intrinsic Motivation Inventory. *Perceptual and Motor Skills*. 97 (1): 271-280.

Williams, G. C., Deci, E. L. (1998). Supporting autonomy to motivate glucose control in patients with diabetes. *Diabetes Care*. 21: 1644-1651.

7 Autonomy

This chapter describes a randomized controlled trial designed to target the basic psychological need of autonomy, thought to be the most critical need for proper adjustment and well-being. Supported by Self-Determination Theory, the work is rooted in prior research conducted using the ASSISTments TestBed examining the effects of choice and hint medium. The effects of autonomy support and hint medium (video vs. text) are considered with regard to math performance and student's prior knowledge. This chapter will read as a strong draft and will be submitted to a journal venue following completion of my degree. As such, further revision may occur before formal presentation. In addition, this paper borrows language from the overview offered in Chapter 5 of this work, which will not be submitted for publication elsewhere.

Proper citation of this chapter is as follows:

Ostrow, K. S. (In Preparation). The Option of Video Hints in an Online Math Assignment Supports the Performance and Motivation of Low Knowledge Students.

7.1 Introduction

7.1.1 Autonomy and Self-Determination Theory

Autonomy, or acting from the perspective of the self, has been an important construct underlying soft and hard sciences for centuries. Value for individual or collective forms of autonomy have been considered in philosophy, politics, ethics, medicine, computing, robotics, and human development (Dryden, 2018). In psychology, Self-Determination Theory (SDT) considers autonomy a basic psychological need, universal to the human condition and critical for well-being (Ryan & Deci, 2000; Ryan & Deci, 2002). SDT maintains a framework that integrates organismic and dialectical backgrounds into a single theory of motivation driven by regulatory responses to need fulfillment (Ryan & Deci, 2002). Effects of perceived autonomy are thought to flourish when perceived relatedness (or belonging) and perceived competence are also present. Of the three basic needs, autonomy is considered the most critical for well-being and growth (Koestner & Losier, 2002). With regard to education, autonomy describes having an internal locus of causality (Reeve, 2002), valuing or finding interest in the task at hand, and maintaining a sense of personal initiative that aligns with external influences (Reeve & Jang, 2006). Together with belonging and competence, autonomy is thought to nurture regulatory and motivational drives that promote success in self-determined behaviors.

When applied to educational tasks, typically rife with external controls and contingencies (i.e., tests and grades), the SDT framework offers an opportunity for intervention on a socio-emotional level. SDT posits that support for autonomy, belonging, and competence can promote greater internalization along a continuum of motivation. This continuum, shown in Figure 7.1 as adapted from Ryan & Deci (2000), spans from amotivation to intrinsic motivation, with incrementally

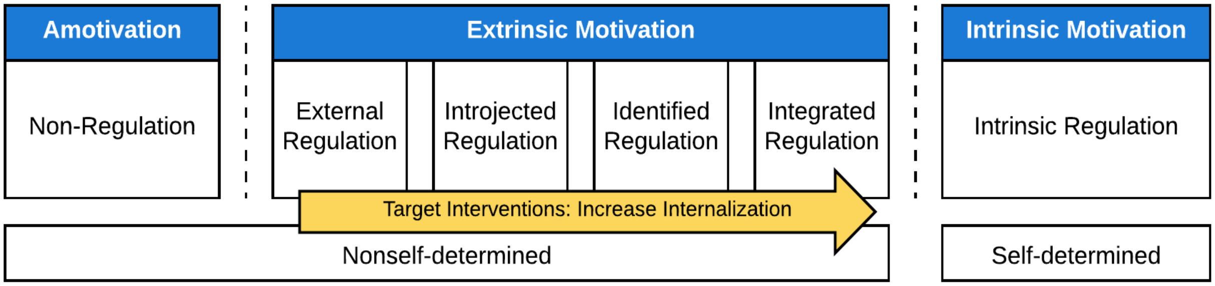


Figure 7.1. Framework of Self-Determination Theory, adapted from (Ryan & Deci, 2000).

internalized extrinsic motivation representing the majority of the scale. Support for competence is thought to result in extrinsic motivation driven by external regulation (Ryan & Deci, 2002); a common example within education is the provision of an A+ to reward strong performance and capability. To achieve introjected regulation, support for competence must be complimented by support for relatedness or belonging (Ryan & Deci, 2002). Students must feel as though they belong in their class, associating a sense of trust and safety with their teacher, peers, and academic community. To reach identified and integrated regulation, students must begin to internalize the task and it must align with their personal values and beliefs. Ownership of internalized volition can be achieved by complimenting supports for competence and belonging with support for autonomy (Ryan & Deci, 2002). In this sense, perceived autonomy is the most crucial of the basic psychological needs, promoting success and well-being through more intrinsic motivation. Past work has associated support for autonomy with greater enjoyment of educational tasks, higher self-esteem (Deci, Schwartz, Sheinman, & Ryan, 1981), greater school satisfaction (Vallerand et al., 1989), greater school performance, and higher conceptual understanding (Deci, Vallerand, Pelletier & Ryan, 1991). When supports for autonomy are removed and regulation grows more external and controlling, consequences can include higher anxiety and maladjustment (Ryan & Connell, 1989). Thus, by providing support for autonomy, especially in environments that promote relatedness and competence, it may be possible to affect learning in positive ways.

It is important to note that intrinsic motivation is an ideal and is not necessary for success and well-being to be attained. In fact, theories of cognitive development suggest that children are less capable of achieving higher levels of self-regulation (Piaget, 1971). Coupled with this theory, the understanding that most educational tasks are uninteresting and externally regulated makes it evident that motivational interventions in educational contexts should be focused on internalizing external regulation. Past work in this area noted stronger learning outcomes amongst students exhibiting identification than those exhibiting intrinsic motivation (Koestner & Losier, 2002). Further, while intrinsic motivation is associated with greater short-term progress and stronger immediate emotion, identification affords stronger long-term outcomes (Koestner & Losier, 2002). Identification has also been linked to greater engagement with educational activities, less drop out, and more effective transitional periods (Koestner, Losier, et al., 1996).

7.1.2 Supporting Autonomy in Education

It is also important to note that in the pursuit of helping students to internalize external regulation, not all goals and motives promote well-being and happiness. Goals that undermine autonomy, belonging, or competence can actually suppress well-being, as suggested by Basic Needs Theory (Ryan & Deci, 2002). This poses a complex issue within education, where most tasks are externally regulated and fail to support the three basic needs. Standardized testing is an excellent example of this conflict: educational stakeholders value standardized tests as measures of value added and as markers of student proficiency and teacher efficacy. However, standardized testing elicits a controlling environment where students can easily suffer from diminished perceptions of competence, establishing high-stress learning contexts that lead students to exhibit poorer conceptual learning and long-term retention (Grolnick & Ryan, 1987). External, controlling regulations lead to compliance or defiance rather than self-determined behavior (Deci, Vallerand, Pelletier, & Ryan, 1991), leaving students feeling imprisoned by academic expectations. Moving away from testing and toward tasks more supportive of autonomy and competence has been shown to lead to stronger learning and memory outcomes (Grolnick & Ryan, 1987; Grolnick, Ryan, & Deci, 1991). Deci, Vallerand, Pelletier, & Ryan (1991), leaders in the area of SDT, noted that "the highest quality of conceptual learning seems to occur under the same motivational conditions that promote personal growth and adjustment" (p. 326). Just as a flower needs the nutriments of earth, sun, and water to bloom, students need environments rich in autonomy, belonging, and competence to nurture their motivation for learning.

Although perception of autonomy has been shown to strengthen intrinsic motivation for learning (Pekrun, 2006; Frenzel, Pekrun, & Goetz, 2007; Patall, Cooper & Robinson, 2008; Murayama, Pekrun, Lichtenfeld, & Hofe, 2013), support for autonomy is not prevalent in education. In an influential meta-analysis of 41 studies examining the effects of choice interventions, Patall, Cooper, & Robinson (2008) noted a small to moderate effect on average, d = .30, 95% CI = [0.25, 0.35] using a fixed-error model. Collected works offered evidence of positive effects on various outcomes including intrinsic motivation, effort, task performance, and perceived competence. Collected works also varied with regard to the benefits of choice, among 91 effect sizes from 41 studies, 78 represented positive effects (e.g. Becker, 1997; Cordova & Lepper, 1996; Iyengar & Lepper, 1999; Kernan, Heimann, & Hanges, 1991), while 13 represented negative effects (e.g., Schwartz, 2000, Baumeister, Bratslavsky, Muraven, & Tice, 1998). Observed effects were stronger for instructionally irrelevant choices, when multiple choices were provided in succession, when choices were not supplemented by rewards, when experimental groups were compared to strict controls, in samples of children, and in laboratory settings (Patall, Cooper & Robinson, 2008). The present work was designed to consider many of these issues within an online learning environment.

A similarly influential review of choice interventions by Katz & Assor (2007) also suggested that choice can be either motivating or demotivating, based on how well the context of that choice supports the basic psychological needs defined by SDT. The authors suggested that while teachers are widely open to the positive effects of choice (Flowerday & Schraw, 2004) choice alone does not define autonomy (Katz & Assor, 2007). For instance, empty choices or choices that lack structure are often ineffective in producing positive performance outcomes or improving intrinsic motivation. Choices can also vary in efficacy based on whether they entail actions or options (Reeve, Nix, & Ham, 2003), with action choices associated with greater perceived autonomy. Action choices, which delegate how time is spent, may function through stronger interest, which Flowerday et al. (2004) suggested was a primary mechanism underlying the effect of choice. In addition, choices that undermine competence or that overload students' cognitive capacities can have deleterious effects (Payne 1976; Sweller, 2004). In contrast, choices that impart meaning or that make a task more valuable on a personal level have been shown to promote learning (Cordova & Lepper, 1996). Further, the effect of choice can vary based on students' age and level of cognitive development (Katz & Assor, 2007; Piaget, 1971) and based on cultural considerations (Iyengar & Lepper 1999). Still, when administered appropriately, evidence suggests that choice offers a potentially powerful intervention by which teachers can instill opportunities for internalized regulation.

7.1.3 Technology in the Classroom

Online learning environments offer unique, cost effective opportunities for supporting students' autonomy in educational tasks. An increase in the popularity of one-to-one programs providing all students with laptops (Singer, May 13, 2017), coupled with an influx of readily accessible online learning environments (Escueta, Quan, Nickow, & Oreopoulos, 2017), has helped to ensure that the future of most learning tasks will involve technology.

Online learning environments tend to pair rich feedback with robust data collection and reporting, inspiring a new age of investigation into how people learn (Baker, 2016). Content and feedback often involve hypermedia that more primitive approaches to education have lacked. For instance, online learning environments allow for rich animations and narrations, interactive and collaborative experiences, easy access to supplementary resources, the support of immediate feedback or heavily scaffolded learning paths, and opportunities for both cognitive and socio-emotional intervention. However, bigger, brighter, and more connected does not always translate to better learning. To reduce the misuse of hypermedia elements, Mayer (2005) established a set of multimedia principles to optimize the design of technology supported learning environments and to address issues like cognitive load (Sweller, 2004) and working memory (Baddeley & Hitch, 1974). Motivational elements considered alongside these principles have also shown efficacy (Mayer, 2014). As such, the present work considers the influence of hypermedia by comparing how hint medium (video vs. text) in an online learning environment influences students' cognitive and motivational outcomes.

The present work is conducted within ASSISTments, an online learning environment for middle school mathematics that provides students with immediate feedback and teachers with formative assessment data to drive class instruction (Heffernan & Heffernan, 2014). ASSISTments is unique in that it also supports educational research at scale through its capacity for student-level randomized controlled trials and A/B testing (Ostrow, Heffernan, & Williams, 2017). The system is used by more than 50,000 students around the world and has a vast library of K-12 content mapped to the Common Core Stand Standards (NGACBP & CCSSO, 2010). A recent efficacy trial revealed that ASSISTments usage was associated with powerful learning gains, essentially doubling students' mathematics learning over the course of a single academic year (Roschelle, et al., 2016).

The present work builds on two areas of research in ASSISTments: assessing the effects of supporting student autonomy and assessing the effect of video-enriched hint content. Past work examining the effects of choice within ASSISTments suggested that students who were able to invest in their learning experience outperformed those who did not receive support for autonomy. Initial work suggested that students given a choice of hint medium (video vs. text) averaged higher accuracy on a mathematics assignment while using fewer hints and attempts (Ostrow & Heffernan, 2015). Further, the effect of choice remained significant even when hints were not ultimately utilized, leaving students' choices instructionally irrelevant. In that work, the effect of feedback medium was not significant, contrary to earlier work on hint medium as a mechanism for efficacy within ASSISTments (Ostrow & Heffernan, 2015; Ostrow & Heffernan, 2014). Follow-up work revealed that choice was potentially less effective than first anticipated but was limited by a small sample size and low power (Ostrow, 2015). That work instead isolated a main effect of hint medium, again associating video hints with stronger assignment scores, less hint usage overall, and higher efficiency in assignment completion, aligning with earlier work on the differential effects of hint medium (Ostrow 2015; Ostrow & Heffernan, 2014). Thus, incongruence in past work emphasizes the need for further exploration into how support for autonomy and variation in hint medium might influence performance and intrinsic motivation.

7.1.4 The Present Study

Building on past work, the theoretical framework for the present work is shown in Figure 7.2. Despite inconsistencies in past work, it is hypothesized that support for autonomy by providing students a choice in their hint medium will be associated with stronger cognitive performance immediately following a math assignment, as measured by a posttest comprised of near and far transfer items. It is also hypothesized that hint medium (video vs. text, chosen or randomly assigned) will influence learning outcomes, with stronger performance anticipated from video-enriched content. The present work also extends on past work to investigate how choice and hint-medium affect students' motivation by collecting self-reported perceptions of autonomy and interest/enjoyment using subscales of the Intrinsic Motivation Inventory. It is hypothesized that support for autonomy will be associated with heightened perceptions of autonomy and greater interest and enjoyment. Considering past work with ASSISTments, it is also hypothesized that effects of choice and hint medium will differ across levels of prior knowledge, with a stronger effect expected among low knowledge students. In addition, it is suspected that observed effects may vary based on hint usage during the assignment, as hints are provided on-demand and not all

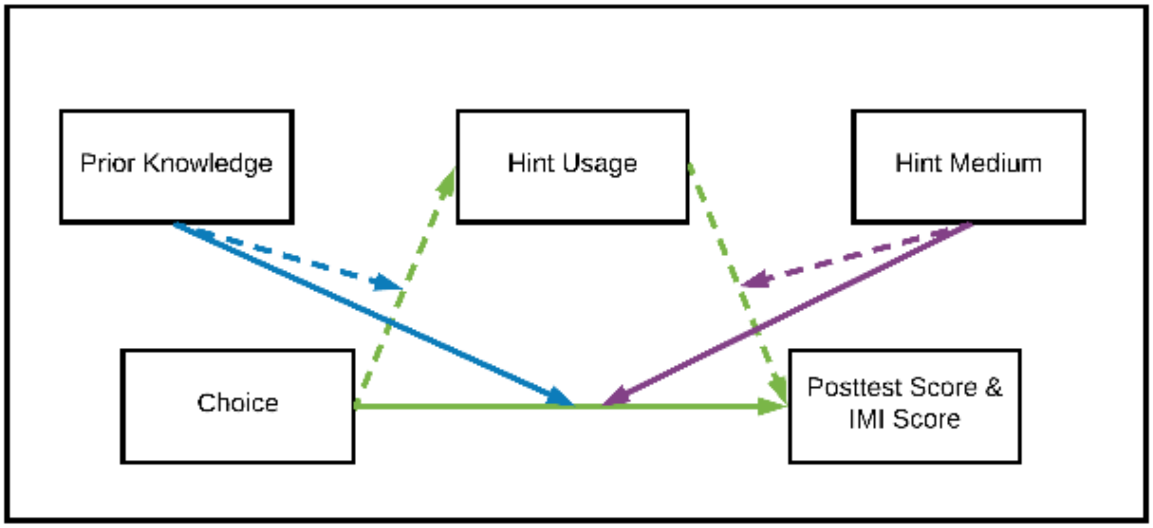


Figure 7.2. The theoretical framework guiding the present work. Choice is expected to influence posttest score and IMI subscale scores. This effect likely differs by prior knowledge levels and may differ by the hint medium requested or assigned. The effect may or may not change when considering the subsample of students accessing hints.

students request assistance while completing their assignments. As such, analyses are conducted on both the intent-to-treat and treated samples for comparison of observed effects. The following research questions guide the present work:

RQ 1. Does offering students a choice in their hint medium affect their cognitive performance as measured at posttest? In addition, does hint medium itself affect later performance, or is there an interaction effect between choice and medium? Finally, do observed effects differ by levels of prior knowledge?

RQ 2. Does offering students a choice in their hint medium affect self-reported perceptions of autonomy and intrinsic motivation? Is there a main effect of hint medium, or an interaction effect between choice and medium? And again, do observed effects differ by levels of prior knowledge?

RQ 3. To supplement RQ1 and RQ2 and to calibrate the present work with previous research in ASSISTments, does choice and/or hint medium affect within-assignment metrics of performance including accuracy, hint usage, attempt count, or overall response time? How do observed effects differ by levels of prior knowledge?

7.2 Methods

7.2.1 Learning Environment

This study was conducted within ASSISTments (www.ASSISTments.org), a popular online learning platform with over 50,000 student users around the world. The platform has over 25,000 problems mapped to the Common Core State Standards (NGACBP & CCSSO, 2010) that are delivered as certified material and offered as a free service of WPI. Teachers assign content to students as classwork or homework and benefit from robust reports on student performance. As students complete their assignments, they benefit from immediate feedback and tutoring. Past work has shown that ASSISTments reliably improves students' learning in comparison to traditional paper and pencil approaches (Mendicino, Razzaq, Heffernan, 2009; Koedinger, McLaughlin & Heffernan, 2010; Singh et al., 2011; Kelly, Heffernan, Heffernan, et al., 2013; Soffer, et al., 2014). A recent efficacy trial also revealed that students using the platform for a year experienced large, reliable learning gains on standardized assessments (Roschelle, et al., 2016), leading reviewers of online learning platforms to list ASSISTments as one of the most effective online learning interventions currently available to teachers and students (Escueta, Quan, Nickow, & Oreopoulos, 2017).

In addition, ASSISTments serves as a powerful tool for educational research at scale (Ostrow, Heffernan, & Williams, 2017). Randomized controlled trials can be embedded in Skill Builders to assess the efficacy of learning interventions. Skill Builders are certified mastery-based fluency assignments tailored to specific skills. Skill Builders are completed as classwork or homework, allowing researchers to collect data from authentic learning environments using minimally invasive approaches.

7.2.2 Participants

Participants were student users of ASSISTments, sourced through naturalistic assignment of a modified ASSISTments Certified Skill Builder. Data was collected between January and March of 2018. The assignment was assigned to 17 classes by 12 teachers from 9 schools in 3 states, establishing a full sample of 556 students. Three of these teachers were listed as teaching 6th grade, four were listed as teaching 7th grade, two were listed as teaching 8th grade, one was listed as teaching both 7th and 8th grades, and two did not list their class grade levels. Based on ASSISTments' name-based query, the sample was 40% female, 42% male, and 18% unidentified.

7.2.3 Procedures

7.2.3.1 Math Content

The present study was embedded within an ASSISTments Certified Skill Builder with mathematics content mapped to the Common Core State Standard (NGACBP & CCSSO, 2010) of "Finding the Percent of a Number," a sixth grade skill in Ratios and Proportional Relationships (6.RP.A.3c). This skill was chosen based on its popularity in the previous year during the data collection period of interest.

Problems within the assignment covered finding the percent of numbers when the percentage is less than or greater than 100% (e.g., "What is 41% of 90?" or "What is 120% of 80?"), offering two problem templates that differed slightly in hint structure. Hint content was modified for ten problems (five of each type), establishing text and video copies with perfectly matched feedback. In text feedback conditions, hints were comprised of screenshots of text formatted in Microsoft

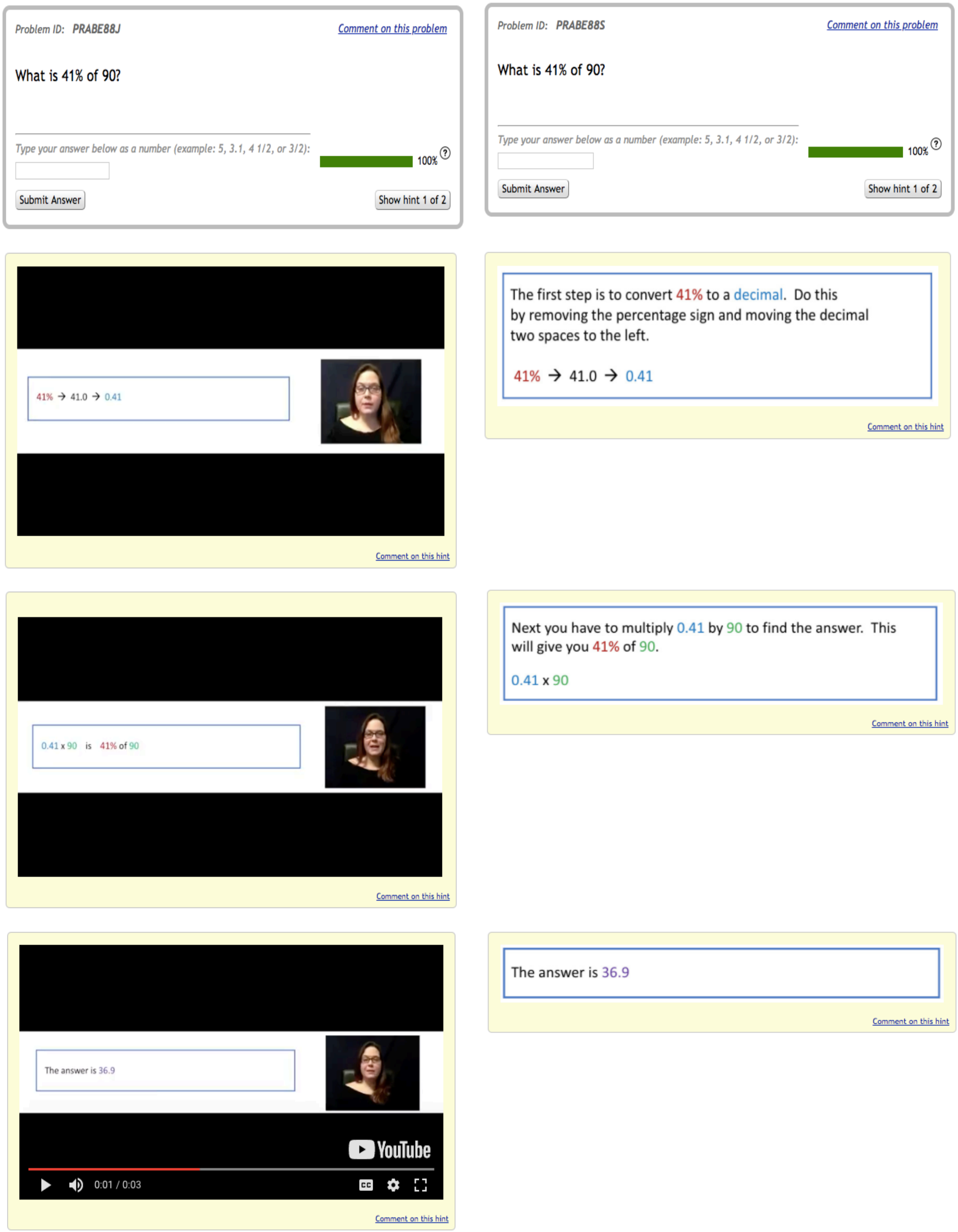
PowerPoint. In video feedback conditions, hints were comprised of YouTube videos of the primary author relaying the same exact hint content. A comparison of these mediums is presented in Figure 7.3. All videos used a structured PowerPoint presentation format, with the "talking head" of the primary author embedded next to the featured hint content. The primary author read the text of the hint, with any pertinent mathematical equations listed on screen. Cues and colors were kept identical in the transition across mediums, as shown in Figure 7.3. Additional examples of problem and hint content are available at (Ostrow, 2018) for further reference.

7.2.3.2 Experimental Design

The experimental design for the study underlying the present work is presented in Figure 7.4. Students were initially subjected to a "video check" problem to ensure that they had access to video content hosted on YouTube. Students reporting difficulty accessing video content were provided an alternative assignment (the traditional Skill Builder content) and did not participate in the posttest or IMI. Those who passed the video check by confirming their ability to see and hear video content were given an introduction to the assignment (see Figure 7.5, left) and were randomly assigned to receive either choice or no choice of their hint medium. The introduction was intended to provide additional structure to the assignment by outlining expectations and procedures for success, as autonomy is thought to be effective only in the presence of high structure (van Loon, Ros, & Martens, 2012). Those randomly assigned to the choice condition were asked to select their preference of video or text hints (see Figure 7.5, right) and were routed into the appropriate condition (Text Choice (TC) or Video Choice (VC)). Students that did not get to choose were randomly assigned to receive either video feedback (Video No Choice (VNC)) or text feedback (Text No Choice (TNC). Hint feedback for all problems was provided on demand and with increasing specificity, with the answer provided as the final (or {|"bottom out") hint.

All conditions were structured such that video or text versions of the 10 problems with modified hint content were delivered randomly. If students required more than 10 problems to reach mastery, they were routed into additional content pulled from the original Skill Builder. These "spill over" problems included the original, unaltered text hints for all conditions. This approach was used to limit the amount of video content that needed to be created for the study to 35 videos. A possible limitation of this approach is the potential washout effect in struggling students requesting or assigned to video feedback but needing more than 10 problems.

Following mastery of the Skill Builder (i.e., accurately answering three consecutive problems), all participating students were transitioned to a posttest containing three transfer problems of progressive difficulty. Posttest problems were delivered with correctness feedback but without additional hint feedback; if students got stuck they were able to request the answer and move on. Following the posttest, students could opt-in to provide responses to IMI subscales measuring their interest/enjoyment and perceived autonomy. Survey items were embedded into ASSISTments using Qualtrics and the ASSISTments Survey System (see Ostrow & Heffernan, C. 2015 for more information). This integration allows for the collection of more robust survey elements that can be



Problem ID: PRABE88J

What is 41% of 90?

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

41% → 41.0 → 0.41

0.41 x 90 is 41% of 90

The answer is 36.9

Problem ID: PRABE88S

What is 41% of 90?

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

The first step is to convert 41% to a decimal. Do this by removing the percentage sign and moving the decimal two spaces to the left.

41% → 41.0 → 0.41

Next you have to multiply 0.41 by 90 to find the answer. This will give you 41% of 90.

0.41 x 90

The answer is 36.9

Figure 7.3. Example of math content and matched video or text hints.

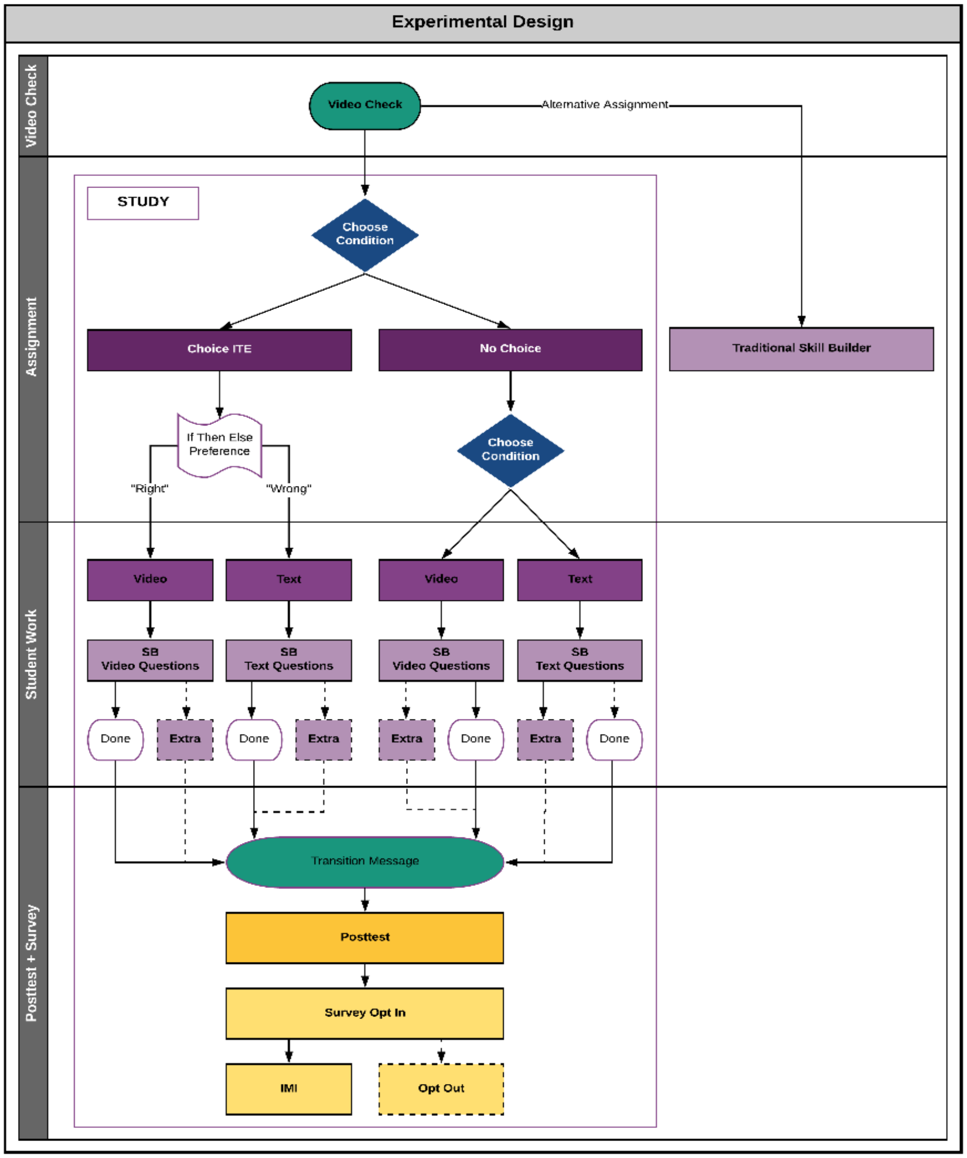
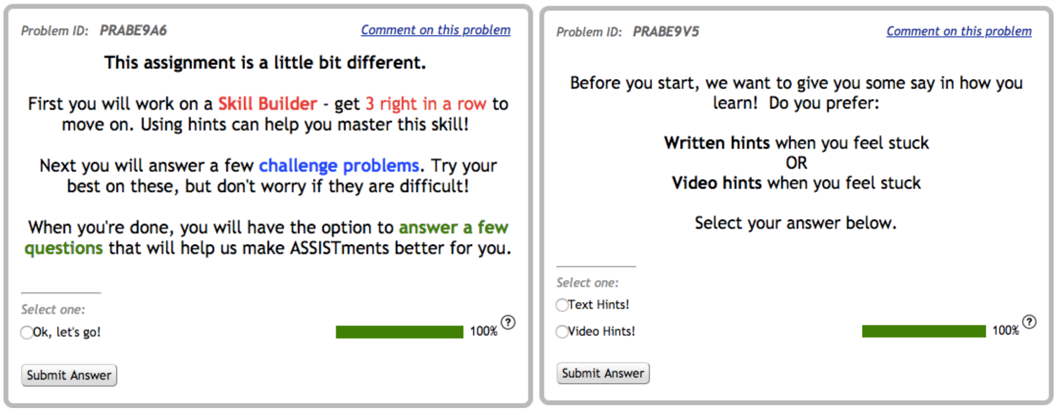


Figure 7.4. Experimental design.


Problem ID: PRABE9A6 Comment on this problem

This assignment is a little bit different.

First you will work on a Skill Builder - get 3 right in a row to move on. Using hints can help you master this skill!

Next you will answer a few challenge problems. Try your best on these, but don't worry if they are difficult!

When you're done, you will have the option to answer a few questions that will help us make ASSISTments better for you.

Select one:

○Ok, let's go!

Problem ID: PRABE9V5 Comment on this problem

Before you start, we want to give you some say in how you learn! Do you prefer:

Written hints when you feel stuck
OR
Video hints when you feel stuck

Select your answer below.

Select one:

○Text Hints!

○Video Hints!

Figure 7.5. Introduction providing structure (left) and preference routing question (right).

linked to students' math performance through unique student identifiers. Those who opted out of the IMI were simply prompted to submit their assignment.

7.2.4 Measures

7.2.5 Intrinsic Motivation Inventory

The IMI is a multidimensional scale intended to measure participants' subjective experiences of task participation (Intrinsic Motivation Inventory, 2018; Ryan, 1982). The IMI is well established, with 30 years of use lending evidence to subscale validity and reliability across tasks, conditions, and settings (Intrinsic Motivation Inventory, 2018). The scale has six primary subscales that can be used independently or in combination: interest/enjoyment, perceived competence, effort, value/usefulness, felt pressure and tension, and perceived choice. A seventh subscale intended to measure perceived relatedness or belonging was added in recent years and is not well-established as valid or reliable.

Students were asked to indicate how true each statement was for them using a Likert scale (1 = *Not at All True*, 7 = *Very True*). In the present work, all scale items were modified slightly to reflect an academic task or setting; such modifications are thought to be inconsequential to outcomes and interpretations (Intrinsic Motivation Inventory, 2018). Past work has suggested that order effects are also negligible, and that subscales and items can be included or excluded as necessary (Schell & Oswald, 2013; Intrinsic Motivation Inventory, 2018). The two subscales considered in the present work are described more thoroughly in the subsections below.

7.2.5.1 Interest/Enjoyment

This subscale is the primary measure of intrinsic motivation. It has previously been shown to be predicted by Autonomy and Competence reports in respective scales (Intrinsic Motivation Inventory, 2018), suggesting these scales are correlated. In its original form, this subscale includes seven items regarding intrinsic motivation (i.e., "I enjoyed doing this assignment very much"), with two items reverse scored (i.e., "This assignment did not hold my attention at all"). A reduced version of the subscale is used in the present work, comprised of only five items and achieving higher reliability, establishing a slightly more efficient measure (Ostrow & Heffernan, In Press).

7.2.5.2 Perceived Autonomy

This subscale is the primary measure of autonomy, also known as choice, volition, or task-based locus of control. It has previously been shown to predict Interest/Enjoyment (Intrinsic Motivation Inventory, 2018). In its original form, the subscale includes seven items regarding perceived autonomy (i.e., "I did this assignment because I wanted to"), with five items reverse scored (i.e., "I did this assignment because I had to"). A reduced version of the subscale is used in the present work, comprised of six items and achieving higher reliability, establishing a slightly more efficient measure (Ostrow & Heffernan, In Press).

7.2.6 Student Performance Measures

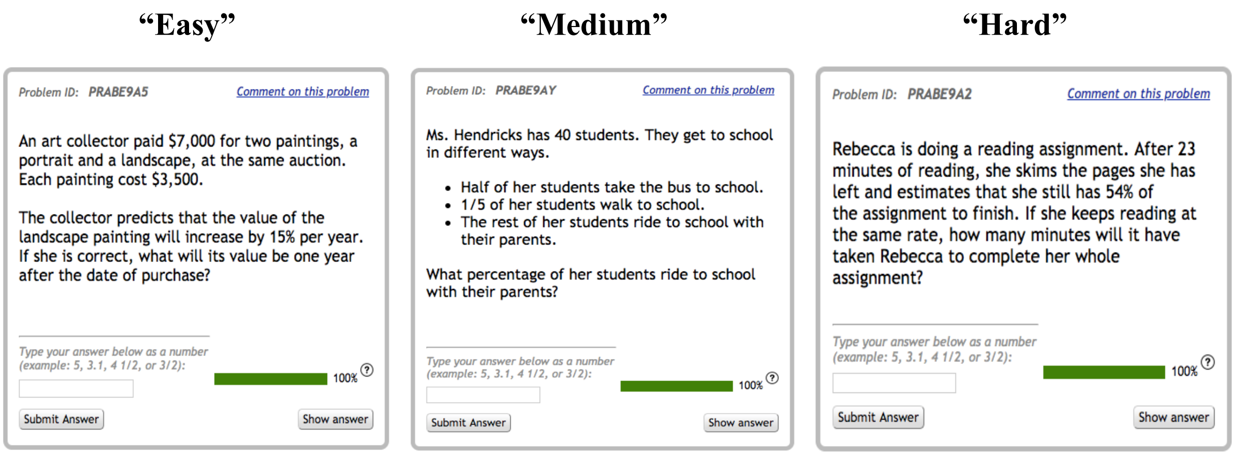
Student performance measures are processed from data logged by ASSISTments and represent student abilities before, during, and after their math assignment. Each metric considered in the present work is described in detail below.

7.2.6.1 Prior Knowledge

All problem accuracy data from a student's work within ASSISTments prior to their participation in the present study is averaged into a single value meant to serve as a metric of their general math ability. For instance, a student who has accurately solved 75 out of 100 attempted problems within various ASSISTments assignments would have a prior knowledge score of 0.75, or 75%. This metric does not control for the amount of work the student has completed (i.e., another student accurately solving 3 out of 4 attempted problems would share the score of 0.75) or for the difficulty level of problems solved. However, previous work supports that when the value is dichotomized on its median within a dataset, resulting groups of "low prior knowledge" and "high prior knowledge" students exhibit significantly different performance within and immediately following experimental assignments (Razzaq & Heffernan, 2009; Ostrow, Heffernan, Heffernan, & Peterson; 2015). High knowledge students generally outperform low knowledge students, work more efficiently, and use fewer instances of hint feedback. In addition, past work has shown experimental effects can be more impactful for low knowledge students, potentially speaking to the achievement gap (Razzaq & Heffernan, 2009; Ostrow, Heffernan, Heffernan, & Peterson; 2015). As such, this categorical measure of prior knowledge is used in the present work with the expectation that choice and feedback medium may have more pronounced effects in low knowledge students.

7.2.6.2 Posttest Score

Following mastery of their Skill Builder content, all study participants were given a transition message telling them to try their best on a few challenge problems. These challenge problems, a set of three transfer problems considered to be of "Easy," "Medium," and "Hard" difficulty, formed the posttest. Problems were presented using the same order of progressive difficulty for all students. Problems were sourced from 7th and 8th grade ASSISTments content mapped to the Ratios and Proportional Relationships standard, with the intention of serving as an extension of the 6th grade skill. These problems required students to apply the skill of finding the percentage of a number in more complex formulas or alongside distractors, as shown in Figure 7.6. Posttest problems did not feature hint feedback; students could make multiple attempts and if they got stuck they were able to request the answer and move on. ASSISTments logged the accuracy of each student's first response on each problem and values were averaged to arrive at their overall posttest score (i.e, .80 or 80%). Posttest score was assessed in the full analytic sample as well as in the



“Easy”

Problem ID: PRABE9A5 Comment on this problem

An art collector paid $7,000 for two paintings, a portrait and a landscape, at the same auction. Each painting cost $3,500.

The collector predicts that the value of the landscape painting will increase by 15% per year. If she is correct, what will its value be one year after the date of purchase?

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

“Medium”

Problem ID: PRABE9AY Comment on this problem

Ms. Hendricks has 40 students. They get to school in different ways.

- Half of her students take the bus to school.
- 1/5 of her students walk to school.
- The rest of her students ride to school with their parents.

What percentage of her students ride to school with their parents?

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

“Hard”

Problem ID: PRABE9A2 Comment on this problem

Rebecca is doing a reading assignment. After 23 minutes of reading, she skims the pages she has left and estimates that she still has 54% of the assignment to finish. If she keeps reading at the same rate, how many minutes will it have taken Rebecca to complete her whole assignment?

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

Figure 7.6. Posttest problems and delivery of progressive difficulty.

subsample of students who requested feedback during their assignment (fidelity of feedback medium treatment).

7.2.6.3 Metrics within Assignment

ASSISTments log files also provided information on students’ performance within the Skill Builder portion of their assignment. In theory, choice or feedback medium may have altered how students interacted with their assignments (i.e., students with choice may have shown higher accuracy or performed slower, with greater diligence). Four metrics were considered as dependent variables in a set of secondary analyses: assignment accuracy, hint usage, attempt count, and overall response time.

“Assignment Score” reflects the average performance across problems in the Skill Builder portion of the assignment. This variable is the sum of correct responses divided by the total number of problems seen. Based on teacher preferences, accuracy may be binary (1 = *Correct*, 0 = *Incorrect*) or use a partial credit algorithm based on students’ attempt and hint counts. Although the assignment is mastery based, assignment score provides insight into the quality of students’ overall performance. Higher assignment scores suggest stronger or more persistent performance, while lower assignment scores suggest either a lack of understanding or persistence, or gaming behavior.

“Assignment Hints” reflects the total number of hints students requested while working through their math problems. Each problem in the assignment included two to three hints, delivered with progressive specificity, with the final hint providing the correct answer. Higher hint counts are not necessarily bad, as moderate hinting is usually indicative of learning. However, very low hint counts are likely due to ceiling effects, and very high hint counts may indicate amotivation, inattention, or attempts to ‘game’ the assignment.

“Assignment Attempts” reflects the total number of solution attempts students made while working through their math problems. Although students are marked incorrect if they are unable to accurately solve a problem on their first attempt, multiple attempts can then be used to arrive at the correct answer, allowing students to move on to the next problem. Like hints, moderate attempt counts may be indicative of learning, while lower counts are likely due to ceiling effects and higher counts may indicate gaming.

"Assignment Time" reflects the sum of students' first response times while solving problems within their assignment. The distributions of time-based measures in ASSISTments usually have strong positive skew because students can walk away from their assignments without closing the tutor, thereby inflating response times dramatically. As such, Assignment Time is the natural log of the trimmed total time (in minutes) that students take to respond to all problems. Moderate assignment times suggest student struggle or persistence, while higher times suggest inactivity and lower times suggest either strong incoming knowledge or gaming behavior.

7.3 Results

Of the full sample of 556 students, 15 opened and immediately closed the assignment before making any progress. Additionally, 190 did not pass the video check and participated in an alternative assignment. Two students were assigned to choose their feedback medium and dropped out before making a selection, therefore reducing the analytic sample to 349 students. Based on ASSISTments' name-based query, the analytic sample was 43% female, 40% male, and 17% unidentified. The sample was also representative of 12 teachers at nine schools in three states. Teachers were listed as teaching 6th grade (three), 7th grade (four), or 8th grade (three), with two teachers not listing their class grade levels. Random assignment left the groups somewhat unbalanced at the highest level, with 185 students assigned to a choice condition and 164 assigned to a condition without choice. Among those not given a choice, 76 were then randomly assigned to receive text hints, while 88 were randomly assigned to receive video hints. Groups were especially unbalanced within the subsample of students given a preference, with 122 choosing text hints and 63 choosing video hints.

Attrition and treatment fidelity were assessed prior to undertaking more robust analyses. Attrition rates were not significantly different across experimental groups, χ^2 (3) = 1.22, p = .748, with completion rates of 75-81%, as shown in Table 7.1. The treated sample for hint medium (those requesting hint feedback) was small, with only 16-22% of students using hints across groups. An a priori power analysis revealed that in order to find a moderate to small effect (f = .25 - .1), 128-787 students would be required to observe a true effect, as shown in Figure 7.7. A moderate proportion of students opted-in to the IMI following their posttests, with 43-58% of students providing responses across groups. Of the analytic sample of 349 students, 301 completed their posttests and 174 completed their IMI subscales. Within assignment measures also carry higher sample sizes due to inclusion of the "Alternative" group that received a traditional, business as usual Skill Builder assignment and were excluded from posttest and IMI activities. This group serves as a "business as usual" control for within assignment analyses. Descriptive statistics for the analytic sample are provided in Table 7.2. Correlation analyses revealed that variables

Table 7.1. Attrition analysis and percentage of hint usage and survey fidelity by condition.

	Incomplete	**Complete**	**Total**	**% Completion**	**Hints***	**Survey***
Choice Text	23	99	**122**	81	17%	43%
No Choice Text	19	57	**76**	75	17%	46%
Choice Video	12	51	**63**	81	22%	54%
No Choice Video	18	70	**88**	80	16%	58%
Total	**72**	**277**	**349**	79		

Note. *Hints and Survey participation samples are proportions based on students that completed.

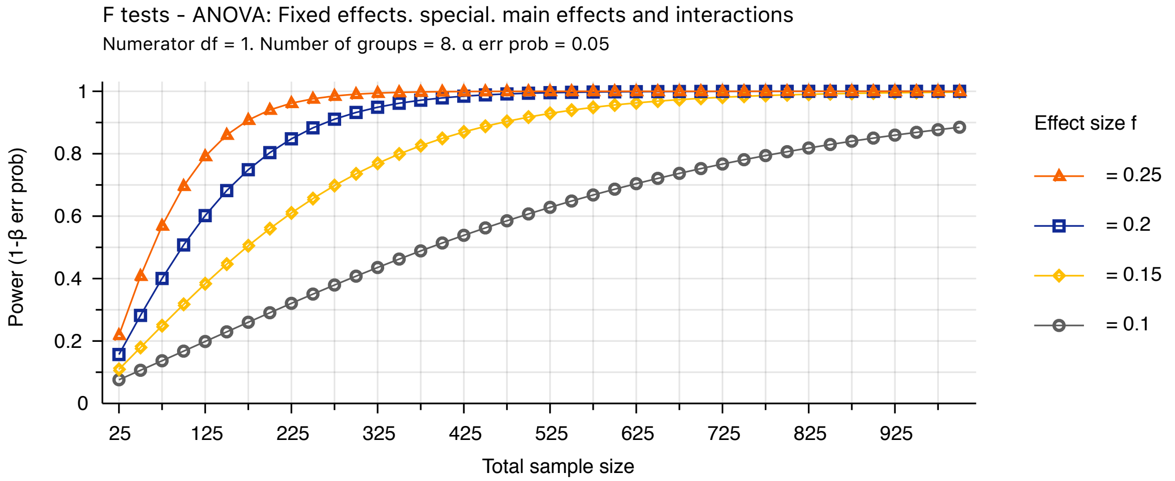


Figure 7.7. Observable power by sample size for a range of effect sizes.

Table 7.2. Descriptive statistics of the analytic sample.

	Analytic Sample				
	N	Mean	SD	Min	Max
Independent Variables					
Video	349	.43	.50	0.00	1.00
Choice	349	.53	.50	0.00	1.00
Knowledge	539	.58	.49	0.00	1.00
Within Assignment Dependent Variables					
Accuracy	539	.86	.26	0.00	1.00
Average Hint Usage	539	.51	1.26	0.00	7.00
Average Attempt Count	539	1.16	.67	0.00	11.00
Average First Response Time (ln, in min.)	515	.95	1.26	-2.59	4.09
After Assignment Dependent Variables					
Posttest Score	301	.43	.34	0.00	1.00
Interest/Enjoyment	174	3.14	1.48	1.00	7.00
Perceived Autonomy	174	3.48	1.55	1.00	7.00
IMI Score	173	3.31	1.11	1.00	5.73
Valid N (listwise)	173				

Table 7.3. Correlation analysis of modeled variables.

	1	2	3	4	5	6	7	8	9	10
Independent Variables										
1. Video										
2. Choice	**-.20**									
3. Knowledge		**.22**								
Within Assignment Dependent Variables										
4. Assignment Score			.11							
5. Assignment Hints			**-.18**	**-.42**						
6. Assignment Attempts			**-.19**	-.13	**.61**					
7. Assignment ln(FRT)	.11		**-.25**		.11	**.26**				
After Assignment Dependent Variables										
8. Posttest Score			**.17**	**.36**	**-.35**	**-.25**				
9. Interest/Enjoyment			**-.25**					**.21**		
10. Perceived Autonomy										

*Note***.** Bold correlations, p < .01. Others, p < .05. Suppressed correlations were not significant, p > .05.

Table 7.4. Item analysis of IMI subscales.

	Analytic Sample				
IMI Question	N	Mean	SD	Min	Max
Interest/Enjoyment					
I enjoyed doing this assignment very much.	176	3.47	1.62	1.00	7.00
This assignment was fun to do.	175	3.32	1.66	1.00	7.00
I would describe this assignment as very interesting.	175	3.33	1.69	1.00	7.00
I thought this assignment was quite enjoyable.	174	3.19	1.67	1.00	7.00
While I was doing this assignment, I was thinking about how much I enjoyed it.	175	2.39	1.69	1.00	7.00
Perceived Autonomy					
I believe I had some choice about doing this assignment.	175	3.25	1.90	1.00	7.00
I felt like it was not my own choice to do this assignment. (Reversed)	174	3.96	2.03	1.00	7.00
I didn’t really have a choice about doing this assignment. (Reversed)	174	3.63	2.13	1.00	7.00
I felt like I had to do this assignment. (Reversed)	174	2.85	1.80	1.00	7.00
I did this assignment because I had no choice. (Reversed)	174	3.72	2.12	1.00	7.00
I did this assignment because I had to. (Reversed)	174	3.44	2.13	1.00	7.00
Valid N (listwise)	173				

modeled together were not highly correlated, reducing concern about multicollinearity, as shown in Table 7.3.

7.3.1 IMI Subscale Reliability

The IMI subscales were calculated using the reduced scale format established in (Ostrow & Heffernan, In Press). Students took the Interest/Enjoyment subscale first, and then the Perceived Autonomy subscale. Reliability matched that of past work, with both the five item Interest/Enjoyment scale and the six item Perceived/Autonomy scale reaching high reliability, $\alpha = .93$ and .86, respectively. This aligns well with previously reported values of $\alpha = .92$ and .83, respectively (Ostrow & Heffernan, In Press). Overall mean scores were 3.14 and 3.48,

Table 7.5. Correlations of IMI subscale items.

	IE1	IE2	IE3	IE4	IE5	PA1	PA2	PA3	PA4	PA5
Interest/Enjoyment										
1. I enjoyed doing this assignment very much.										
2. This assignment was fun to do.	**.85**									
3. I would describe this assignment as very interesting.	**.69**	**.70**								
4. I thought this assignment was quite enjoyable.	**.79**	**.85**	**.78**							
5. While I was doing this assignment, I was thinking about how much I enjoyed it.	**.62**	**.65**	**.69**	**.73**						
Perceived Autonomy										
1. I believe I had some choice about doing this assignment.	.15				**.21**					
2. I felt like it was not my own choice to do this assignment. (Reversed)						**.32**				
3. I didn't really have a choice about doing this assignment. (Reversed)						**.40**	**.61**			
4. I felt like I had to do this assignment. (Reversed)						**.39**	**.49**	**.54**		
5. I did this assignment because I had no choice. (Reversed)					.15	**.38**	**.53**	**.66**	**.55**	
6. I did this assignment because I had to. (Reversed)			.18	.16	.17	**.34**	**.52**	**.58**	**.56**	**.68**

*Note***.** Bold correlations, p < .01. Others, p < .05. Suppressed correlations were not significant, p > .05.

respectively. Item level analysis is presented in Table 7.4. Correlation analyses (shown in Table 7.5) revealed a pattern similar to that observed in past work on scale validation, showing that subscale items cluster well, while low or non-significant correlations tend to arise across subscales. Scale items were reverse scored as necessary and averaged to form Interest/Enjoyment and Perceived Autonomy scores, with overall mean scores of 3.14 and 3.48, respectively.

7.3.2 RQ 1 – Effects of Choice and Medium on Posttest Score

Past work has suggested that instilling autonomy promotes higher posttest scores (Ostrow & Heffernan, 2015), and that feedback medium may or may not be influential in learning outcomes (Ostrow & Heffernan, 2014; Ostrow & Heffernan, 2015; Ostrow, 2015). In addition, differential effects have been observed in past work for high and low knowledge students (Razzaq & Heffernan, 2009; Ostrow, Heffernan, Heffernan, & Peterson; 2015). As such, the present work used a three-way independent factors ANOVA to examine the effects of choice, feedback medium, and knowledge level on posttest score. Within the full, intent-to-treat sample, Levene's test was not significant, suggesting that variance was equal across groups, F(7,293) = 1.10, p = .366. These students were treated for choice, but not necessarily treated for feedback medium, as not all students actually requested hints while working. Within this sample, the main effect of feedback medium was not significant, F(1, 293) = .004, p = .950, nor was the main effect of choice, F(1, 293) = 0.39, p = .531. However, the main effect of prior knowledge level was statistically significant, F(1, 293) = 5.58, p = .019, partial η^2 = .02. High knowledge students (M = .49, SD = .35) outperformed low knowledge students (M = .38, SD = .33) as expected considering the operationalization of this variable and findings from past work.

The interaction effect of medium and knowledge level was statistically significant, F (1, 293) = 5.60, p = .019, partial η^2 = .02. Simple effects tests revealed that low knowledge students requesting or assigned to video feedback (M = .43 SD = .33) outperformed those requesting or assigned to text feedback (M = .33, SD = .32), with results nearing significance, F(1, 293) = 3.06, p = .081. Medium was not significant amongst high knowledge students, p > .05. A statistically significant difference was also observed in high knowledge students requesting or assigned to text

(M = .52, SD = .34) and low knowledge students requesting or assigned to text (M = .33, SD = .32), F(1, 293) = 12.74, p = .000, partial $\eta^2 = .02$. This interaction is shown in Figure 7.8. No other interaction effects were significant, as shown in Table 7.6, but the full interaction is visualized in Figure 7.9 for reference.

A secondary analysis was conducted to assess differences in posttest score among students that requested hint feedback during their assignment. These students were treated for choice as well as for hint medium. This sample was small, with only 62 students requesting hints during their assignments. As such, knowledge level was not considered in order to improve observable power. A two-way independent factors ANOVA was conducted to examine the effects of feedback medium and the provision of choice within the treated sample. Levene's test was not significant, suggesting equal variance across groups, F(3, 58) = 1.10, p = .366. Neither the interaction, nor the main effects were significant, as shown in Table 7.6. This may be due in part to exceptionally low observed power. Power analysis revealed that if the true expected effect is small (*f* = .1), a total sample size of 787 would be required to observe it, $\beta = .20$, $\alpha = .05$. If the true effect were anticipated to be more moderate (*f* = .25), the total required sample size would be 128, still larger than that available in the treated sample by 66 students. Figure 7.10 displays the interaction of choice and feedback medium within the full and treated samples. Interestingly, scores are virtually the same for the full sample across groups, with video feedback linked to slightly stronger performance, and choice linked to slightly stronger performance. However, the treated sample exhibited lower posttest scores overall, and those with choice performed better after choosing and

Table 7.6. Three-way independent factors ANOVA results of hint medium, choice, and prior knowledge level on posttest score in the intent to treat and treated samples.

	SS	df	MS	F	p	1-β
Intent-to-Treat Sample						
Medium	0.00	1	0.00	.004	.950	.05
Choice	0.04	1	0.04	0.39	.531	.10
Knowledge	0.63	1	0.63	5.58	.019	.65
Medium X Choice	0.02	1	0.02	0.19	.664	.07
Medium X Know	0.63	1	0.63	5.60	.019	.66
Choice X Know	0.04	1	0.04	0.32	.571	.09
Medium X Choice X Know	0.07	1	0.07	0.61	.434	.12
Error	32.99	293	.11			
Treated Sample						
Medium	0.05	1	0.05	0.64	.425	.13
Choice	0.01	1	0.01	0.04	.851	.05
Medium X Choice	0.04	1	0.04	0.46	.501	.11
Error	4.66	58	.08			

Table 7.7. Means (SDs) of posttest score by hint medium, choice, and prior knowledge level.

	Choice	No Choice	Total
Intent-to-Treat Sample	(n = 166)	(n = 135)	(n = 301)
Low Knowledge *(n = 150)*	.37 (.34)	.39 (.31)	.37 (.33)
Text *(n = 81)*	.34 (.32)	.32 (.32)	.33 (.32)
Video *(n = 69)*	.42 (.38)	.44 (.31)	.43 (.33)
High Knowledge *(n = 151)*	.51 (.35)	.45 (.34)	.49 (.35)
Text *(n = 92)*	.52 (.34)	.53 (.35)	.52 (.34)
Video *(n = 59)*	.48 (.38)	.38 (.31)	.44 (.35)
Total (n = 301)	.45 (.35)	.41 (.32)	.43 (.34)
Treated Sample	(n = 35)	(n = 27)	(n = 62)
Text (n = 34)	.17 (.24)	.21 (.31)	.18 (.27)
Video (n = 28)	.28 (.33)	.21 (.26)	.25 (.29)
Total (n = 62)	.21 (.28)	.21 (.28)	.21 (.28)

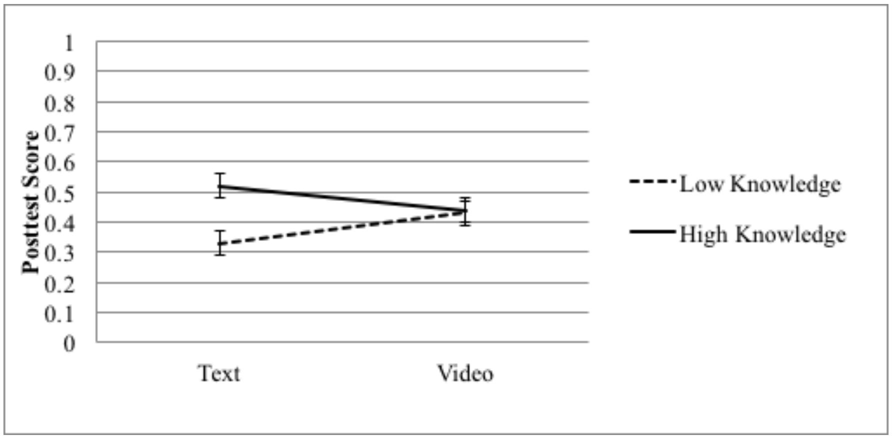


Figure 7.8. Posttest score by hint medium and prior knowledge level.

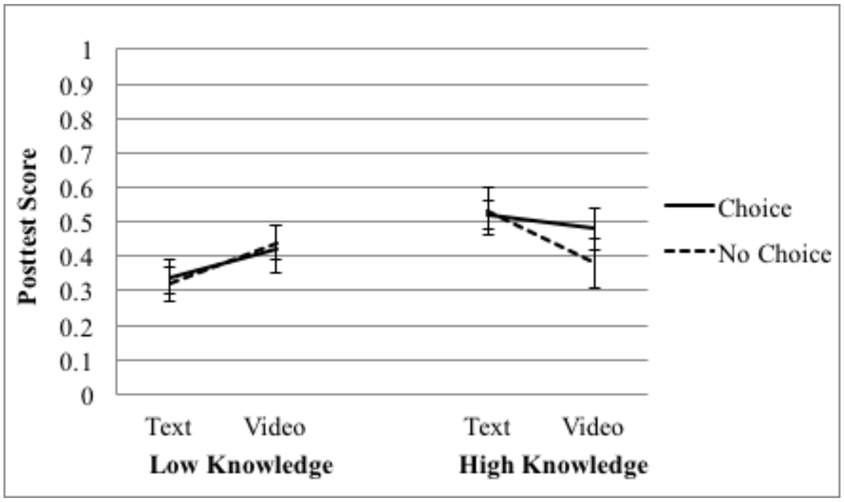


Figure 7.9. Posttest score by hint medium, prior knowledge level, and choice condition.

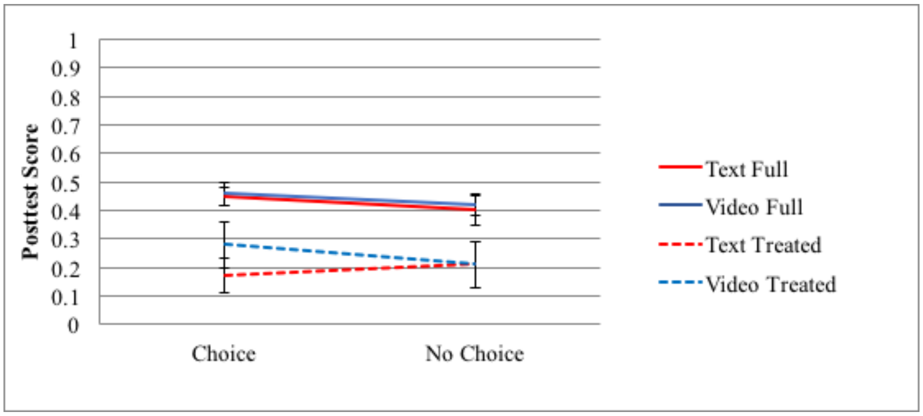


Figure 7.10. Posttest score by choice condition and hint medium in the intent to treat (full) and treated samples. Treated samples utilized their chosen or assigned hint feedback.

utilizing video feedback. This difference is not observable in the no choice condition, with students performing at approximately the same level. These findings were not significantly different in the treated sample, potentially due to low power and small sample size. Means and standard deviations for both samples are provided in Table 7.7.

7.3.3 RQ 2 – Effects of Choice and Medium on Intrinsic Motivation Subscales

Considering how past work on the effects of choice within ASSISTments (Ostrow & Heffernan, 2015; Ostrow, 2015) parallel the framework of Self-Determination Theory, it was also hypothesized that the provision of choice would alter students' responses on IMI subscales following their assignment and posttest. Presumably, those allowed choice of their hint medium would have higher scores on the Perceived Autonomy subscale, and either choice or hint medium may influence their self-reported Interest/Enjoyment. It is also possible that prior knowledge level may alter how students respond about their experience within a math assignment. As such, two three-way independent factors ANOVAs were used to examine the effects of choice, feedback medium, and knowledge level on IMI subscales for Perceived Autonomy and Interest/Enjoyment.

7.3.3.1 Perceived Autonomy

Within the portion of the intent-to-treat sample who opted-in to the IMI, Levene's test was not significant, suggesting that variance was equal across groups, F(7, 166) = 1.80, p = .091. Interestingly, none of the interaction effects or main effects were significant, as shown in Table 7.8. Means and standard deviations are included in Table 7.9 for reference. Extremely low power was achieved across effects, with only 174 students completing the IMI within the intent-to-treat sample. As noted for the posttest analysis following the same structure, 128-787 students would be required to observe moderate to small true effects when $\beta = .20$ and $\alpha = .05$. If true effects do exist they are likely small, given that the choice provided is minor and the wording of the IMI subscale items more broadly address students' sense of choice in completing their assignment. A larger sample size would help distinguish a small effect from truly insignificant findings.

7.3.3.2 Perceived Interest/Enjoyment

Within the portion of the intent-to-treat sample who opted-in to the IMI, Levene's test was significant, suggesting that variance was not equal across groups, F(7, 166) = 2.09, p = .047. In smaller samples this would be worrisome, but as Field suggests, the test is losing popularity in larger samples as its sensitivity is artificially influenced by sample size (2013).

Within this sample, the main effects of feedback medium and choice were not significant, as shown in Table 7.8. However, the main effect of knowledge was statistically significant, F(1, 166) = 9.16, p = .003, partial $\eta^2 = .05$. Interestingly, high knowledge students (M = 2.77, SD = 1.25) had significantly lower Interest/Enjoyment scores than their low knowledge peers (M = 3.50, SD = 1.60).

The interaction effect of medium and prior knowledge level was statistically significant, F (1, 166) = 5.55, p = .020, partial $\eta^2 = .03$. Simple effects tests revealed that low knowledge students requesting or assigned to video hints (M = 3.84, SD = 1.51) significantly outperformed those requesting or assigned to text hints (M = 3.12, SD = 1.62), F(1, 166) = 4.94, p = .028. Medium was not significant amongst high knowledge students, p > .05. Low knowledge students reported significantly higher Interest/Enjoyment after requesting or being assigned to video (M = 3.84, SD = 1.51) than their high knowledge peers in the same group (M = 2.65, SD = 1.44), F(1, 166) = 15.04, p = .000, partial $\eta^2 = .08$. This interaction is depicted visually in Figure 7.11. These findings suggest that video feedback may significantly improve the Interest/Enjoyment of struggling students, which aligns with past work within ASSISTments (Ostrow & Heffernan, 2014; Ostrow, 2015). No other interaction effects were significant, as shown in Table 7.8. Means and standard deviations are included in Table 7.9.

Astute readers will note that these findings mirrored those observed for posttest score, despite a low significant correlation between posttest and Perceived Interest/Enjoyment scores. When assessed using a linear regression to predict Interest/Enjoyment from posttest score, the model was significant, F(1, 172) = 8.11, p = .005, and explained 4.5% of the variance in Interest/Enjoyment scores, $R^2 = .045$, Adjusted $R^2 = .039$. Students' Interest/Enjoyment scores increased by .97 for every 1 SD increase in posttest score.

7.3.4 RQ 3– Effects of Experimental Condition on Within-Assignment Performance

Past work on the effects of choice within ASSISTments did not examine posttest. Instead, the focus was within-assignment performance (Ostrow & Heffernan, 2015; Ostrow, 2015). In addition, non-significant trends in past work suggested that students used fewer hints and attempts but spent longer working on each problem, suggesting optimized learning. The goal of research question three was to examine within-assignment performance.

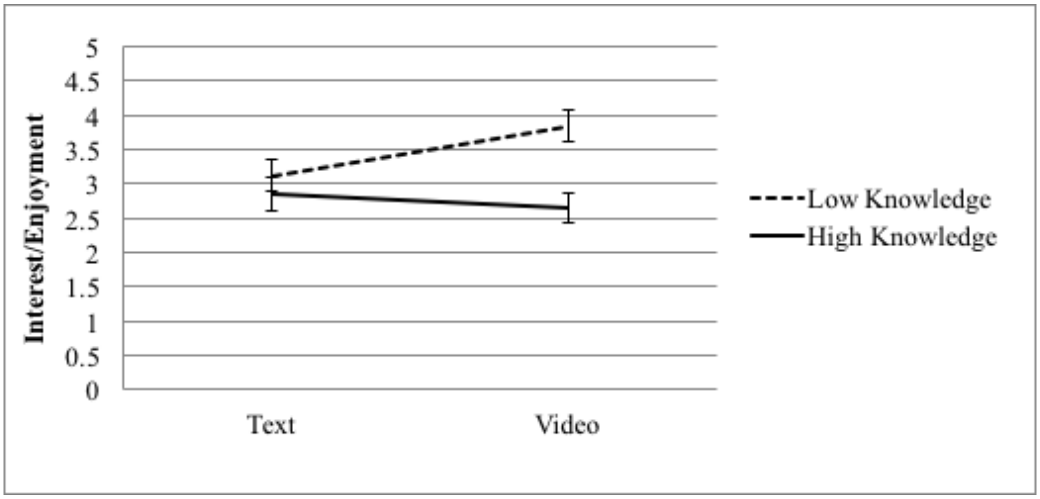


Figure 7.11. Interest/Enjoyment scale score by hint medium and knowledge level.

Table 7.8. Three-way independent factors ANOVA results of hint medium, choice, and prior knowledge level on Perceived Autonomy and Interest/Enjoyment IMI subscale scores.

	SS	df	MS	F	p	1-β
Perceived Autonomy						
Video	1.04	1	1.04	0.43	.515	.10
Choice	0.03	1	0.03	0.01	.919	.05
Knowledge	2.31	1	2.31	0.94	.334	.17
Video X Choice	0.04	1	0.04	0.02	.895	.05
Video X Know	0.86	1	0.86	0.35	.555	.09
Choice X Know	0.54	1	0.54	0.22	.640	.08
Video X Choice X Know	2.03	1	2.03	0.83	.365	.15
Error	407.88	166	2.46			
Interest/Enjoyment						
Video	0.99	1	0.99	0.49	.486	.11
Choice	1.12	1	1.12	0.56	.457	.12
Knowledge	18.50	1	18.50	9.16	.003	.85
Video X Choice	1.07	1	1.07	0.53	.468	.11
Video X Know	11.22	1	11.22	5.55	.020	.65
Choice X Know	2.04	1	2.04	1.01	.317	.17
Video X Choice X Know	3.83	1	3.83	1.90	.170	.28
Error	335.23	166	2.02			

Table 7.9. Means (SDs) of Perceived Autonomy and Interest/Enjoyment IMI subscale scores by hint medium, choice, and prior knowledge level.

	Choice (n = 88)	No Choice (n = 86)	Total (n = 174)
Perceived Autonomy			
Low Knowledge *(n = 88)*	3.26 (1.42)	3.45 (1.43)	3.39 (1.42)
Text *(n = 41)*	3.20 (1.45)	3.15 (1.53)	3.17 (1.48)
Video *(n = 47)*	3.32 (1.44)	3.67 (1.34)	3.57 (1.36)
High Knowledge *(n = 86)*	3.59 (1.58)	3.51 (1.90)	3.57 (1.67)
Text *(n = 48)*	3.49 (1.70)	3.67 (1.60)	3.52 (1.67)
Video *(n = 38)*	3.78 (1.34)	3.41 (2.10)	3.61 (1.70)
Total (n = 174)	3.48 (1.53)	3.47 (1.58)	3.48 (1.55)
Interest/Enjoyment			
Low Knowledge *(n = 88)*	3.72 (1.92)	3.40 (1.42)	3.50 (1.60)
Text *(n = 41)*	3.46 (1.98)	2.90 (1.34)	3.12 (1.62)
Video *(n = 47)*	4.08 (1.88)	3.77 (1.38)	3.84 (1.51)
High Knowledge *(n = 86)*	2.78 (1.26)	2.74 (1.27)	2.77 (1.25)
Text *(n = 48)*	2.74 (.99)	3.30 (1.38)	2.86 (1.09)
Video *(n = 38)*	2.85 (1.67)	2.41 (1.11)	2.65 (1.44)
Total (n = 174)	3.09 (1.56)	3.19 (1.40)	3.14 (1.48)

Four two-way independent factors ANOVAs were used to examine the effects of experimental condition and knowledge level on within-assignment scores, hint usage, attempt count, and response time. Within-assignment metrics have the added benefit of a strong, business-as-usual control in the Alternative group, students that failed the video check and did not participate in the posttest or IMI survey but completed a traditional Skill Builder. As such, these analyses compare performance metrics across five groups: Alternative, Text Choice (TC), Text No Choice (TNC), Video Choice (VC), and Video No Choice (VNC). All results are summarized in Table 7.10, with means and standard deviations presented in Table 7.11.

7.3.4.1 Assignment Score

Levene's test was significant, suggesting unequal variance across groups, F(9, 529) = 5.82, p = .000, but the sample was rather large and this was not necessarily of concern (Field, 2013). When considering students' average scores within their assignments (problems correct/total problems completed), the interaction effect of condition and knowledge level was significant, F(4, 529) = 2.94, p = .020, partial η^2 = .02. Simple effects tests revealed that this effect was driven in part by significant differences across groups amongst high knowledge students, F(4, 529) = 2.44, p = .046, partial η^2 = .02. Pairwise comparisons revealed significant differences (p < .05) between VNC (M = .78, SD = .35) and Alternative (M = .91, SD = .24), TC (M = .93, SD = .14), and TNC (M = .93, SD = .19). Differences across groups amongst low knowledge students were not significant, p > .05. Significant differences were also observed between high and low knowledge students in the Alternative group, F(1, 529) = 7.74, p = .006, partial η^2 = .01, and in the TC group, F(1, 529) =

8.02, p = .005, partial η^2 = .02, with trending differences in the TNC group, F(1, 529) = 3.40, p = .066, partial η^2 = .01. This interaction is displayed visually in Figure 7.12.

Main effects are less critical in the context of a significant interaction, but the main effect of condition was not significant, F(4, 529) = .48, p = .748. The main effect of knowledge level was significant, F(1, 529) = 5.18, p = .023, partial η^2 = .01, with high knowledge students (M = .89, SD = .24) outperforming low knowledge students (M = .82, SD = .29), as expected.

7.3.4.2 Assignment Hints

Levene's test was again significant when examining the total number of hints students used while completing their assignments, F(9, 529) = 11.35, p = .000. For this dependent variable, the interaction effect of condition and knowledge level was not significant, F(4, 529) = .91, p = .455, nor was the main effect of condition, F(4, 529) = 1.97, p = .098. The main effect of knowledge level remained significant, F(1, 529) = 14.01, p = .000, partial η^2 = .03, with high knowledge students using significantly fewer hints (M = .30, SD = .87) than their low knowledge peers (M = .80, SD = 1.61). Differences in hint usage are depicted visually in Figure 7.13.

7.3.4.3 Assignment Attempts

When considering the total number of attempts students used while completing their assignments, Levene's test remained significant, F(9, 529) = 6.29, p = .000. The interaction effect of condition and knowledge level was not significant, F(4, 529) = 1.15, p = .330. However, the main effect of condition was significant, F(1, 529) = 5.11, p = .000, partial η^2 = .04. Pairwise comparisons revealed significant differences (p < .05) between Alternative (M = 3.39, SD = 2.05) and TC (M = 4.61, SD = 3.25), TNC (M = 4.75, SD = 3.41), VC (M = 4.38, SD = 3.27), and VNC (M = 5.11, SD = 3.53). No other significant differences were observed between groups, suggesting that the Alternative group was out of place, with fewer attempts on average than all other groups (which should not have been the case). The main effect of prior knowledge level was also significant, F(1, 529) = 13.07, p = .000, partial η^2 = .02, with low knowledge students using more attempts on average (M = 4.93, SD = 3.72) than their high knowledge peers (M = 3.77, SD = 2.30). Observed effects of attempt count are depicted in Figure 7.14.

7.3.4.4 Assignment Time

This variable was slightly more difficult to interpret because it was heavily positively skewed (a typical issue with time measures in ASSISTments, as students can easily walk away from their homework assignments without closing the tutor, leaving high end outliers). In order to normalize this variable, 34 high end cases were trimmed to one hour of activity and the natural log of this adjusted variable was taken to address skew. As such, means represent the natural log of students' total time across first responses to all problems.

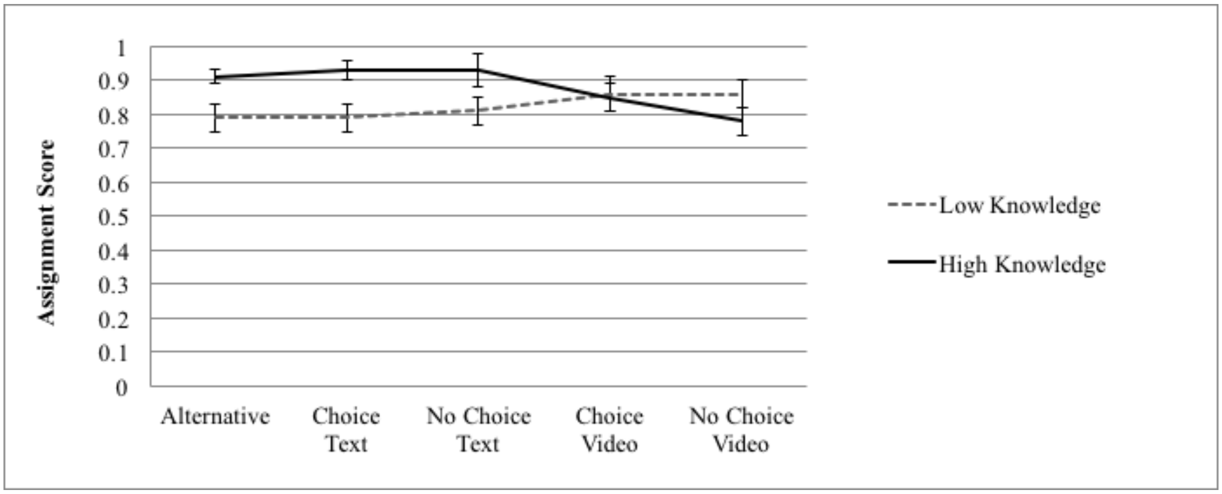


Figure 7.12. Assignment score by condition and knowledge level.

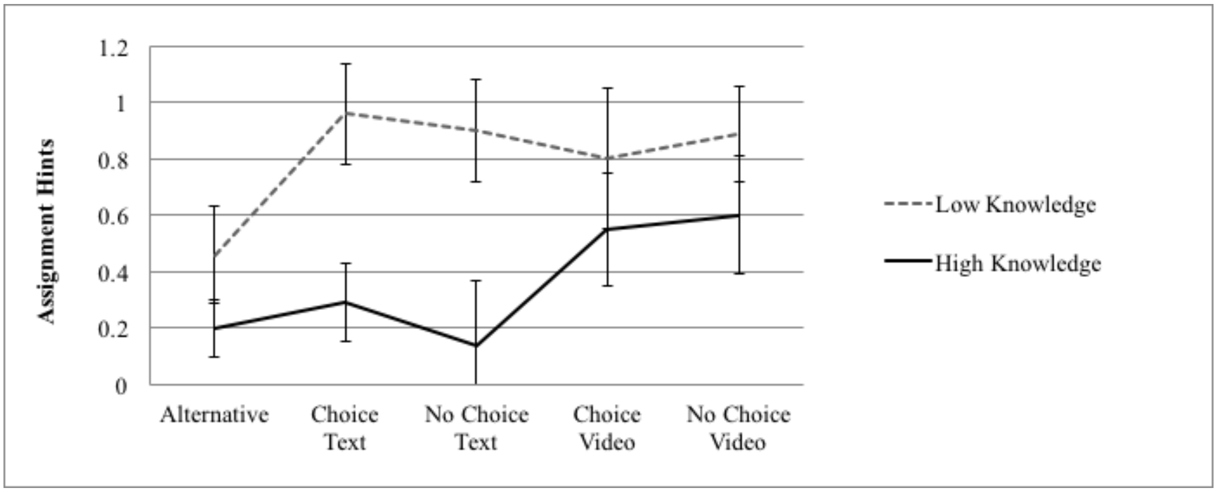


Figure 7.13. Assignment hint usage by condition and knowledge level.

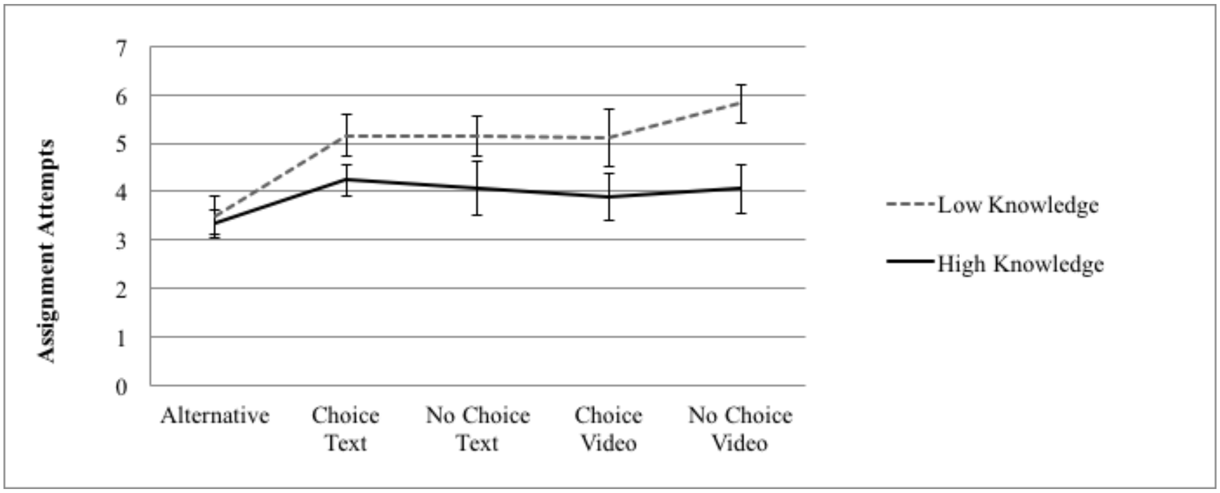


Figure 7.14. Assignment attempt usage by condition and knowledge level.

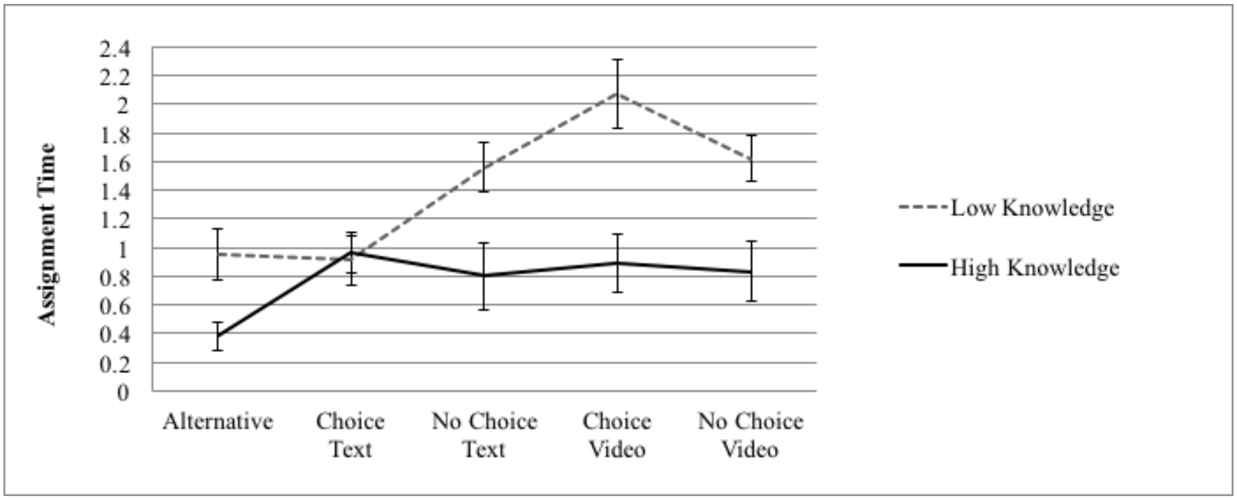


Figure 7.15. Assignment first response time (ln(minutes)) by condition and knowledge level.

Table 7.10. Three-way independent factors ANOVA results of hint medium, choice, and prior knowledge level on within assignment metrics.

	SS	df	MS	F	p	1-β
Assignment Score						
Condition	.13	4	0.03	0.48	.748	.17
Knowledge	.35	1	0.35	5.18	.023	.62
Condition X Know	.80	4	0.20	2.94	.020	.79
Error	36.08	529	.068			
Assignment Hints						
Condition	11.94	4	2.99	1.97	.098	.59
Knowledge	21.26	1	21.26	14.01	.000	.96
Condition X Know	5.55	4	1.39	0.91	.455	.29
Error	802.72	529	1.52			
Assignment Attempts						
Condition	176.28	4	44.07	5.11	.000	.97
Knowledge	112.64	1	112.64	13.07	.000	.95
Condition X Know	39.77	4	9.94	1.15	.330	.36
Error	4,558.41	529	8.62			
Assignment Time						
Condition	34.28	4	8.57	6.28	.000	.99
Knowledge	43.35	1	43.35	31.74	.000	1.00
Condition X Know	17.91	4	4.48	3.28	.011	.84
Error	689.67	505	1.37			

Table 7.11. Means (SDs) of within assignment metrics by hint medium, choice, and knowledge level.

	Low Knowledge (n = 225)	**High Knowledge (n = 314)**	**Total (n = 539)**
Assignment Score			
Alternative *(n = 190)*	0.79 (0.35)	0.91 (0.24)	0.87 (0.28)
Choice Text *(n = 122)*	0.79 (0.31)	0.93 (0.14)	0.88 (0.24)
No Choice Text *(n = 76)*	0.81 (0.29)	0.93 (0.19)	0.86 (0.27)
Choice Video *(n = 63)*	0.86 (0.22)	0.85 (0.28)	0.86 (0.26)
No Choice Video *(n = 88)*	0.86 (0.22)	0.78 (0.35)	0.83 (0.28)
Total (*n = 539*)	0.82 (0.29)	0.89 (0.24)	0.86 (0.26)
Assignment Hints			
Alternative *(n = 190)*	0.46 (1.15)	0.20 (0.61)	0.27 (0.79)
Choice Text *(n = 122)*	0.96 (1.76)	0.29 (0.86)	0.56 (1.33)
No Choice Text *(n = 76)*	0.90 (1.70)	0.14 (0.59)	0.62 (1.44)
Choice Video *(n = 63)*	0.80 (1.66)	0.55 (1.03)	0.65 (1.31)
No Choice Video *(n = 88)*	0.89 (1.75)	0.60 (1.48)	0.77 (1.65)
Total (*n = 539*)	0.80 (1.61)	0.30 (0.87)	0.51 (1.26)
Assignment Attempts			
Alternative *(n = 190)*	3.48 (2.64)	3.36 (1.80)	3.39 (2.05)
Choice Text *(n = 122)*	5.16 (3.95)	4.23 (2.64)	4.61 (3.25)
No Choice Text *(n = 76)*	5.15 (3.79)	4.07 (2.54)	4.75 (3.41)
Choice Video *(n = 63)*	5.12 (4.19)	3.89 (2.45)	4.38 (3.27)
No Choice Video *(n = 88)*	5.81 (3.81)	4.06 (2.81)	5.11 (3.53)
Total (*n = 539*)	4.93 (3.72)	3.77 (2.30)	4.25 (3.03)
Assignment Time (ln)			
Alternative *(n = 190)*	0.95 (1.59)	0.38 (1.18)	0.52 (1.31)
Choice Text *(n = 122)*	0.91 (0.64)	0.96 (1.03)	0.94 (0.90)
No Choice Text *(n = 76)*	1.56 (1.14)	0.80 (0.99)	1.27 (1.14)
Choice Video *(n = 63)*	2.07 (1.37)	0.89 (1.07)	1.36 (1.32)
No Choice Video *(n = 88)*	1.62 (1.39)	0.83 (1.05)	1.32 (1.32)
Total (*n = 539*)	1.36 (1.31)	0.67 (1.12)	0.95 (1.25)

Levene's test remained significant in this analysis, F(9, 505) = 3.38, p = .000. The interaction effect of condition and knowledge level was significant, F(4, 505) = 3.28, p = .011, partial η^2 = .03. Simple effects tests revealed that this effect was driven in part by significant differences across groups amongst high knowledge students, F(4, 505) = 3.74, p = .005, partial η^2 = .03. Pairwise comparisons revealed significant differences (p < .05) between Alternative (M = .38, SD = 1.18) and TC (M = .96, SD = 1.03), VC (M = 0.89, SD = 1.07), and VNC (M = .83, SD = 1.05). Differences were also significant across groups amongst low knowledge students, F(4, 505) = 6.20, p = .000, partial η^2 = .05. Pairwise comparisons revealed significant differences (p < .05) between Alternative (M = .95, SD = 1.59) and TNC (M = 1.56, SD = 1.14), VC (M = 2.07, SD = 1.37), and VNC (M = 1.62, SD = 1.39), and between TC (M = .91, SD = .64), and TNC (M = 1.56, SD = 1.14), VC (M = 2.07, SD = 1.37), and VNC (M = 1.62, SD = 1.39). Significant differences were also observed between high and low knowledge students in the Alternative group (F(1, 505) = 7.93, p = .005, partial η^2 = .02), in the TC group (F(1, 505) = 7.19, p = .008, partial η^2 = .01), in the VC group (F(1, 505) = 14.80, p = .000, partial η^2 = .03), and in the VNC group (F(1, 505 = 9.06, p = .003, partial η^2 = .02). Low knowledge students took longer in all conditions other than chosen text, with means and standard deviations presented in Table 7.11 and shown in Figure 7.15.

This significant interaction effect was supported by significant main effects of both condition and prior knowledge level. Pairwise comparisons for the main effect of condition, F(4, 505) = 6.28, p = .000, partial η^2 = .05, revealed significant differences ($p < .05$) between Alternative (M = .52, SD = 1.31) and TNC (M = 1.27, SD = 1.14), VC (M = 1.36, SD = 1.32), and VNC (M = 1.32, SD = 1.32), with trending significance compared to TC (M = .94, SD = .90) (p = .071). A significant difference was also observed between TC (M = .94, SD = .90) and VC (M = 1.36, SD = 1.32), with students taking longer to respond when requesting video feedback. The main effect of prior knowledge level was also significant, F(1, 505) = 31.74, p = .000, partial η^2 = .06, with low knowledge students (M = 1.36, SD = 1.31) taking longer than high knowledge (M = .67, SD = 1.12), as anticipated.

7.4 Discussion

Past work has provided mixed evidence for the significant effects of choice on within-assignment performance in ASSISTments (Ostrow & Heffernan, 2015; Ostrow, 2015). Even when differences were not significant, these studies associated choice with stronger performance, lower overall hint usage, and fewer problems required for mastery. In addition, past work also provided mixed evidence for the effectiveness of video hints, especially amongst low knowledge students (Ostrow & Heffernan, 2014; Ostrow & Heffernan, 2015; Ostrow, 2015). However, sample sizes have been too small to properly interpret interaction effects. Further, it is possible that mixed evidence for these effects may have been due in part to variation in factors like skill or grade level. The present work sought to clarify these effects using a well-controlled experimental design including a posttest and self-report items from the IMI.

The sample in the present work was much like that observed in past work, with two thirds of students assigned to the choice condition selecting text hints (Ostrow & Heffernan, 2015; Ostrow, 2015). Completion rates in the present work were also adequate, with approximately 80% of students finishing their assignments across conditions. Hint usage was relatively low (16-22%), just as observed in previous work, suggesting either that the skill was not very difficult or that the assignment was used as review material and ceiling effects caused few students to require extra assistance. Opt-in for the IMI was also relatively high, with approximately half of the analytic sample participating. Interestingly, this participation rate was considerably lower than that observed when survey opt-in was placed at the beginning of a similar study (see Chapter 8), possibly due to fatigue or anticipated teacher expectations. Recent work suggesting reduced subscales for the IMI (Ostrow & Heffernan, In Press, see Chapter 6) was supported by the current work, with high reliability observed in the responses of the present sample. In addition, results suggested high within-assignment accuracy (86% on average) but moderate posttest scores (43% on average), suggesting that the posttest was well developed and more difficult than the assignment. Future analysis assessing each posttest item independently would be helpful to reveal the distribution of accuracy across transfer items of varied difficulty and levels of persistence in the face of challenge.

The first research question of the present work considered the effects of choice and hint medium on posttest score. Primary analyses revealed an interaction between hint medium and knowledge level in which low knowledge students excelled following video hint conditions (requested or assigned). Although this was only a trending effect, the same was not observed in high knowledge students. Instead, high knowledge students excelled beyond their low knowledge peers following text hints (requested or assigned). This finding suggested that although choice may not have played a critical role in students' performance, hint medium was associated with differential performance by knowledge level. High knowledge students showed lower performance following video hints, especially if they were randomly assigned to the condition. As such, the element of choice seemed to calibrate the effect of hint medium, even if not in a statistically significant manner.

When considering only the students who requested hints, small sample size led to underpowered analyses and no significant main effects or interactions were observed. Future work would need a sample size at least double that considered in the present work to reach adequate power, suggesting a powerful limitation of the present work. Still, this treated sample exhibited lower posttest scores overall and students with choice performed better after choosing and utilizing video feedback. Across hint mediums, students were performing at approximately the same level when not given choice, suggesting that a more robust sample might reveal a significant effect of choice. Choice also seemed to calibrate the effect of hint medium in the treated sample, where the effect of video increased when it was chosen, despite the lack of a significant interaction effect. As such, recommendations based on the findings of the present work are to increase the amount of video hints available in ASSISTments for low knowledge students, while allowing students choice of their hint medium to enhance its efficacy. Such a course of action would likely promote learning outcomes among low knowledge students while leaving high knowledge students without consequence.

The second research question considered the effects of choice and hint medium on students' Perceived Autonomy and Interest/Enjoyment scale scores. Based on the principles of Self-Determination Theory, it was theorized that supporting autonomy through the provision of hint medium might increase Perceived Autonomy and Interest/Enjoyment. Results suggested that no significant differences were observed in Perceived Autonomy scale scores. It is possible that the choices were not meaningful enough, or that the available choice was an ineffective option choice rather than an effective action choice (Reeve, Nix, & Hamm, 2003). It is also possible that the choice of hint medium may not have been impactful enough to influence Perceived Autonomy responses, given that scale items were tailored to students' participation in and completion of the entire assignment and their feelings of ownership over the assignment itself rather than the way the assignment was completed. It is possible that effects of the provided choice were not captured by the scale's tone, posing a potential limitation to the present work.

Further, despite observing no effect of hint medium or choice, results suggested an effect of knowledge level in students' Interest/Enjoyment, or the primary scale signifying intrinsic motivation. Interestingly high knowledge students exhibited lower interest and enjoyment than their low knowledge peers. This observation was made more complex by the significant interaction observed between hint medium and prior knowledge level: low knowledge students requesting or assigned to video hints exhibited higher interest/enjoyment than low knowledge students requesting or assigned to text hints. This effect was not observed within high knowledge students. In addition, if assigned to video, low knowledge students reported higher interest and enjoyment than their high knowledge peers, by more than one full scale point on average. Coupled with the effects observed for posttest score performance, where video hints were associated with stronger learning outcomes among low knowledge students, these results suggest the potential benefit of choice and video among this population.

IMI findings suggest that for struggling students, video feedback may significantly improve performance and thereby strengthen perceptions of competence, which ultimately improves measures of intrinsic motivation like interest and enjoyment. Given that the posttest took place before the IMI, it was not shocking that regression analysis revealed that Interest/Enjoyment scores increased by .97 for every 1 SD increase in posttest score. However, coupled with the finding that high knowledge students were reporting significantly lower Interest/Enjoyment, significant correlation may have been driven by the performance outcomes of low knowledge students. It is also possible that low knowledge students cared more about the posttest challenge problems based on expectations of teacher perception. They may have persisted longer or focused more, believing that their teacher had assigned the content and that the challenge problems might influence their assignment scores. Higher knowledge students may have more appropriately interpreted the structure of the assignment, understanding that they had already completed the primary content and that the challenge problems were essentially a waste of time, thereby reducing effort and score, and undermining interest and enjoyment overall. It is also possible that the effect was driven by mechanisms that were not directly measured, like students' perceptions of the structure level of the assignment or their perceptions of the skill topic itself. Perhaps higher interest and enjoyment was observed for low knowledge students receiving or requesting video feedback because of the visual presence of the author, rather the content's delivery. While inability to pinpoint a mechanism is not necessarily a limitation of the present work given it was not a primary goal, future work in this area should consider additional metrics and tailor moderation and mediation analyses to better define why these effects were observed.

The third research question considered the effects of experimental condition (hint medium X choice) on within assignment performance metrics. These analyses were included in the present work to mirror past work that considered only within-assignment metrics (Ostrow & Heffernan, 2015; Ostrow, 2015). Past work suggested that choice and hint medium influenced assignment accuracy, hint usage, attempt count, and time on task (Ostrow, 2015). In the present work, these metrics were observed among five groups of students: the four experimental conditions and the alternative sample that participated in a business-as-usual Skill Builder.

Analysis of assignment score revealed a significant interaction effect between condition and knowledge level, in which high knowledge students assigned video hints exhibited lower assignment scores than those in the alternative, choice text, and assigned text conditions. Low knowledge students did not exhibit significant differences across groups, but trends suggested that video hints improved within-assignment performance. Low knowledge students choosing or assigned to video hints showed performance on par with their high knowledge peers in other conditions. However, video undermined the performance of high knowledge students. It is possible that high knowledge students were stronger readers, or were capable of greater focus with text material, making video hints an annoyance that slowed their progress and distracted them from the task at hand.

Analysis of assignment hint usage only revealed greater hint usage within low knowledge students across conditions, which was to be expected. However, when considering assignment attempts, students in the alternative group used one attempt less on average than all other groups. It is not clear why this effect was observed, as it should not have been inherent to the experimental design. The Alternative group would have shared the same structure as the assigned text condition (but without modifications to the original hints in the first ten problems). It was anticipated that these groups would look approximately equivalent, which was not the case. More in line with anticipated results, there was also a significant main effect of prior knowledge with low knowledge students using more attempts on average than their high knowledge peers.

Results for the analysis of assignment first response time were slightly more difficult to interpret given that the variable was transformed using the natural log to address positive skew. However, exponentiation of the resulting means should return the scale to minutes for comparison. The significant interaction effect of condition and knowledge level suggested that high knowledge students exhibited faster overall response times when completing the alternative assignment (1.46 minutes) than when participating in the chosen text (2.61 minutes), chosen video (2.44 minutes), or assigned video (2.29 minutes) conditions. Further, low knowledge students exhibited faster overall problem solution in the alternative assignment (2.59 minutes) than when assigned to text (4.76 minutes), when choosing video (7.92 minutes), or when assigned to video (5.05 minutes). Students were also performing faster when text was chosen rather than assigned but slower when video was chosen rather than assigned. As would be expected, low knowledge students were almost always taking longer than their high knowledge peers (chosen text was nearly equivalent). Interestingly, low knowledge students completed the alternative assignment with overall response times on par with their high knowledge counterparts participating in all experimental conditions. However, high knowledge students performed faster across all experimental conditions than low knowledge students. Low knowledge performed slowest when they chose video hints and fastest when they chose text hints. Without choice, low knowledge students spent about as long with text hints as with video hints.

These results suggest that students that chose video may have spent longer within hints, watching and integrating the hint feedback, whereas those that chose text hints may have been gaming through their assignment. This variable reflects first response time across problems and therefore does not take time spent within hints into account, which is a considerable limitation. Future work considering the amount of time students spent within matched content feedback and how time on task aligns with performance outcomes could help to define a potential mechanism driving the observed effects. Further, while low knowledge students spent more time solving problems after choosing video hints, they did not use significantly more hints or attempts and exhibited within-assignment scores as high as their high knowledge peers. Considering these effects together with improved posttest performance and greater interest and enjoyment, providing the option of video hints for low knowledge students may present a low-cost intervention that effectively targets the achievement gap.

The primary overall limitation of this work was small sample size for a complex analysis resulting in low power. In addition, choice was not observed to be statistically significant on posttest performance or motivational metrics, which was possibly because the choice provided was not substantial enough or did not impart value or ownership for students as they completed their assignment. The primary contribution of this work was the practical association of video hints with performance and motivation improvements in low knowledge students, especially in the presence of support for autonomy. These students seemed to attend to their chosen hint medium more, resulting in stronger within-assignment performance metrics and higher posttest scores coupled with higher self-reported interest and enjoyment. As such, the option of video hints may offer a low-cost intervention that targets the achievement gap that could potentially demonstrate synergistic or compounding effects when repeated across assignments.

References

Baddeley, A. D. & Hitch, G. (1974). Working Memory. Psychology of learning and motivation. 8: 47-89.

Baker, R. (2016) Stupid tutoring systems, intelligent humans. International Journal of Artificial Intelligence in Education. 26 (2): 600-614.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego-depletion: Is the active self a limited resource? Journal of Personality and Social Psychology. 74: 1252–1265.

Becker, D. A. (1997). The effects of choice on auditors' intrinsic motivation and performance. Behavioral Research in Accounting. 9: 1–19.

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

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*. 125: 627-668.

Deci, E. L., Schwartz, A. J., Sheinman, L., & Ryan, R. M. (1981). An instrument to assess adults' orientation toward control versus autonomy with children: Reflections on intrinsic motivation and perceived competence. Journal of Educational Psychology. 73: 642-650.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist.* 26: 325–346.

Dryden, J. (2018). Autonomy. Internet Encyclopedia of Philosophy. Retrieved from: https://www.iep.utm.edu/autonomy/

Escueta, M., Quan, V., Nickow, A. J., & Oreopoulos, P. (2017). Education Technology: An Evidence-Based Review. National Bureau of Economic Research Working Paper No. 23744.

Field, A.: Discovering statistics using IBM SPSS Statistics. 4th edn. SAGE Publications, London (2013).

Flowerday, T., & Schraw, G. (2000). Teacher beliefs about instructional choice: A phenomenological study. Journal of Educational Psychology. 92: 634–645.

Flowerday, T., Schraw, G., & Stevens, J. (2004). The role of choice and interest in reader engagement. The Journal of Experimental Education. 72: 93–114.

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): 497-514.

Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology. 52*: 890–898.

Grolnick, W. S., Ryan, R. M., & Deci, E. L. (1991). Inner Resources for School Achievement: Motivational Mediators of Children's Perceptions of Their Parents. Journal of Educational Psychology. 83(4): 508-517.

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. *International Journal of Artificial Intelligence in Education*. 24 (4): 470-497.

Intrinsic Motivation Inventory, http://selfdeterminationtheory.org/questionnaires/, last accessed 2018/02/07.

Iyengar, S. S., & Lepper, M. R. (1999). Rethinking the value of choice: A cultural perspective on intrinsic motivation. Journal of Personality and Social Psychology. 76: 349–366.

Katz, I. & Assor, A. (2007). When Choice Motivates and When It Does Not. *Educational Psychology Review*. 19 (4): 429-442.

Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, G. & Soffer, D. (2013). Estimating the effect of web-based homework. In Lane, Yacef, Mostow & Pavlik (eds)

Proceedings of the 16th International Conference on Artificial Intelligence in Education. Springer Berlin Heidelberg. 7926: 824-827.

Kernan, M. C., Heimann, B., & Hanges, P. J. (1991). Effects of goal choice, strategy choice, and feedback source on goal acceptance, performance, and subsequent goals. Journal of Applied Social Psychology. 21: 713–733.

Koedinger, K., McLaughlin, E. & Heffernan, N. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. Journal of Educational Computing Research. Baywood Publishing. 4: 489-510.

Koestner, R. & Losier, G. F. (2002). Distinguising Three Ways of Being Internally Motivated: A Closer Look at Introjection, Identification, and Intrinsic Motivation. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 101-121). Rochester, NY: University of Rochester Press.

Koestner, R., Losier, G. F., Vallerand, R. J., & Carducci, D. (1996). Identified and introjected forms of political internalization: Extending self-determination theory. *Journal of Personality and Social Psychology*. 70: 1025-1036.

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*. 29: 171-173.

Mendicino, M., Razzaq, L. & Heffernan, N.T. (2009). Improving learning from homework using intelligent tutoring systems. *Journal of Research on Technology in Education*. 41 (3): 331-346.

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): 1475-1490.

National Governors Association Center for Best Practices (NGACBP) & Council of Chief State School Officers (CCSSO). (2010). Common Core State Standards. Washington, DC: Authors.

Ostrow, K. S. (2015). A Multifaceted Consideration of Motivation and Learning within ASSISTments. MS Thesis, Social Science & Policy Studies, Worcester Polytechnic Institute.

Ostrow, K. S. (2018). Autonomy within ASSISTments: OSF Registration Page. https://osf.io/wg2d6/, last accessed 2018/04/09.

Ostrow, K. & Heffernan, C. (2015). ASSISTments TestBed Resource Guide. https://www.assistmentstestbed.org, last accessed 2018/02/06.

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. 296-299.

Ostrow, K. & Heffernan, N. (2015). The Role of Student Choice Within Adaptive Tutoring. In Conati, Heffernan, Mitrovic & Verdejo (eds.) Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015). Springer International Publishing. Madrid, Spain. June 22-26. 752-755.

Ostrow, K. S. & Heffernan, N. T. (In Press). Testing the Validity and Reliability of Intrinsic Motivation Inventory Subscales within ASSISTments. Submitted to the Proceedings of the 2018 International Conference of AIED.

Ostrow, K., Heffernan, N. T., Heffernan, C., & Peterson, Z. (2015). Blocking vs., Interleaving: Examining Single-Session Effects within Middle School Math Homework. In Conati, Heffernan, Mitrovic & Verdejo (eds.) Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015). Springer International Publishing. Madrid, Spain, June 22-26. 388-347.

Ostrow, K. S., Heffernan, N. T., & Williams, J. J. (2017). Tomorrow's EdTech Today: Establishing a Learning Platform as a Collaborative Research Tool for Sound Science. *Teachers College Record.* 119(3): 1-36.

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): 270-300.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior Performance*. 16: 366-387.

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review*. 18 (4): 315-341.

Piaget, J. (1971). *Biology and knowledge*. Chicago: University of Chicago Press.

Razzaq, L. & Heffernan, N. (2009). To Tutor or Not to Tutor: That is the Question. In Dimitrova, Mizoguchi, du Boulay & Graesser (Eds.) Proceedings of the 2009 Artificial Intelligence in Education Conference. IOS Press. 457-464.

Reeve, J. (2002). Self-determination theory applied to educational settings. In E. L. Deci & R. M. Ryan (Eds.), Handbook of self-determination research (pp. 184-203). Rochester: University of Rochester Press.

Reeve, J., & Jang, H. (2006). What teachers say and do to support students' autonomy during a learning activity. *Journal of Educational Psychology. 98:* 209–218.

Reeve, J., Nix, G., & Hamm, D. (2003). Testing models on the experience of self-determination in intrinsic motivation and the conundrum of choice. *Journal of Educational Psychology. 95*: 375–392.

Roschelle, J., Feng, M., Murphy, R. & Mason, C. (2016). Online Mathematics Homework Increases Student Achievement. AERA OPEN. October-December 2016. 2 (4): 1–12. DOI: 10.1177/2332858416673968

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology. 43*: 450–461.

Ryan, R. M. & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*. 57: 749-761.

Ryan, R. M. & Deci, E. L. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist*. 55(1): 68-78.

Ryan, R. M., & Deci, E. L. (2002). An overview of self-determination theory. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 3-33). Rochester, NY: University of Rochester Press.

Schell, K. L., Oswald, F. L. (2013). Item grouping and item randomization in personality measurement. *Personality and Individual Differences.* 55: 317-321.

Schwartz, B. (2000). Self-determination: The tyranny of freedom. *American Psychologist*. 55: 79–88.

Singer, N. (May 13, 2017). How Google Took Over the Classroom. New York Times. https://www.nytimes.com/2017/05/13/technology/google-education-chromebooks-schools.html last accessed 2018/04/09.

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, Bull, Kay, & Mitrovic (eds) Proceedings of the 15th International Conference on Artificial Intelligence in Education Conference. Springer Berlin Heidelberg. 6738: 328–336.

Soffer, D., Das, V., Pellegrino, G., Goldman, S., Heffernan, N., Heffernan, C., & Dietz, K. (2014). Improving long-term retention of mathematical knowledge through automatic reassessment and relearning. American Educational Research Association Conference, Division C - Learning and Instruction / Section 1c: Mathematics.

Sweller, J. (2004). Instructional design consequences of an analogy between evolution by natural selection and human cognitive architecture. *Instructional Science. 32*: 9–31.

Vallerand, R. J., Blais, M. R., Brière, N. M. & Pelletier, L. G. (1989). Construction and validation of the Academic Motivation Scale. Canadian Journal of Behavioral Sciences. 21: 323-349.

Van Loon, A. M., Ros, A., & Martens, R. (2012). Motivated learning with digital learning tasks: what about autonomy and structure? *Educational Technology Research and Development*. 60 (6): 1015-1032.

8 Belonging

This chapter describes a randomized controlled trial designed to target the basic psychological need of relatedness or belonging. Supported by Self-Determination Theory, the work employs a value affirmation intervention intended to enhance learning outcomes. The effect of the intervention is examined with regard to math performance across students' self-reported baseline belonging and prior knowledge levels. This chapter will be submitted to a journal venue following dissertation defense. This chapter will read as a strong draft and will be submitted to a journal venue following completion of my degree. As such, further revision may occur before formal presentation.

Proper citation of this chapter is as follows:

Ostrow, K. S. (In Preparation). Value Affirmation as an Intervention for Online Learning: Assessing the Roles of Prior Knowledge and Sense of Belonging.

8.1 Introduction

A sense of belonging is a critical part of human existence that has been acknowledged openly in psychological study since Maslow's (1943) Hierarchy of Needs. Maslow's work considered the feeling of belonging, or being accepted by and affiliated with others, a prerequisite to one's ability to achieve competency or accomplish intellectual goals (Maslow, 1943). While various researchers have extended definitions of the needs underlying this hierarchy, all maintain that a sense of belonging plays a critical role in development (Alderfer, 1972; Wahba & Bridgewell, 1976; Thompson, Grace, & Cohen, 2001; Nohria, Lawrence, & Wilson, 2001; Ryan & Deci, 2009). Essentially, humans are social creatures (Vygotsky, 1978) that are best able to thrive, learn, and grow when they feel safe and connected.

Within education, a student's sense of belonging plays a similarly critical role. Goodenow (1993b) defined scholastic belonging as "the extent to which students feel personally accepted, respected, included, and supported by others in the school social environment." Reduced academic belonging has been linked to lower class participation by at-risk students (Wehlage, 1989), higher dropout rates (Finn, 1989), and increased delinquent behavior, substance abuse, and gang related activity (Burnett & Walz, 1994; Goff & Goddard, 1999). It has been hypothesized that inducing a sense of belonging in educational contexts affects learning outcomes by increasing students' confidence through altered expectancy-value (Goodenow, 1993a). The present work considers and attempts to intervene upon students' sense of belonging within the context of an online learning assignment to broaden the field's understanding of belonging to include effects observed within digital learning environments.

8.1.1 Motivational Interventions in Education

Researchers have noted the importance and fragility of student motivation for decades, with early work noting motivation as a systemic problem for education (Nicholls, 1979). That problem persists today, and despite calls for the study of controlled motivational interventions within educational contexts, published work in the area has declined over the past decade (Lazowski & Hulleman, 2016). Concurrent research on classroom norms and culture has focused more broadly on positive learning climates that promote interaction and support learning (Pianta & Hamre, 2009). For instance, the Classroom Assessment Scoring System (CLASS) was developed to measure emotional supports, classroom organization, and instructional supports in pre-K-12 classrooms (Pianta, La Paro, & Hamre, 2008; Pianta & Hamre, 2009). Regardless of researchers' approach to the issue, it is generally understood that motivation begets persistence, effort, higher quality work, and ultimately, learning and grades (Lazowski & Hulleman, 2016; Fredericks et al., 2011; Maehr & Midgley, 1999; Goodenow, 1993b; Schwartz, Tsang, & Blair, 2016) and that the construct is an important area of focus for research in education.

High socio-emotional functioning and motivation are also increasingly linked to school readiness and productive learning behaviors (Pianta & Hamre, 2009). However, it has been shown that motivation tends to decline as students enter adolescence and transition from elementary to middle school, with significant reductions observed from 3rd to 8th grade (Lepper, Corpus, Iyengar, 2005). It has also been observed that only approximately 25% of certified teachers support socio-emotional functioning in effective ways (Pianta, Belskhy, Houts, Morrison & NICHD ECCRN, 2007). Together, these factors speak to the need for additional research into effective motivational interventions, especially during the critical period of adolescence.

Examinations of motivational interventions have consistently revealed vast benefits of interventions intended to boost student motivation coupled with minor (if any) consequence. Lazowski & Hulleman (2016) conducted a meta-analysis of 74 motivational interventions within educational contexts. These interventions covered 15 commonly referenced theoretical frameworks, including achievement goals, expectancy-value, implicit theories of intelligence (mindsets), self-affirmation, self-determination, and social belongingness, among others. Findings revealed a powerful effect size ($d = 0.49$, 95% CI = [0.43, 0.56]) for motivational interventions on average (Lazowski & Hulleman, 2016). Given that motivational interventions tend to carry low costs, requiring few resources and minimal time on task, additively or recursively implementing motivational supports for students may offer a simple, empathetic approach toward greater learning.

Fluctuations in the study of students' sense of belonging have mirrored that of other motivational frameworks (Ma, 2003). Motivating students within socially connected contexts increases their likelihood of positive social and academic trajectories (Hamre & Pianta, 2001). Understanding this potential, Goodenow (1993b) established the Psychological Sense of School Membership (PSSM) scale as the first valid and reliable measure of the social and contextual elements influencing students' sense of belonging. Since the creation of this measure nearly 25 years ago, it has been implemented in more than 41 published studies (You, et al., 2011) in which researchers have examined the implications of a sense of belonging amongst diverse groups of students. In their meta-analysis, Lazowski & Hulleman (2016) included five studies specifically examining the efficacy of belongingness interventions, observing a moderate average effect size (*d* = 0.35, 95% CI = [0.07, 0.63]). Research has also suggested that sense of belonging varies primarily at the student-level, with the majority of this variance linked to students' self-esteem and health factors (more so than gender, race, or ethnicity), and the majority of school-level variance linked to locale or urbanicity (Ma, 2003; Goodenow, 1993b). As such, research focused on strengthening students' sense of belonging through self-esteem and value interventions, or that which considers the moderating effects of belonging, is likely to prove effective in fostering students' motivation and enhancing their academic trajectories.

8.1.2 Belonging and Self-Determination Theory

One popular motivational framework that supports the promotion of relatedness and students' sense of belonging is Self-Determination Theory (SDT). Originally postulated by Deci & Ryan (1985), SDT offers a macro-theory of human motivation that combines the innate need to belong (Maslow, 1943) with earlier work on students' feelings of competence (or, their ability to meet academic challenges: Harter, 1978; White, 1963) and autonomy (or, participating of their own volition: deCharms, 1968; Deci, 1975). The theory promotes intrinsic motivation and the internalization of extrinsic motivators by shifting one's locus of control (Rotter, 1954) through supports for autonomy, belonging, and competence. Lazowski & Hulleman's recent meta-analysis (2016) examined 11 motivational interventions rooted in Self-Determination Theory, observing a high effect size (d = 0.70, 95% CI = [0.53, 0.87]) on average.

Internalization of extrinsic motivators is an appropriate approach for education. Few students find their mathematics homework intrinsically motivating, but there may be room to strengthen ownership of skills and abilities, potentially leading to integrated task regulation (Ryan & Deci, 2000; Niemiec & Ryan, 2009). As it stands, classroom performance is commonly driven by external controls (i.e., grades) that can undermine students' sense of belonging and instill anxiety, boredom, and alienation (Niemiec & Ryan, 2009). However, teachers that genuinely connect to their students through respect, patience, and clear communication of classroom norms tend to increase students' sense of belonging and de-emphasize their value of extrinsic controls (Niemiec & Ryan, 2009; Pianta & Hamre, 2009). As such, two common approaches support successful belonging interventions: changing students' attributions or changing the classroom social structure to promote relatedness (Schwartz, Tsang, & Blair, 2016). It may also be possible to help students integrate extrinsic controls through the tone of content and curricula via learning opportunities that enrich the classroom. The present work examines whether an online learning platform can serve as this vehicle.

8.1.3 Belonging and Self-Affirmation Theory

Bandura (1977) rightly stated that while learners with self-belief may not succeed, learners without self-belief will almost always fail. Emphasis on self-belief and self-esteem via self-affirmation falls at a crossroads with research on students' sense of belonging as a common and effective intervention (Cohen & Garcia, Apfel, & Master, 2006; Cohen & Garcia, 2008; Walton & Cohen, 2011). Aguilar, Walton, & Wieman (2014) depicted two recursive cycles that inflate or deflate a students' sense of belonging as influenced by performance and social context (see Figure 8.1 for an adaptation of this framework). Within these recursive cycles, success and positive feedback from others can motivate students and enhance task performance, while failure and negative feedback from others can demotivate students and prompt anxiety or task avoidance (Schwartz, Tsang, & Blair, 2016).

Considering the recursive nature of students' sense of belonging, researchers argue that a well-placed intervention can halt a negative feedback loop and reorient students to the task at hand, potentially preventing downward socio-emotional spirals that ultimately hinder performance (Cohen & Garcia, 2008; Cohen, et al. 2009). This process has proven true for combating stereotype threat, or students' fear of performing in a manner that confirms a negative stereotype about them or a group that they identify with, amongst minority college freshmen (Cohen, Garcia, Apfel, & Master, 2006; Walton & Cohen, 2011) and struggling middle school students (Cohen & Garcia, 2006; Cohen & Garcia, 2008), resulting in lasting improvements to academic performance in both cases. This intervention is thought to work by allowing students to maintain their self-integrity by bolstering and affirming their values, thereby making the human condition more apparent and enhancing a sense of belonging or acceptance within the learning environment. The present work employs an abbreviated version of a previously validated self-affirmation intervention (Cohen, et al. 2009) within a single mathematics assignment in an online learning platform with the goal of enhancing student performance. As past work has also shown that this intervention is especially potent for high-risk students, students' prior knowledge level will be assessed as a potential moderator.

8.1.4 Technology in the Classroom

Recent government reports have suggested that technology will soon serve as a key driver in delivering educational interventions to students, and that those interventions will likely serve to personalize the learning experience (U.S. DoE, 2010). Intelligent tutoring systems and online learning platforms are gaining momentum through use in classrooms around the world, and the increased popularity of Massive Open Online Courses for higher education and distance learning is challenging popular educational contexts. Many of these systems were established in pursuit of a future in which artificial intelligence would make decisions regarding students' education. However, the Educational Data Mining and Learner Analytics fields have revealed the true power of these platforms: as vehicles for collecting data that teachers can act upon, effectively amplifying what they are able to understand about their students in support of data driven classroom discourse (Baker, 2016). Thus, rather than reinventing the educational context, many of these systems are growing into tools that enrich existing learning processes.

ASSISTments, an online learning platform focused primarily on middle school mathematics, is one such tool that also supports the creation of randomized controlled experiments within classwork and homework assignments, coupling experimental data with the tutor's log files and serving as a shared scientific instrument for researchers studying the efficacy of learning interventions in authentic learning environments (Ostrow, et al. 2016). Capabilities for transparent educational research at scale through online learning environments are increasing and research once resigned to laboratories or costly efficacy trials can now be conducted with fewer obstacles (Ostrow, Heffernan, & Williams, 2017). Following this trend, the present work seeks to extend prior research on students' sense of belonging by validating the PSSM within an online learning platform and examining how belonging and self-affirmations interact in the digital age to effect students' academic performance.

8.1.5 The Present Study

The theoretical framework underpinning this work, as depicted in Figure 8.1, is an extension of the belonging cycle presented by Aguilar, Walton, and Wieman (2014). This framework postulates the effect of a value affirmation on students' performance will differ based on their baseline belonging and prior knowledge levels, as well as by the timing of the intervention (Schwartz, Tsang, & Blair, 2016). As such, variables for baseline belonging, prior knowledge, and delay of treatment or control are used in the present analysis as independent variables, exploring interaction effects using a three-way independent factors ANOVA. This approach was chosen because all independent variables are treated as categorical and interpretation is somewhat simpler than the output of a complex moderation analysis.

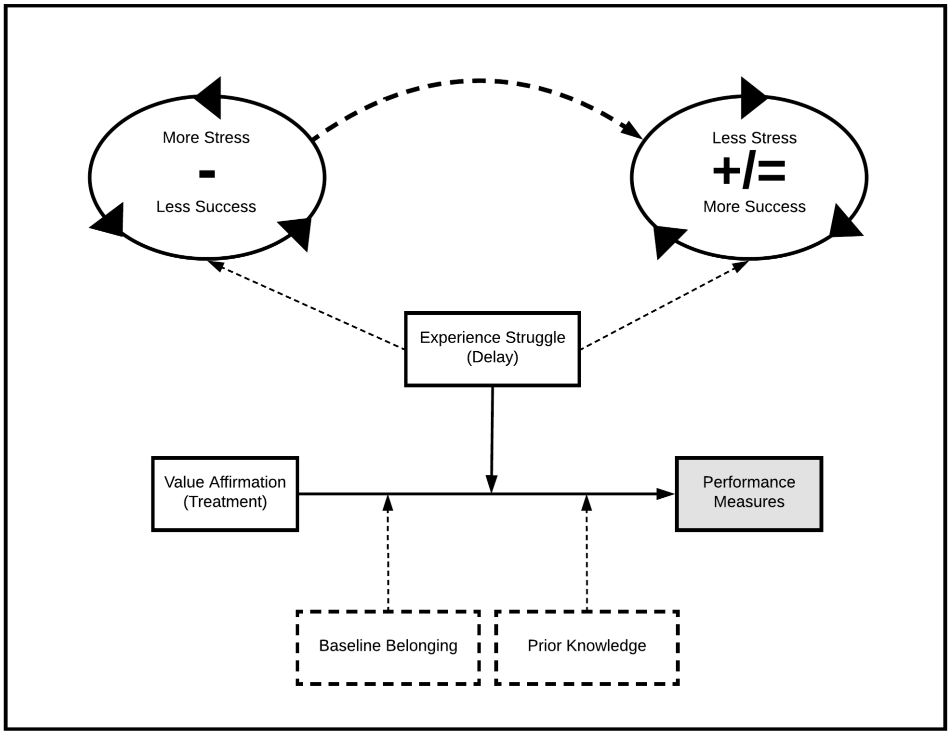


Figure 8.1. Theoretical framework adapted from Aguilar, Walton, and Wieman's (2014) Belonging Cycle. Variables with solid outlines are manipulated in the present study, those with dotted outlines are data collected before randomization, and the shaded box represents dependent variables including posttest score and within assignment metrics.

Based on this framework, students may internalize their performance as positive or negative, feeding the respective cycle through their effort, emotional response, and success or failure (Schwartz, Tsang, & Blair, 2016). Within the present work, a brief value affirmation prompt is intended to break negative belonging cycles, promoting a learning environment with less stress and more success. The hypothesized mechanism for the effect of treatment (reducing stress and increasing success) is likely complex in nature, linking to students' perceptions of their competence and autonomy, and their reactions to the system's feedback and tutoring (delivered as correctness feedback and progressively detailed hints). For students that negatively interpret their performance (-) the affirmation intervention is intended to provide an armor of sorts meant to help transform their interpretation of their performance into a more positive or neutral cycle (+/=). For students that begin with a positive or neutral interpretation of their performance (+/=), the intervention may bootstrap this effect, or may alternatively have a null effect, but is not expected to cause negative consequences (-). The hope is that this transformative process will alter students' academic performance within and immediately following their assignment. It is hypothesized that measures of student performance will be enhanced by the value affirmation intervention, especially for students whose original state reflected a deficit in baseline belonging.

The present work poses three research foci to investigate aspects of student belonging and self-affirmation within a mathematics assignment delivered using an online learning platform. These goals include:

Research Focus 1. Does a common measure of students' sense of belonging (the PSSM) produce valid and reliable results when delivered through an online learning platform? That is, do baseline measures of sense of belonging gathered within an online learning platform align with findings from past implementations of the scale?

Research Focus 2. Does a self-affirmation intervention intended to boost belonging alter student performance within a single homework assignment? Does the effect of the intervention interact with students' baseline measures of belonging and their prior knowledge? Do students perform differently within their assignments as measured by hint count, attempt count, and overall problem count?

Research Focus 3. Expand on the initial analysis of Goal 2 to understand whether the timing of the self-affirmation intervention alters its effect on posttest scores. That is, are differences observed when the intervention is provided immediately vs. delayed (i.e., in response to struggle)?

8.2 Methods

8.2.1 Learning Environment

The present work was conducted within ASSISTments (www.ASSISTments.org), an online learning platform used by hundreds of teachers assigning mathematics classwork and homework to over 50,000 student users. The platform has over 25,000 problems mapped to the Common Core State Standards (NGACBP & CCSSO, 2010) that are delivered as certified material and offered as a free service of WPI. The goal of platform is to provide students with instructional *ASSISTance* while serving as a powerful assess*MENT* tool for teachers, thus establishing the moniker "ASSISTments" (Heffernan & Heffernan, 2014). Use of ASSISTments has been shown to reliably improve students' learning in comparison to traditional paper and pencil approaches (Mendicino, Razzaq, Heffernan, 2009; Koedinger, McLaughlin & Heffernan, 2010; Singh et al., 2011; Kelly, Heffernan, Heffernan, et al., 2013; Soffer, et al., 2014). Most recently, SRI International showed that use of the platform for homework over the course of a single school year resulted in large, reliable learning gains on standardized assessments (Roschelle, et al., 2016). Given its success and growth, ASSISTments offers a unique environment in which to examine how sense of belonging might moderate students' academic performance.

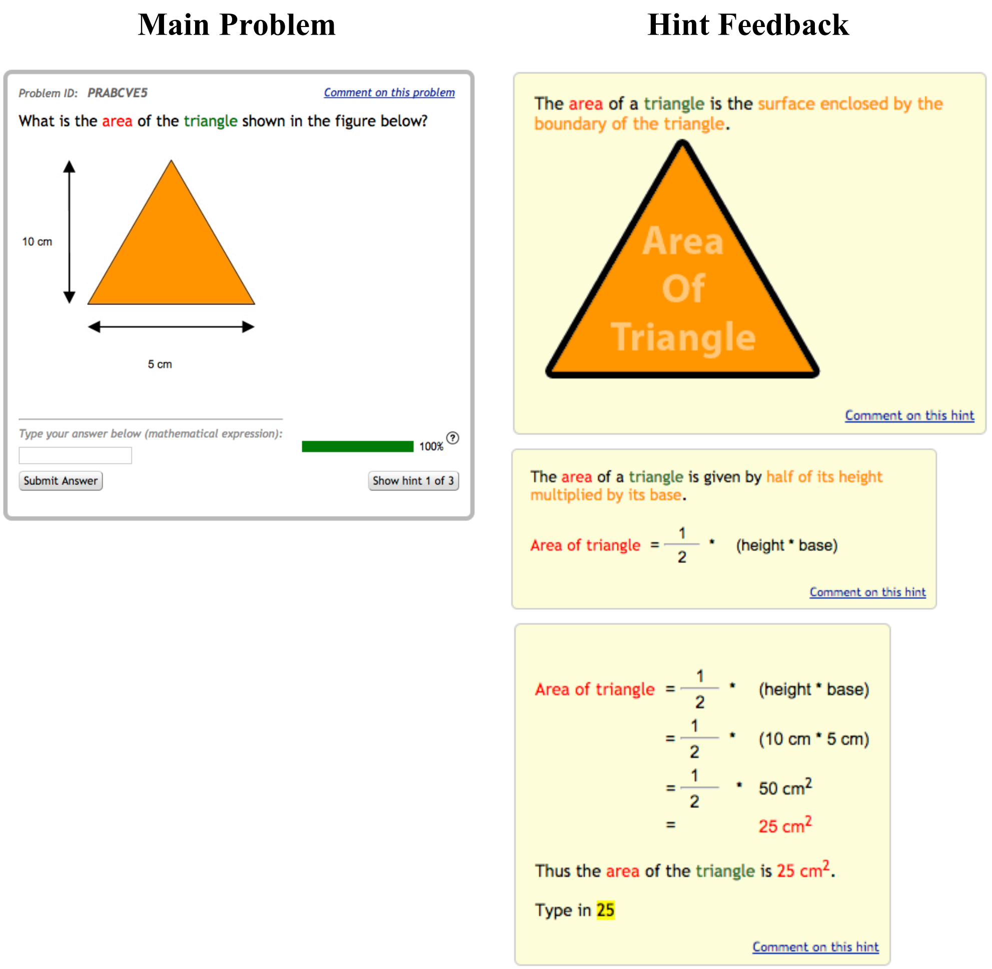
8.2.2 Participants

Participants were collected using ASSISTments by embedding survey content and a randomized controlled affirmation intervention into an ASSISTments Certified Skill Builder. Skill Builders are mastery-based learning assignments tailored to specific skills that students complete as classwork or homework. Embedding research within Skill Builders allows researchers to collect data from authentic learning environments in a manner that is minimally invasive to learning (www.ASSISTmentsTestBed.org). As teachers assign Skill Builders to their students naturalistically, in line with their existing curricula, sample populations grow steadily over time. Using this approach, a sample of 378 students was collected between March 2017 and March 2018. The sample was representative of 13 teachers, from 11 schools across 4 states. Two of these teachers were listed as teaching 6th grade, five were listed as teaching 7th grade, three were listed as teaching 8th grade, one was listed as teaching 10th grade, and two did not list their class grade levels. Based on ASSISTments' name-based query, the sample was 42% female, 48% male, and 10% unidentified.

8.2.3 Procedures

8.2.3.1 Math Content

The present study was embedded within an ASSISTments Certified Skill Builder with mathematics content mapped to the Common Core State Standard (NGACBP & CCSSO, 2010) of "Finding the Area of a Triangle or a Quadrilateral," a sixth grade Geometry skill (6.G.A.1). This skill was chosen because of its difficulty and its versatility in problem types. Problems within the assignment covered finding the area of trapezoids, triangles, rectangles, parallelograms, and squares. Problems were delivered using a consistent pattern to interleave alternating area formulas and thereby increase the desirable difficulty of the assignment, requiring that students routinely practice solution recall rather than developing rote strategies (Rohrer, 2012; Bjork, 1994). Examples of a triangle problem and concurrent tutoring feedback are presented in Figure 8.2.



Problem ID: PRABCVE5

What is the area of the triangle shown in the figure below?

Type your answer below (mathematical expression):

The area of a triangle is the surface enclosed by the boundary of the triangle.

The area of a triangle is given by half of its height multiplied by its base.

$$\text{Area of triangle} = \frac{1}{2} \cdot (\text{height} \cdot \text{base})$$

$$\begin{aligned}\text{Area of triangle} &= \frac{1}{2} \cdot (\text{height} \cdot \text{base})\\ &= \frac{1}{2} \cdot (10\text{ cm} \cdot 5\text{ cm})\\ &= \frac{1}{2} \cdot 50\text{ cm}^2\\ &= 25\text{ cm}^2\end{aligned}$$

Thus the area of the triangle is 25 cm^2.

Type in 25

Figure 8.2. Examples of assignment problems and concurrent hint feedback.

8.2.3.2 Experimental Design

The experimental design employed in the present study is depicted in Figure 8.3. At the start of the assignment students were notified that their assignment would include survey questions intended to help improve their experience in ASSISTments and were able to opt-out of this extra content. Those that opted-out were provided a traditional Skill Builder and their data was excluded from the present analysis. Students that chose to participate began their assignment by completing the Psychological Sense of School Membership (PSSM) Scale. Following this survey, students were randomly assigned to one of four conditions: Immediate Control (IC), Delayed Control (DC), Immediate Treatment (IT), or Delayed Treatment (DT). Students assigned to a treatment condition received an abbreviated value affirmation prompt, while those assigned to a control condition received an abbreviated null affirmation prompt to control for time on task (Cohen, et al. 2009). Students assigned to an immediate condition received their prompt directly following the PSSM. Those assigned to a delayed condition only received their prompt after exhibiting struggle in their assignment, operationalized as not being able to complete their assignment (which requires that students reach 'mastery' by accurately solving three consecutive problems) within the first three problems. Following mastery of the traditional skill content, students were given a posttest comprised of four transfer problems of progressive difficulty. These problems required students to solve for the area of irregular figures, modifying and extending the area formulas practiced within the assignment, as shown in Figure 8.4. Posttest problems were not delivered with hint feedback; if students got stuck they were able to request the answer and move on.

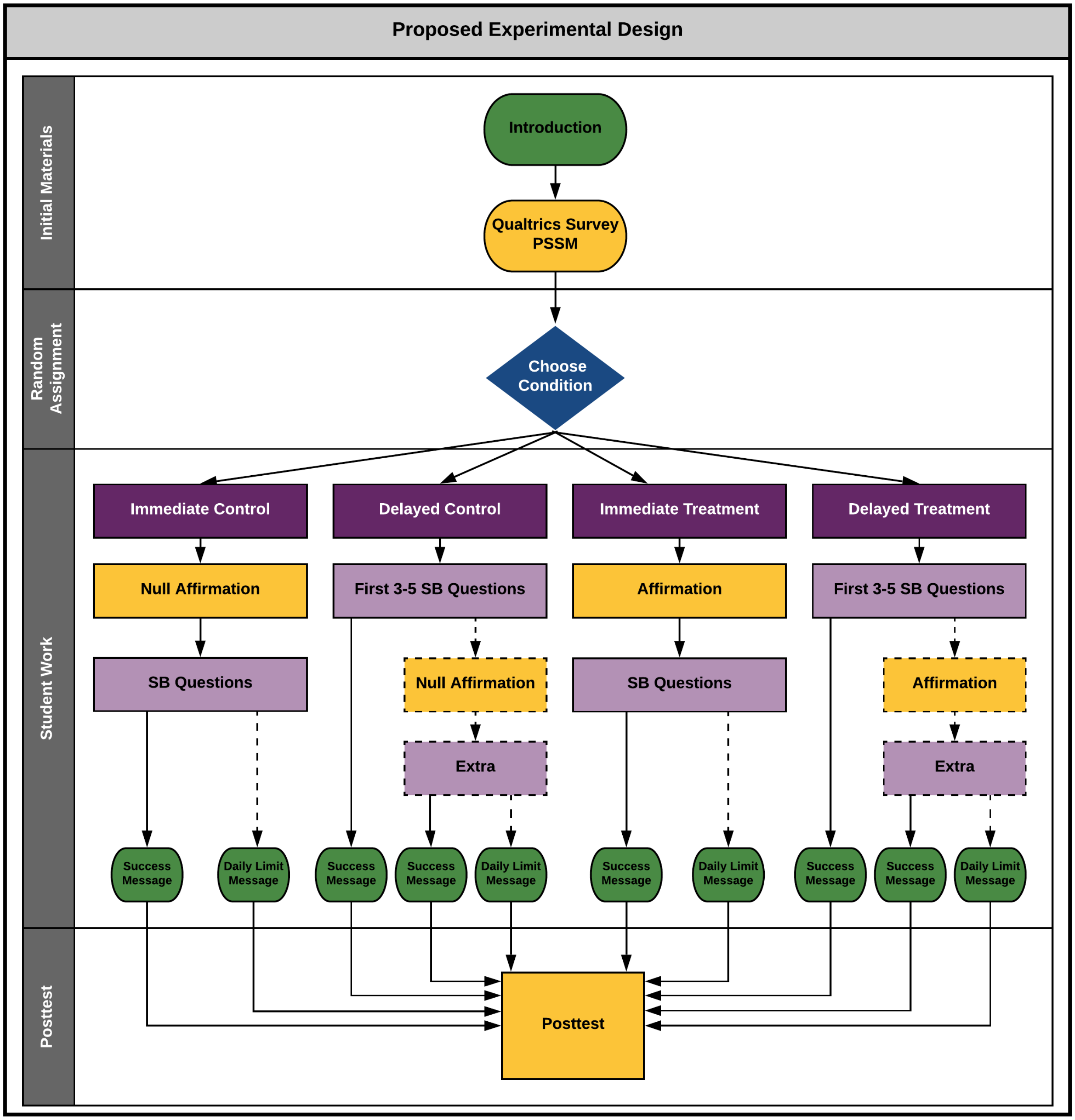
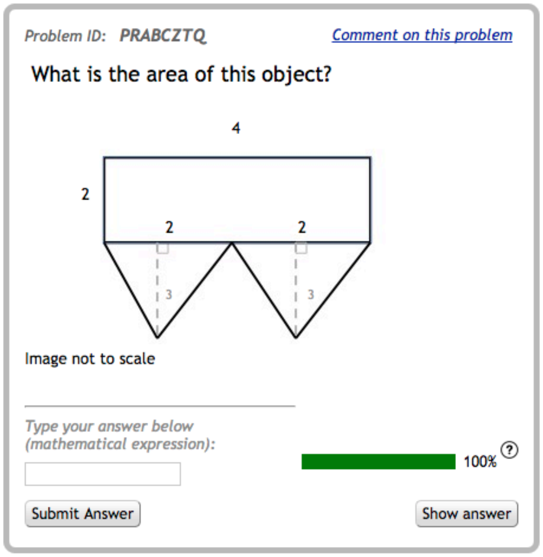


Figure 8.3. Experimental design.

Posttest Problem 1



Problem ID: PRABCZTQ

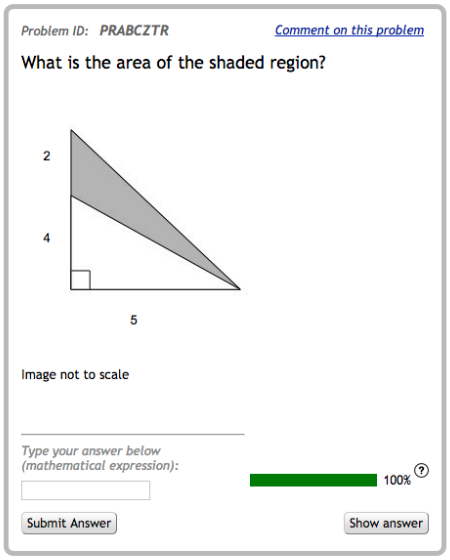
What is the area of this object?

Image not to scale

Type your answer below (mathematical expression):

100%

Posttest Problem 2



Problem ID: PRABCZTR

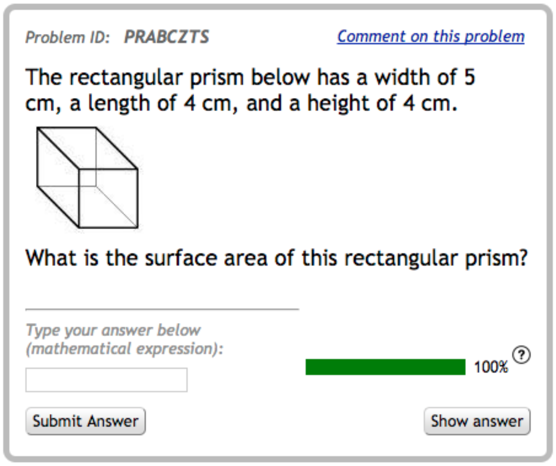
What is the area of the shaded region?

Image not to scale

Type your answer below (mathematical expression):

100%

Posttest Problem 3



Problem ID: PRABCZTS

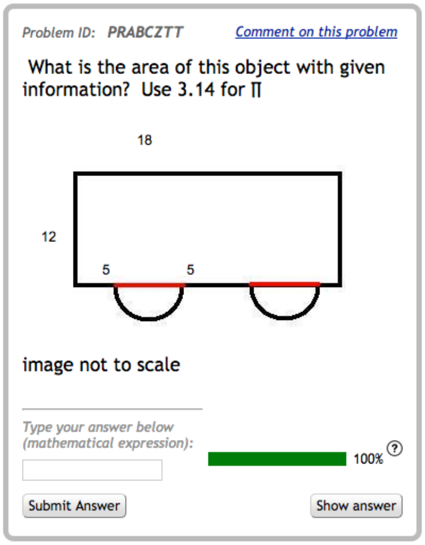
The rectangular prism below has a width of 5 cm, a length of 4 cm, and a height of 4 cm.

What is the surface area of this rectangular prism?

Type your answer below (mathematical expression):

100%

Posttest Problem 4



Problem ID: PRABCZTT

What is the area of this object with given information? Use 3.14 for π

image not to scale

Type your answer below (mathematical expression):

100%

Figure 8.4. Posttest problems and order of delivery.

8.2.4 Measures

8.2.4.1 Student Self-Report

Psychological Sense of School Membership Scale. In order to measure students' baseline sense of belonging, participating students began their assignment by completing the Psychological Sense of School Membership (PSSM) Scale. The full scale includes 18 Likert scaled items (1= "*Not at all true*" to 5 = "*Completely true*") that have been shown to have high internal consistency (α = 0.77 to 0.88) and the scale is well documented as a valid measurement tool amongst middle school students (grades 6-8) (Goodenow, 1993b). The scale is typically used by averaging Likert scores (after reverse scoring items 3, 6, 9, 12, and 16) across all 18 items to arrive at a single value. Higher values indicate a stronger sense of belonging. This overall score has been shown to carry high construct validity, with significant correlations to students' self-reported motivation, grades and GPA, and teacher-rated student effort (Goodenow, 1993b). Past work has also suggested that the items in this scale cluster into three constructs for multidimensional use, representing caring relationships (i.e., Item 4, "Other students in this school take my opinions seriously"), acceptance (i.e., Item 13, "I can really be myself at this school"), and rejection (i.e., Item 6, "Sometimes I feel as if I don't belong here.") (You, et al., 2011). Items from the PSSM are presented in Table 8.5 with their descriptive statistics.

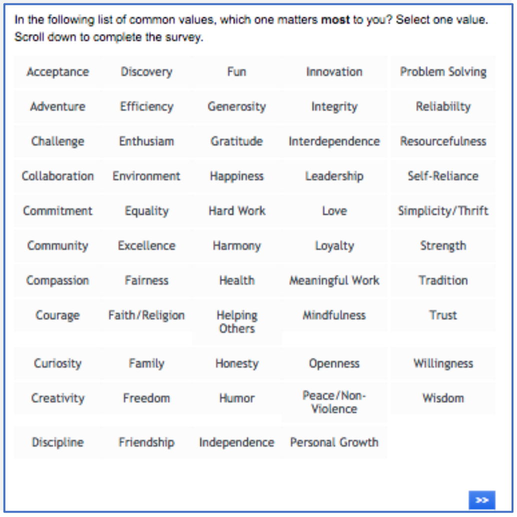
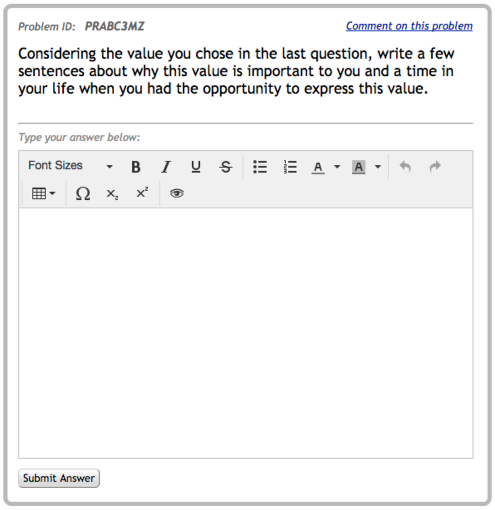
The present work offers the first examination (to our knowledge) of implementing the PSSM within an online learning environment. The scale was built and delivered using Qualtrics, a survey system that can be fully integrated within ASSISTments to collect more complex student data than the system typically captures. Students were asked to rate all 18 items but were able to skip items if desired. Survey data collected through Qualtrics was merged with students' ASSISTments logs via unique student identification numbers.

Affirmation and Null Affirmation Interventions. Students were randomly assigned to receive either a brief value affirmation writing prompt (Treatment conditions) or a brief null affirmation writing prompt (Control conditions). The delay of these prompts also varied by condition (Immediate vs. Delayed). These writing interventions were adapted from Cohen, et al. 2009. Original versions of the intervention suggested reliable improvements to students' academic performance with long-term effects after a 15-30 minute intervention, often delivered recursively (Cohen, et al. 2009). In the present work, the prompts were shortened such that students wrote only two to three sentences. All students were presented with a list of 54 common values (e.g., Courage, Loyalty, Wisdom, see Table 8.1 for a full list) and directions. Students in treatment conditions were asked to select the value that was most important to them and write two to three sentences regarding why the value they chose holds importance in their life. Students in control conditions were instead asked to select the value that was least important to them and to write two to three sentences regarding why the value they chose may be important for others. The null affirmation served to control for time on task, with the subtle shift in value ownership between conditions linked to potential differences in performance outcomes.

8.2.4.2 Student Performance Measures

Student performance measures are those metrics logged by ASSISTments that define students' mathematics abilities before, during, and after their math assignment. Each metric considered in the present work is described in detail below.

Table 8.1. Affirmation prompt and values.

Question 1	Question 2
	

In the following list of common values, which one matters **most** to you? Select one value. Scroll down to complete the survey.

Acceptance	Discovery	Fun	Innovation	Problem Solving
Adventure	Efficiency	Generosity	Integrity	Reliabiilty
Challenge	Enthusiam	Gratitude	Interdependence	Resourcefulness
Collaboration	Environment	Happiness	Leadership	Self-Reliance
Commitment	Equality	Hard Work	Love	Simplicity/Thrift
Community	Excellence	Harmony	Loyalty	Strength
Compassion	Fairness	Health	Meaningful Work	Tradition
Courage	Faith/Religion	Helping Others	Mindfulness	Trust
Curiosity	Family	Honesty	Openness	Willingness
Creativity	Freedom	Humor	Peace/Non-Violence	Wisdom
Discipline	Friendship	Independence	Personal Growth	

Problem ID: PRABC3MZ

Comment on this problem

Considering the value you chose in the last question, write a few sentences about why this value is important to you and a time in your life when you had the opportunity to express this value.

Type your answer below:

Prior Knowledge. All problem accuracy data from a student's work within ASSISTments prior to their participation in the present study was averaged into a single value meant to serve as a metric of their general math ability. For instance, a student who had accurately solved 75 out of 100 attempted problems within various ASSISTments assignments would have a prior knowledge score of 0.75, or 75%. This metric does not control for the amount of work a student has completed (i.e., another student accurately solving 3 out of 4 attempted problems would share the score of 0.75), or the difficulty levels of problems solved. However, the metric defines significantly different types of learners when the value is dichotomized on its median within a dataset; resulting groups of "low prior knowledge" and "high prior knowledge" students exhibit significantly different reactions to learning interventions (Razzaq & Heffernan, 2009; Ostrow, Heffernan, Heffernan, & Peterson, 2015). As such, this categorical measure of prior knowledge is used in the present work with the expectation that affirmation interventions will be more effective for students with lower prior knowledge. Further work is necessary to assess whether this measure significantly correlates with other measures that would define students as "at-risk."

Posttest Score. After finishing the Skill Builder assignment, all study participants were subjected to a four-item posttest. This test consisted of transfer items meant to measure students' ability to solve area problems of a higher difficulty level than those practiced within the assignment. Problems on finding the area of irregular figures were sourced from 7th and 8th grade ASSISTments Certified Skill Builders within the same Common Core strand (Geometry: NGACBP & CCSSO, 2010). Problems were presented using progressive difficulty and without hint feedback; students were allowed to make multiple attempts and to request the correct answer before moving on to the next problem. ASSISTments logged the accuracy of each student's first response on each problem and values were averaged to arrive at their overall posttest score (i.e., 0.50, or 50%). All posttest problems are presented in Figure 8.4.

Metrics within Assignment. ASSISTments also logged student's' performance as they completed the mathematics problems within their assignment. Considering these metrics within the immediate treatment and control (i.e., students that experienced the affirmation or null affirmation directly following the PSSM and before working on any math content), it is possible to determine whether the intervention significantly altered how students completed their assignment. Three metrics were considered: hint usage, attempt count, and problem count. These metrics are used as dependent variables within secondary analyses presented in the present work. Only the immediate conditions are considered when examining these metrics because the definition of struggle required of students reaching the delayed treatment or control would naturally influence these values.

“Hint Usage” reflects the total number of hints students requested while working through their math problems. Each problem in the assignment included two to three hints, delivered with progressive specificity, with the final hint providing students the correct answer. With regard to this measure, moderate hint counts likely reflect learning, while higher hint counts may reflect students’ amotivation, inattention, or attempts to ‘game’ their assignment.

“Attempt Count” reflects the total number of solution attempts students made while working through their math problems. Although students were marked incorrect if they were unable to accurately solve a problem on their first attempt, multiple attempts could then be used to arrive at the correct answer and allow them to move on to the next problem. These attempts were logged by the system. Higher attempt counts may reflect students’ persistence.

“Problem Count” reflects the total number of problems students solved while working through their assignment. As Skill Builders require students to accurately solve three consecutive problems on their first solution attempts, students may complete the assignment in as few as three problems. Moderate total problem counts likely reflect students’ struggle and persistence, with higher counts potentially reflecting amotivation.

8.2.5 Approach to Data Analysis

Data Preprocessing. Data collected in Qualtrics (PSSM scale items) was merged with ASSISTments mathematics data and descriptive statistics were considered for traditional assumptions of normality. Reverse scored items in the PSSM were recoded and the 18 items were averaged to form a single PSSM score. PSSM scores was then used to form a tri-cut distribution, with resulting labels of Low, Moderate, and High Belonging. Data from ASSISTments was processed to establish dependent measures including average posttest score and within assignment totals for hint count, attempt count, and problem count. In addition, the “prior percent correct” variable within ASSISTments output was dichotomized on its median to form High and Low Knowledge groups. Condition labels were determined using ASSISTments’ log files by transcribing the path of problems completed by each student. Correlations were examined for multicolinearity, and a chi-squared test was conducted to check for differential attrition across conditions prior to more robust analyses.

Research Focus 1. Internal reliability was recorded using Cronbach's alpha, and inter-item correlations were examined. Overall mean score was considered in comparison to past research to determine the reliability of this measure within an online learning environment. Belonging scores were then transformed into categorical labels of low, moderate, or high using three quantiles.

Research Focus 2. In order to determine whether an overall treatment effect exists for the value affirmation intervention, and to examine potential interaction effects between treatment, prior knowledge, and baseline belonging, a three-way independent factors ANOVA was performed with posttest score as the dependent measure. This analysis did not consider intervention delay. This approach was used because all independent variables were treated as categorical, and a multiple regression using moderation techniques would produce the same results with arguably more complex interpretation. Within-assignment metrics of hint count, attempt count, and problem count were also assessed using this approach within the subsample of students receiving the immediate treatment or control.

Research Focus 3. In order to determine whether the timing of the intervention altered its effect, posttest score will also be modeled with all four conditions while considering the interaction effects of students' baseline belonging and prior knowledge level.

8.3 Results

8.3.1 Dataset

Of the 378 students in the original sample, 263 students participated in the study, with 88 students opting-out of PSSM response and participating in the alternative assignment, 21 students opting-in but dropping out during the PSSM, and 6 students opening the assignment and quitting before taking any action. Among the 263 participants, 184 completed the PSSM and the full assignment, making up the analytic sample for posttest analyses. A chi-squared analysis was used to assess attrition across groups given the loss of approximately 30% of the sample. Results suggested that although attrition was high, it was not significantly different across groups, χ^2 (3) = 6.61, p = .086. Attrition rates are represented by completion percentage across groups in Table 8.2. A total of 241 students completed the PSSM but not necessarily the remainder of the assignment, creating an inflated analytic sample for assessment of survey reliability.

The resulting dataset compiled from ASSISTments and Qualtrics has nine primary features of interest, including five independent variables: 1) a categorical belonging variable constructed using a tri-cut of the raw distribution of students' initial belonging scores averaged from the 18 PSSM items (values between 1 and 5, tri-cut, categorized, and labeled as high, moderate, or low); 2) a categorical knowledge variable constructed by dichotomizing the raw distribution of students' prior knowledge scores on its median (values between 0.00 and 1.00, dichotomized to labels of high or low); 3) a dummy coded feature for the presence of a delay (0 = immediate, 1 = delay); 4) a dummy coded feature for the presence of the treatment (0 = control, 1 = treatment); and 5) a full condition variable representing all four conditions (DT = Delayed Treatment, IT = Immediate Treatment, DC = Delayed Control, and IC = Immediate Control). The dataset also includes four dependent variables: 1) posttest score (average accuracy across the four posttest questions ranging from 0.00 to 1.00); 2) within assignment hint count; 3) within assignment attempt count; and 4) within assignment problem count. Descriptive statistics of these variables are presented in Table 8.3. Further, correlation analysis revealed that none of the independent or dependent variables that would be modeled together were significantly correlated, relieving any worry of multicollinearity, as shown in Table 8.4. Significant correlations existed between Full Condition and its Delay and Treatment components (as expected), and between dependent variables of within assignment performance, all modeled separately. All data was merged in Microsoft Excel before being preprocessed and analyzed in IBM SPSS. Anonymized data has been made available at (Ostrow, 2018) for further reference.

8.3.2 Focus 1 - Baseline Belonging

The first goal of this work was to determine whether baseline measures of belonging as measured by the PSSM showed reliability with past implementations of the instrument. Item level analysis is presented in Table 8.5. Student recruitment rate exceeded reports in past implementations, with 71% of students opting to participate (a recruitment rate of 59% based on parental consent and

Table 8.2. Attrition rates across groups.

	Incomplete	Complete	**Total**	% Completion
Delayed Control	17	53	**70**	76
Delayed Treatment	19	60	**79**	76
Immediate Control	19	41	**60**	68
Immediate Treatment	23	31	**54**	57
Total	**78**	**185**	**263**	70

Table 8.3. Descriptive statistics for the analytic sample.

	Analytic Sample				
Variable	N	Mean	SD	Min	Max
Belonging (High = 3, Moderate = 2, Low = 1)	236	1.99	0.81	1.00	3.00
Knowledge (High = 1, Low = 0)	263	0.46	0.50	0.00	1.00
Delay (Delay = 1, Immediate = 0)	263	0.57	0.50	0.00	1.00
Treatment (Treatment = 1, Control = 0)	263	0.51	0.50	0.00	1.00
Full Condition (DC = 1, DT = 2, IC = 3, IT = 4)	263	2.37	1.09	1.00	4.00
Posttest Score	202	0.49	0.28	0.00	1.00
Hint Total	263	1.24	2.40	0.00	15.00
Attempt Total	263	5.83	5.31	0.00	34.00
Problem Count Total	263	3.74	2.32	0.00	17.00
Valid N (listwise)	184				

Table 8.4. Correlations among independent and dependent variables.

	1	2	3	4	5	6	7	8
1. Belonging (High = 3, Moderate = 2, Low = 1)								
2. Knowledge (High = 1, Low = 0)								
3. Delay (Delay = 1, Immediate = 0)								
4. Treatment (Treatment = 1, Control = 0)								
5. Full Condition (DC = 1, DT = 2, IC = 3, IT = 4)			**-.89**	**.41**				
6. Posttest Score	.15							
7. Hint Total	-.14					**-.32**		
8. Attempt Total						**-.23**	**.69**	
9. Problem Count Total						**-.19**	**.67**	**.84**

*Note***.** Bold correlations, p < .01. Suppressed correlations were not significant at p < .05.

Table 8.5. Item level analysis of the PSSM.

	Analytic Sample				
PSSM Question	N	Mean	SD	Min	Max
1. I feel like a real part of (name of school).	249	3.69	1.06	1.00	5.00
2. People here notice when I'm good at something.	250	3.58	1.13	1.00	5.00
3. It is hard for people like me to be accepted here. (Reversed)	249	3.94	1.24	1.00	5.00
4. Other students in this school take my opinions seriously.	250	3.30	1.10	1.00	5.00
5. Most teachers at (name of school) are interested in me.	249	3.66	1.06	1.00	5.00
6. Sometimes I feel as if I don't belong here. (Reversed)	247	3.99	1.28	1.00	5.00
7. There's at least one teacher or other adult in this school I can talk to if I have a problem.	248	3.91	1.34	1.00	5.00
8. People at this school are friendly to me.	247	4.02	0.92	1.00	5.00
9. Teachers here are not interested in people like me. (Reversed)	248	4.20	1.15	1.00	5.00
10. I am included in lots of activities at (name of school).	248	3.60	1.21	1.00	5.00
11. I am treated with as much respect as other students.	249	3.90	1.13	1.00	5.00
12. I feel very different from most other students here. (Reversed)	248	3.38	1.31	1.00	5.00
13. I can really be myself at this school.	249	3.64	1.16	1.00	5.00
14. The teachers here respect me.	250	4.17	0.96	1.00	5.00
15. People here know I can do good work.	249	4.08	0.94	1.00	5.00
16. I wish I were in a different school. (Reversed)	249	4.22	1.18	1.00	5.00
17. I feel proud of belonging to (name of school).	248	3.85	1.14	1.00	5.00
18. Other students here like me the way I am.	247	3.89	1.04	1.00	5.00
Valid N (listwise)	241				

student assent was observed by You et al. (2011), but only student assent is considered in the present work). Reliability analyses revealed that the items showed high internal consistency amongst 241 full responses (63.4% of the current sample), α = 0.90. This value adequately compared to Goodenow's (1993b) findings (α = 0.77 to 0.88, varying by sample type). Mean score for the present sample across items was 3.83, which was on par with Goodenow's (1993b) suburban samples (3.9 and 3.8). Further, correlation analyses (see Table 8.6) did not reveal any significant correlations higher than r = 0.70 within the current sample, similar to findings of past work with the scale in an unrelated sample (You, et al. 2011).

8.3.3 Focus 2 – Value Affirmation Main Effects

Posttest Score. Goal 2 was established to determine whether a self-affirmation intervention intended to boost belonging altered student performance at posttest following a single mathematics assignment. A three-way independent factors ANOVA was used to assess differences in posttest score by treatment condition while considering the effects of students' initial baseline measure of belonging and their prior knowledge. This analysis disregarded the timing of the intervention (immediate vs. delayed) in order to examine whether the intervention itself was linked to differences in students' posttest performance.

Results suggested that the main effect of treatment on the full sample was not significant, F(1, 172) = 1.48, p = .226. However, on a practical note, posttest scores were higher for students assigned to the treatment (M = .52, SD = .26, n = 92) than those assigned to the control (M = .48, SD = .27, n = 92). Post hoc power analyses revealed that the power of this effect was low, at only 23%. Given the complex interactions considered (resulting in 12 analytic groups), a sample of at least 787 students would have been required to observe a small effect size (*f* = .1) with a standard power of .8 and alpha value of .05.

The main effect of prior knowledge was not significant, F(1, 172) = .16, p = .694, but students' baseline measure of belonging had an effect trending toward significance, F(2, 172) = 2.72, p = .069, partial η^2 = .03. Students with moderate belonging slightly underperformed (M = .46, SD = .24) those with low belonging (M = .49, SD = .28) although the difference was not significant in pairwise comparisons. Those with moderate belonging performed significantly worse than those with high belonging (M = .56, SD = .26), p < .05.

The interaction effect of treatment and baseline belonging was not significant, F(2, 172) = .03, p = .971, nor was the interaction effect of prior knowledge and baseline belonging, F(2, 172) = .94, p = .391. However, the interaction effect of treatment and prior knowledge was significant, F(1, 172) = 4.83, p = .029, partial η^2 = .03. Simple effects tests revealed that low knowledge students performed significantly better at posttest after receiving the affirmation intervention (M = .55, SD = .25) than after receiving the control affirmation (M = .43, SD = .27), F(1, 172) = 6.18, p = .014, partial η^2 = .04.

High knowledge students were not significantly affected, although those receiving the control affirmation (M = 0.53, SD = .26) outperformed those receiving the treatment affirmation (M = .49, SD = .26). This finding could suggest that high knowledge students were distracted by the personal emphasis of the value affirmation, but not enough to cause significant harm to their performance. Low knowledge students performed better (M = .55, SD = .25) than their high knowledge peers (M = .49, SD = .26) when assigned to the treatment group, but this difference was not statistically significant, F(1, 172) = 1.61, p = .206. Still, a treatment that raises the performance of low knowledge students to at or above that of their high knowledge counterparts is of considerable practical importance. High knowledge students performed better (M = .53, SD = .26) than their

Table 8.6. Correlations of PSSM items.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. I feel like a real part of my school.																	
2. People here notice when I'm good at something.	**.57**																
3. It is hard for people like me to be accepted here (R)	**.25**	**.25**															
4. Other students in this school take my opinions seriously	**.43**	**.43**	**.22**														
5. Most teachers at my school are interested in me	**.45**	**.38**	**.21**	**.37**													
6. Sometimes I feel as if I don't belong here (R)	**.39**	**.30**	**.56**	**.25**	**.30**												
7. There's at least one teacher or other adult in this school I can talk to if I have a problem	**.29**	**.22**	**.17**	**.20**	**.34**	**.20**											
8. People at this school are friendly to me	**.44**	**.36**	**.37**	**.36**	**.31**	**.36**	**.28**										
9. Teachers here are not interested in people like me (R)	**.27**		**.24**		**.21**	**.25**	**.22**	.13									
10. I am included in lots of activities at my school	**.44**	**.41**	**.22**	**.34**	**.31**	**.35**	**.18**	**.45**									
11. I am treated with as much respect as other students	**.44**	**.39**	**.31**	**.38**	**.41**	**.39**	**.38**	**.47**	**.23**	**.35**							
12. I feel very different from most other students here (R)	**.24**	**.23**	**.42**	**.23**	**.19**	**.49**		**.31**	**.24**	**.28**	**.26**						
13. I can really be myself at this school	**.56**	**.50**	**.34**	**.39**	**.44**	**.47**	**.37**	**.55**	**.22**	**.45**	**.52**	**.27**					
14. The teachers here respect me	**.46**	**.31**	**.23**	**.28**	**.56**	**.28**	**.49**	**.40**	**.35**	**.26**	**.54**	.15	**.52**				
15. People here know I can do good work	**.48**	**.40**	**.23**	**.32**	**.46**	**.27**	**.27**	**.41**	**.19**	**.32**	**.44**	**.23**	**.41**	**.47**			
16. I wish I were in a different school (R)	**.38**	.14	**.37**	**.17**	**.28**	**.53**	**.24**	**.33**	**.31**	**.27**	**.26**	**.34**	**.36**	**.33**	**.27**		
17. I feel proud of belonging to my school	**.52**	**.34**	**.28**	**.29**	**.47**	**.41**	**.35**	**.45**	**.21**	**.42**	**.45**	**.26**	**.54**	**.50**	**.42**	**.47**	
18. Other students here like me the way I am	**.50**	**.46**	**.33**	**.46**	**.35**	**.40**	**.28**	**.58**	**.17**	**.40**	**.47**	**.25**	**.53**	**.42**	**.53**	**.32**	**.47**

*Note***.** Bold correlations, $p < .01$; all others, $p < .05$. Suppressed correlations were not significant at $p < .05$.

low knowledge peers (M = .43, SD = .27) when assigned to the control condition, nearing statistical significance F(1, 172) = 3.39, p = .067. These effects are depicted visually by the interaction in Figure 8.5.

The three way interaction between Treatment X Prior Knowledge X Belonging was not significant, F(2, 172) = .26, p = .774. Overall, the model only explained 7.7% of the variance in posttest score, suggesting that the variables leave something to be desired with regard to prediction, $R^2 = .077$, Adjusted $R^2 = .018$. ANOVA findings and power are summarized in Table 8.7, while means and SDs across treatment, knowledge level, and belonging level are presented in Table 8.8 and visualized in Figure 8.6.

Variables Within Assignment. With respect to performance differences within students' assignments, three separate ANOVA analyses were conducted considering only students who received immediate treatment or control to assess differences in hint usage, attempt count, and problem count by treatment condition. Immediate conditions were selected for analysis due to fidelity, and natural variation

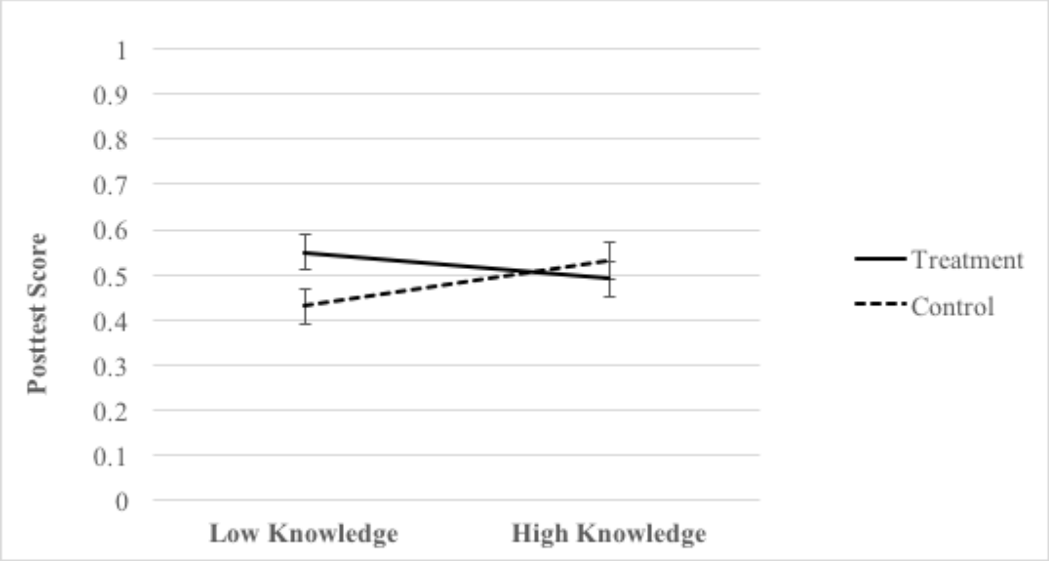


Figure 8.5. Posttest scores by treatment condition and prior knowledge level.

Table 8.7. Three-way independent factors ANOVA results of treatment, prior knowledge level, and baseline belonging on posttest score.

	SS	df	MS	F	p	1-β
Treatment	0.10	1	0.10	1.48	.226	.23
Knowledge	0.01	1	0.01	0.16	.694	.07
Belonging	0.37	2	0.19	2.72	.069	.54
Treatment X Know	0.33	1	0.33	4.83	.029	.60
Treatment X Belong	0.01	2	0.02	0.03	.971	.05
Know X Belong	0.13	2	0.07	0.94	.391	.21
Treatment X Know X Belong	0.04	2	0.02	0.26	.774	.09
Error	11.82	172	0.07			

Table 8.8. Means (SDs) of posttest score by treatment, knowledge level, and baseline belonging.

	Treatment (n = 92)	**Control (n = 92)**	**Total (n = 184)**
Low Knowledge *(n = 97)*	.55 (.25)	.43 (.27)	.49 (.27)
LB *(n = 27)*	.59 (.27)	.44 (.25)	.52 (.27)
MB *(n = 37)*	.50 (.20)	.36 (.28)	.44 (.25)
HB *(n = 33)*	.58 (.31)	.48 (.29)	.53 (.30)
High Knowledge *(n = 87)*	.49 (.26)	.53 (.26)	.51 (.26)
LB *(n = 25)*	.41 (.32)	.49 (.29)	.45 (.30)
MB *(n = 31)*	.46 (.24)	.50 (.27)	.48 (.25)
HB *(n = 31)*	.60 (.22)	.59 (.23)	.59 (.22)
Total (n = 184)	.52 (.26)	.48 (.27)	.50 (.26)

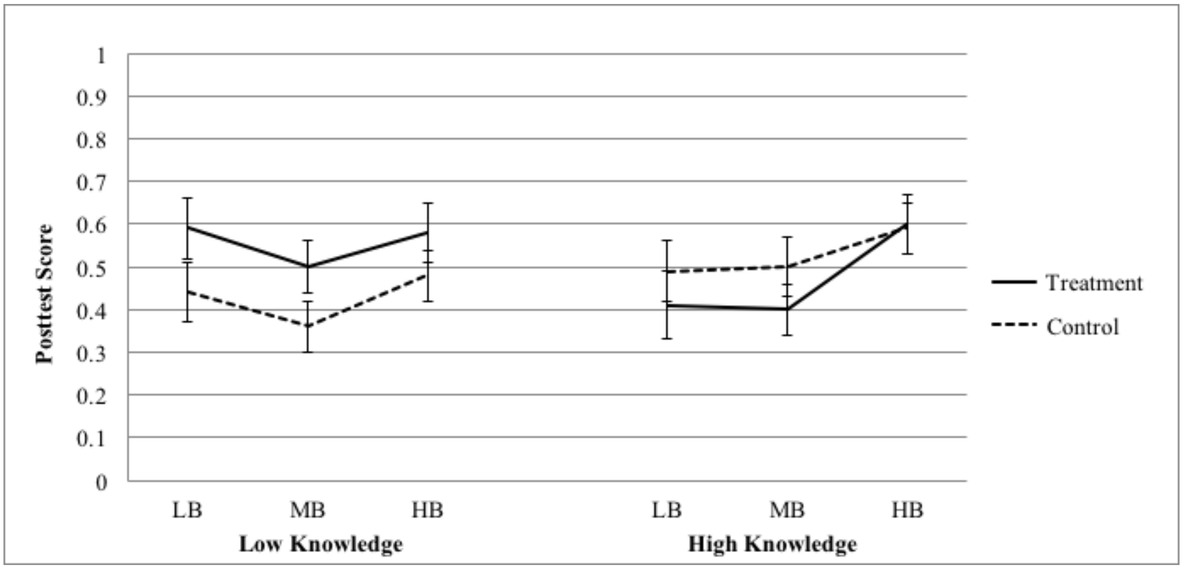


Figure 8.6. Posttest score by baseline belonging, prior knowledge level, and treatment.

within assignment that would lead to inclusion in the delayed condition (higher problem counts and greater attempts). Group sizes at the finest granularity were a substantial limitation to these analyses (M = 8.67, SD = 2.02, Median = 9, Range = [4, 12]). A larger sample would have provided greater power. Still, respective analyses are presented below.

Hint Usage. When considering hint usage during assignment as the dependent variable, none of the main effects or interactions showed significance. Post hoc power analyses were conducted to show the relatively low power of observed effects, with results included in Table 8.9. Given the complex interactions considered that established 12 comparison groups in the overall model, a sample of at least 788 students would have been required to observe small effect sizes ($f = .1$) with a standard power of .8 and alpha value of .05, an increase in sample size by at least 684 students.

In the present analysis, Levene's test was significant, suggesting that variance in hint usage was different across groups, F(11,92) = 3.38, p < .01, likely due to unbalanced and small group sizes. Although the preregistered analysis protocol is followed herein, with such small samples it would likely be more appropriate to consider non-parametric options. The main effect of treatment was not significant, F(1, 92) = 0.72, p = .397, nor were the main effects of knowledge level, F(1, 92) = 0.73, p = .396, or baseline belonging, F(2, 92) = 1.64, p = .200. None of the two-way or three-way interactions showed statistical significance, as summarized in Table 8.9.

Attempt Count. Results were similarly low powered when considering attempt count during assignment as the dependent variable. Levene's test was again significant, suggesting that variance in attempt count was different across groups, F(11,92) = 4.32, p < .001. Non-parametric options may have provided a stronger alternative than the preregistered protocol reported herein.

The main effect of treatment was not significant, F(1, 92) = .01, p = .941, nor was the main effect of knowledge level, F(1, 92) = .01, p = .954. A significant main effect of baseline belonging was

Table 8.9. Three-way independent factors ANOVA results of treatment, prior knowledge level, and baseline belonging on hint usage.

	SS	df	MS	F	p	1-β
Treatment	5.62	1	5.62	0.72	.397	.13
Knowledge	5.64	1	5.64	0.73	.396	.13
Belonging	25.45	2	12.73	1.64	.200	.33
Treatment X Know	13.00	1	13.00	1.67	.199	.24
Treatment X Belong	32.47	2	16.24	2.09	.130	.41
Know X Belong	25.29	2	12.65	1.63	.202	.33
Treatment X Know X Belong	4.09	2	2.04	0.26	.769	.09
Error	715.10	92	7.77			

observed, F(2, 92) = 3.21, p = .045, partial η^2 = .03. Contrasts revealed that a low level of belonging resulted in significantly higher attempt counts (M = 7.32, SD = 8.17) than moderate (M = 5.00, SD = 4.33) or high levels of belonging (M = 4.70, SD = 3.99), p = .013. The difference in attempt count between moderate and high levels of belonging was not significantly different, p = .880.

None of the two-way interactions showed statistical significance, as summarized in Table 8.10. However, the three-way interaction between treatment, knowledge level, and baseline belonging was significant, F(2, 92) = 4.28, p = .017, partial η^2 = .01. Post hoc power analysis revealed that this finding was relatively well powered, 1-β = .75. Simple effects tests revealed a significant difference in attempt counts between low knowledge students with low belonging receiving the treatment (M = 3.00, SD = 2.83) rather than the control (M = 9.67, SD = 10.00), F(1, 92) = 6.06, p = .016. In addition, students with low baseline belonging who were randomly assigned to the treatment condition used significantly more attempts when they were considered high knowledge (M = 12.50, SD = 14.89) rather than low knowledge (M = 3.00, SD = 2.83), F(1,92) = 7.74, p = .007. Finally, high knowledge students in the treatment group showed significant differences in attempt count across levels of belonging, F(2, 92) = 4.12, p = .019. Those with low belonging used more attempts on average (M = 12.50, SD = 14.89) than those with moderate (M = 3.67, SD = 2.40) or high belonging (M = 3.43, SD = 0.79). Means and SDs across treatment, knowledge level, and belonging level are presented in Table 8.11 and visualized in Figure 8.7.

Problem Count. When considering problem count during assignment as the dependent variable, Levene's test was again significant, suggesting that variance was different across groups, F(11,92) = 3.38, p = .001. Non-parametric options may again have provided a stronger alternative than the preregistered protocol reported herein.

The main effect of treatment was not significant, F(1, 92) = 0.04, p = .852, nor were the main effects of knowledge level, F(1, 92) = 1.21, p = .274, or baseline belonging, F(2, 92) = .74, p = .480. Further, none of the two-way interactions showed statistical significance, as summarized in Table 8.12.

Table 8.10. Three-way independent factors ANOVA results of treatment, prior knowledge level, and baseline belonging on attempt count.

	SS	df	MS	F	p	1-β
Treatment	0.17	1	0.17	0.01	.941	.05
Knowledge	0.10	1	0.10	0.01	.954	.05
Belonging	199.27	2	99.64	3.21	.045	.61
Treatment X Know	50.05	1	50.05	1.61	.208	.24
Treatment X Belong	8.26	2	4.13	0.13	.876	.07
Know X Belong	112.25	2	56.13	1.81	.170	.37
Treatment X Know X Belong	266.21	2	133.11	4.28	.017	.75
Error	2859.29	92	31.08			

Table 8.11. Means (SDs) of attempt count by treatment, knowledge level, and baseline belonging.

	Immediate Treatment (n = 49)	**Immediate Control (n = 55)**	**Total (n = 104)**
Low Knowledge *(n = 55)*	5.28 (5.81)	6.58 (6.86)	5.89 (6.30)
LB *(n = 17)*	3.00 (2.83)	9.67 (10.00)	6.53 (8.08)
MB *(n = 22)*	6.75 (6.52)	4.70 (3.97)	5.82 (5.49)
HB *(n = 16)*	5.33 (6.65)	5.29 (4.11)	5.31 (5.51)
High Knowledge *(n = 49)*	5.35 (7.15)	5.21 (3.00)	5.27 (5.05)
LB *(n = 14)*	12.50 (14.89)	6.60 (4.38)	8.29 (8.49)
MB *(n = 18)*	3.67 (2.40)	4.33 (1.50)	4.00 (1.97)
HB *(n = 17)*	3.43 (.79)	4.60 (1.90)	4.12 (1.62)
Total (n = 104)	5.31 (6.32)	5.85 (5.19)	5.60 (5.73)

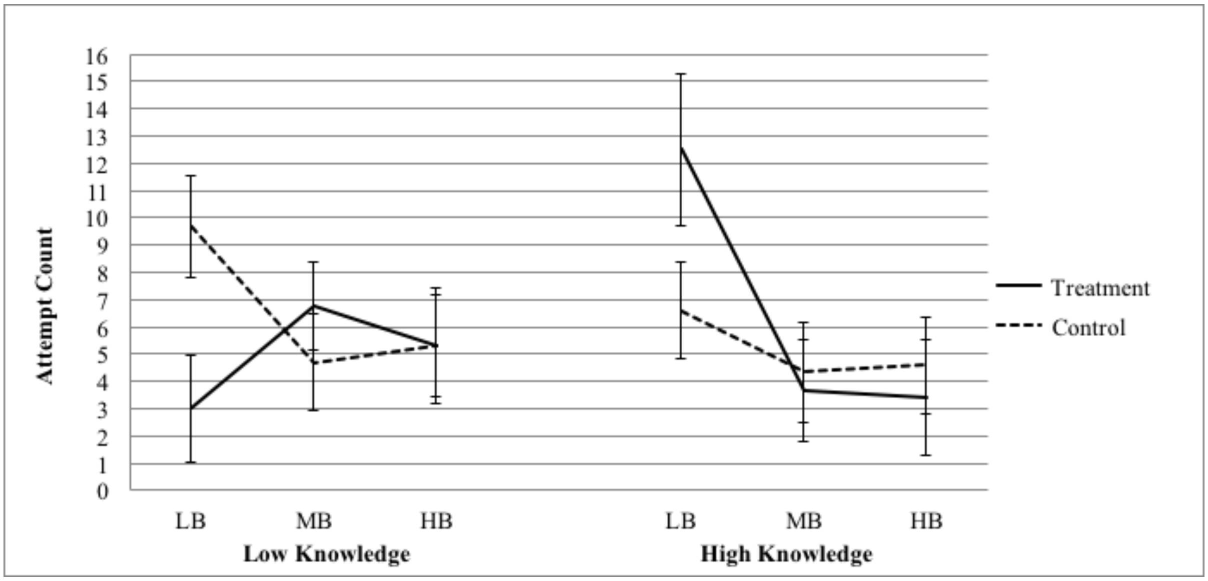


Figure 8.7. Attempt count by baseline belonging, prior knowledge level, and treatment.

Table 8.12. Three-way independent factors ANOVA results of treatment, prior knowledge level, and baseline belonging on problem count.

	SS	df	MS	F	p	1-β
Treatment	0.23	1	0.23	0.04	.852	.05
Knowledge	8.03	1	8.03	1.21	.274	.20
Belonging	9.82	2	4.91	0.74	.480	.17
Treatment X Know	2.76	1	2.76	0.42	.521	.10
Treatment X Belong	2.79	2	1.40	0.21	.811	.08
Know X Belong	28.66	2	14.33	2.16	.121	.45
Treatment X Know X Belong	44.36	2	22.18	3.34	.040	.65
Error	610.32	92	6.63			

Table 8.13. Means (SDs) of problem count by treatment, prior knowledge, and baseline belonging.

	Immediate Treatment (n = 49)	**Immediate Control (n = 55)**	**Total (n = 104)**
Low Knowledge *(n = 55)*	3.34 (3.17)	3.62 (2.48)	3.47 (2.84)
LB *(n = 17)*	1.75 (1.75)	4.33 (3.24)	3.12 (2.89)
MB *(n = 22)*	4.25 (3.72)	3.10 (2.33)	3.73 (3.15)
HB *(n = 16)*	3.56 (3.09)	3.43 (1.51)	3.50 (2.45)
High Knowledge *(n = 49)*	3.70 (2.56)	3.86 (2.12)	3.80 (2.28)
LB *(n = 14)*	6.25 (4.43)	4.20 (3.55)	4.79 (3.77)
MB *(n = 18)*	3.33 (1.80)	3.78 (.97)	3.56 (1.42)
HB *(n = 17)*	2.71 (.95)	3.60 (.52)	3.24 (.83)
Total (n = 104)	3.49 (2.91)	3.75 (2.28)	3.62 (2.59)

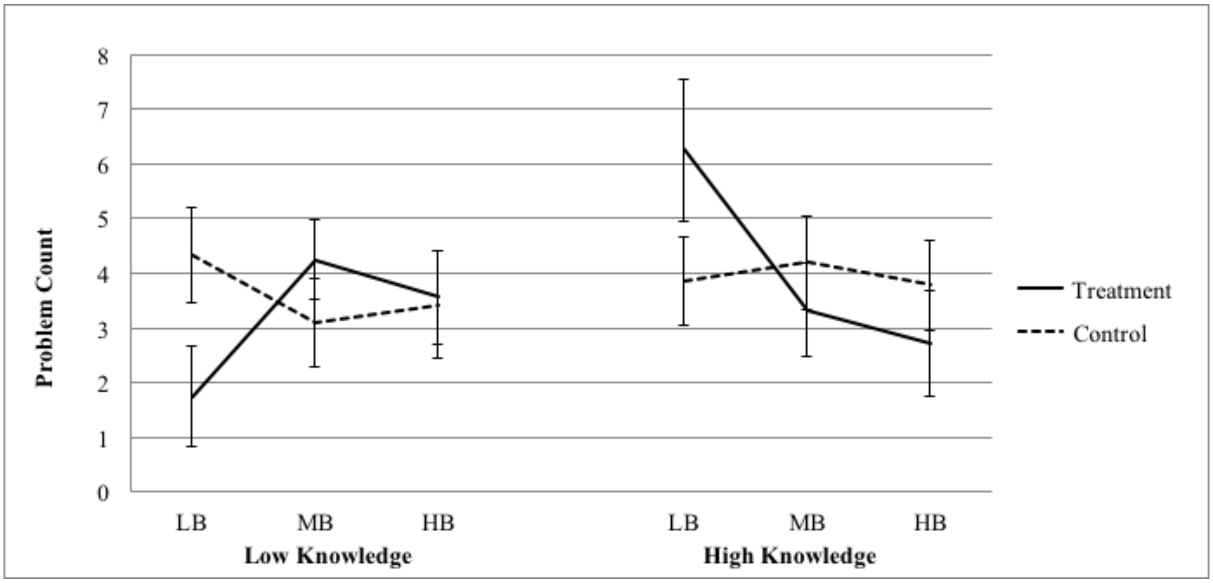


Figure 8.8. Problem count by baseline belonging, prior knowledge level, and treatment.

The three-way interaction between treatment, knowledge level, and baseline belonging was significant, F(2, 92) = 3.34, p = .040, partial $\eta^2 = .07$. Post hoc power analysis revealed that this finding achieved a power of 65%, 1-β = .65. Simple effects tests revealed a significant difference in problem counts between low knowledge students with low belonging receiving the treatment (M = 1.75, SD = 1.75) rather than the control (M = 4.33, SD = 3.24), F(1, 92) = 4.26, p = .042. In addition, students with low baseline belonging who were randomly assigned to the treatment condition saw significantly more problems when they were considered high knowledge (M = 6.25, SD = 4.43) rather than low knowledge (M = 1.75, SD = 1.75), F(1,92) = 8.14, p = .005. Means and SDs across treatment, knowledge level, and belonging level are presented in Table 8.13 and visualized in Figure 8.8.

8.3.4 Focus 3 – Intervention Effects by Timing

Turning back to the full dataset, posttest performance was also analyzed with regard to the timing of the intervention by examining all four experimental conditions (Delay X Treatment) to assess Focus 3 and to determine whether the intervention was more effective in the face of challenge. Fidelity of delayed implementation was critically low based on the definition of 'struggle' required for students to reach the value or control affirmations and prompts. Only 10 students assigned to the delayed treatment group actually saw the value affirmation (13% of the initial sample), with only 8 reaching the posttest and being assessed in the analytic sample (12% of the analytic sample). Similarly, only 12 students assigned to the delayed control group actually saw the control affirmation (17% of the initial sample), with only 11 reaching the posttest and being assessed in the analytic sample (19% of the analytic sample). This likely poses a considerable limitation to any analysis considering the effect of delay due to a lack of power.

Despite this potential issue, Levene's test was not significant, suggesting that variance in posttest was roughly equal across groups, F(23, 160) = 1.01, p = .457. The main effect of condition was not significant, F(3, 160) = 1.975, p = .120, nor was the main effect of knowledge level, F(1, 160) = 0.00, p = .986. However, the main effect of baseline belonging was significant,

Table 8.14. Three-way independent factors ANOVA results of condition, prior knowledge level, and baseline belonging on posttest score.

	SS	df	MS	F	p	1-β
Condition	0.40	3	0.13	1.98	.120	.50
Knowledge	0.00	1	0.00	0.00	.986	.05
Belonging	0.45	2	0.22	3.34	.038	.62
Cond X Know	0.38	3	0.13	1.87	.137	.47
Cond X Belong	0.64	6	0.11	1.59	.153	.61
Know X Belong	0.15	2	0.07	1.10	.335	.23
Cond X Know X Belong	0.20	6	0.03	0.51	.802	.20
Error	10.69	160	0.07			

F(2, 160) = 3.34, p = .038, partial $\eta^2 = .04$. Contrasts revealed that students with a high level of belonging exhibited significantly higher posttest scores (M = .56, SD = .26) than those with moderate (M = .46, SD = .24) or low levels of belonging (M = .49, SD = .28), p = .024. The difference in posttest scores between moderate and low levels of belonging was not significantly different, p = .214. None of the two-way or three-way interactions showed statistical significance, as summarized in Table 8.14.

8.4 Discussion

The present work used three research foci to investigate how participation in a value affirmation intervention effected performance measures within and immediately following an online mathematics assignment when also considering levels of baseline belonging and prior knowledge.

The first research focus centered on the PSSM as a measure of students' sense of school belonging. The scale was found to be internally consistent, $\alpha = .90$, with reliability, mean scale score, and recruitment rates matching prior implementations of the scale in offline environments. As such, results suggested that this scale provides an excellent resource for collecting middle school students' sense of school belonging through ASSISTments and would likely generalize well to other online learning platforms. Belonging scores collected in these environments could serve as valuable covariates or independent variables in models of students' cognitive, behavioral, or socio-emotional learning outcomes. Future work should consider the validity of the PSSM within ASSISTments by correlating overall scale scores to students' self-reported motivation, grades and GPA, and teacher-rated student effort, which have all shown significant correlation to the scale in past work (Goodenow, 1993b).

One potential limitation to the findings presented herein was opt-in bias, in which more motivated or higher performing students may have exhibited a greater propensity for opting to answer the survey items than their less motivated or lower performing peers. A second potential limitation is the high proportion of attrition within the assignment. Although attrition was attributed more to the completion of the math portion of the assignment (24% of students who completed the PSSM failed to complete the assignment), scale attrition represented 15% of the original opt-in sample. This may pose an issue with the generalizability of analyses. Future iterations of this work with larger sample sizes should explore the depth of this issue by examining distributions of student characteristics amongst participants and non-participants to determine whether patterns of participation can be considered random. One approach to dealing with non-random participation would be to include non-responders as a separate condition when conducting within-assignment analyses in order to assess how their Skill Builder performance differed from participants that were assigned to the Immediate Treatment and Immediate Control conditions. In the present analysis, adding groups to within-assignment analyses would have reduced power even further, making findings even less concrete. Among students who opted-in, it is possible that some may not have wanted to provide the personal details of their perceived scholastic experience because they thought that their teachers would be able to access their responses. Although ASSISTments typically produces reports for teachers and students with detailed performance data, teachers were not able to access survey responses and all PSSM related data remained fully anonymized during the analysis process. Making it clear that their responses would remain anonymous might improve fidelity in future iterations of this work. Similarly, students may have felt that the questions were not appropriate for a math assignment and may have viewed them as a distraction. As such, a third potential limitation was scale length. A shorter scale would undoubtedly offer greater opportunity for data collection with less attrition. Past work has suggested the use of an abbreviated 11-item version of the scale that retains its reliability (You, et al. 2011), which could be implemented in future iterations of this work.

The second research focus centered on the use of a value selection task and brief affirmation prompt as an intervention to improve student performance. This intervention, adapted from Cohen, et al. 2009, required students to select the most important (treatment) or least important (control) value from a list and explain why their choice was important to them (treatment) or to others (control). The subtle shift in value ownership is thought to drive the efficacy of the treatment (Cohen, et al. 2009).

Disregarding the timing of the intervention, results suggested that students in the treatment group outperformed those in the control by four points at posttest. Although this difference was not statistically significant, the improvement could be of practical significance for teachers hoping to raise assessment scores. When considering standardized tests, just a few points can mean the difference between a student showing proficiency or being flagged for extra remediation. Results also suggested that students' sense of school belonging influenced learning outcomes. Students with high baseline belonging outperformed those with moderate or low belonging (with trending significance), aligning with past work showing belonging as an effective area for intervention (Cohen & Garcia, Apfel, & Master, 2006; Cohen & Garcia, 2008; Walton & Cohen, 2011). Results suggested that the intervention may have interrupted the negative feedback loop (shown in Figure 8.1) and prepared students to tackle their assignment with greater perseverance (Cohen & Garcia, 2006; Cohen & Garcia, 2008). The significant interaction between treatment and prior knowledge was especially promising and aligned with past work that had shown the intervention was more powerful for high-risk students (Cohen & Garcia, 2006; Cohen & Garcia, 2008). Low knowledge students in the treatment condition significantly outperformed their high knowledge peers, while high knowledge students in the control condition outperformed their low knowledge peers (although this difference was not significant). It is possible that while value ownership benefited low knowledge students by boosting their sense of belonging or self-integrity, the intervention may have been distracting to high knowledge students. While limited, any finding that works to close the achievement gap suggests positive practical importance, and therefore should be explored further by future work.

Within-assignment findings were especially low powered. No significant differences were observed in within-assignment hint usage, and findings for within-assignment attempt count and problem count followed similar trends due to the high correlation between these two dependent variables (r = .84, p < .001). Low levels of baseline belonging were linked to significantly higher attempt counts but were not observed to have a significant main effect in problem count. This may suggest that students with lower levels of belonging struggled more within questions but were able to master their assignment with approximately the same efficiency as their moderate and high belonging peers. The interaction between treatment, knowledge level, and baseline belonging followed the same pattern for attempts and problem count, with significantly lower values observed in low knowledge/low belonging students receiving the treatment condition rather than the control condition. In contrast, within the treatment group, significantly higher values were observed in high knowledge/low belonging students than in low knowledge/low belonging students, suggesting that knowledge level may help to drive persistence.

The primary limitation of the second research focus was a lack of power. Group sizes for the within assignment analyses were substantially limiting (M = 8.67, SD = 2.02, Median = 9, Range = [4, 12]). Post hoc power analysis using G*Power 3.1 revealed that significant main effects and interactions were observed at only 60-75% power, where 80% would be ideal. If the true effects were anticipated to be relatively small (*f* = .1), a much larger sample of 967 students would have been necessary to establish significance at β = .20 and α = .05. A smaller sample would be permissible if the true effects are larger, as shown in Figure 8.9. Moderate effects (*f* = .25) should be observable with a sample of approximately 158 students. Given that the analytic sample for the primary posttest analysis included 184 students, this suggests that power was lost due to random assignment and attrition leading to unequal groups in the present design. It may also suggest that true effect sizes are small or negligible, and that effects were underpowered due to low sample size. However, observed effect sizes were between small and moderate (partial η^2 = .03 – .09, translating to f = .17 – .29), so it is likely that despite not being well powered these findings were not false positives. When considering the within assignment analyses, total sample size dropped to 104 by considering only students receiving immediate treatment or control conditions, suggesting the sample was too small to effectively observe significant differences in student performance.

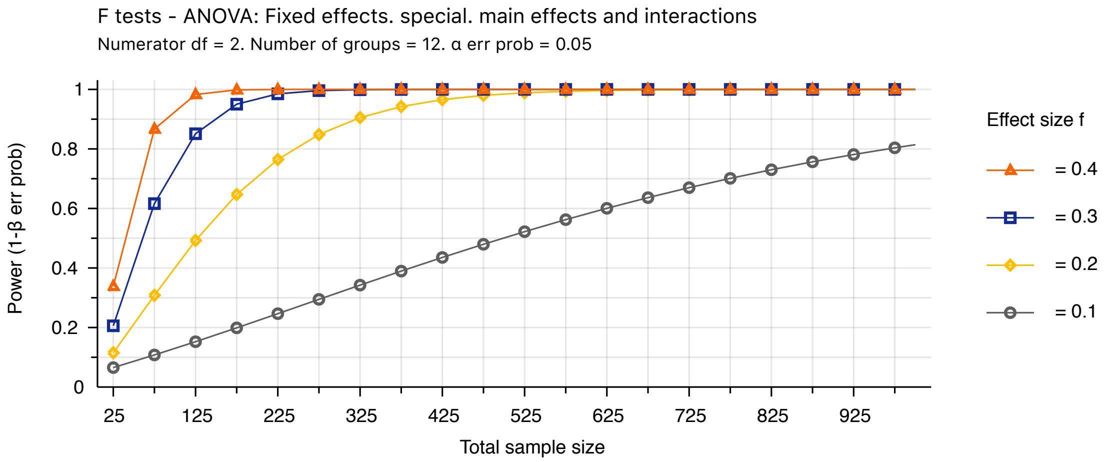


Figure 8.9. Power analysis by sample size for various effect sizes, *f*.

The third and final research focus considered the timing of the value selection task and brief affirmation prompt as an intervention to improve student performance. Fidelity of delayed implementation was critically low based on the need for students to ‘struggle’ to earn their intervention, operationalized as not being able to reach mastery within the first 5 problems. Only 10 students assigned to the delayed treatment group actually saw the value affirmation (13% of the initial sample), with only 8 reaching the posttest and being assessed in the analytic sample (12% of the analytic sample. Similarly, only 12 students assigned to the delayed control group actually saw the control affirmation (17% of the initial sample), with only 11 reaching the posttest and being assessed in the analytic sample (19% of the analytic sample). As such, analysis did not consider the main effect of timing, but instead assessed full condition (Treatment X Delay). The framework of the present work hypothesized that delay would strengthen the effectiveness of the treatment intervention, targeting students who were facing a challenge (Cohen, et al. 2009). Results suggested that full condition was not significant and exhibited no significant interaction effects. A main effect of full condition using this experimental design would have required anywhere from 179 to 1095 students to observe a moderate to small effect when $\beta = .20$ and $\alpha = .05$. This suggests that although the value affirmation intervention was considered effective when timing was disregarded, the effect was washed out in the full sample due to low power, and the true effect may be smaller than anticipated. However, based on the significant main effect of belonging, students with higher levels of baseline belonging averaged an additional 7-10 points on their posttests than their peers with moderate or low baseline belonging. This confirms that belonging should be considered as a critical element of success and may imply that the brief value affirmation was not a powerful enough intervention to improve belonging or its effects were not immediate enough to capture an effect on performance.

This finding leads to the broader limitation of this work: that sense of belonging was not reassessed to verify whether the value affirmation acted through a boost to students’ belonging or through another mechanism (provided it proved effective). The working assumption that the value affirmation affects students’ belonging to alter their academic performance was sourced from prior work (Cohen, et al. 2009) and without a secondary gauge of students’ sense of belonging, this could not be verified. However, repeating the PSSM at the start and end of a single assignment would likely show high reliability and little change, while doubling the assignment length for students. Realistically, the subtle change may not be immediately measureable in a secondary gauge of belongingness or may be driven by another mechanism (such as self-esteem) that is highly correlated to belonging (Ma, 2003). Future iterations of this work should examine this issue further, testing alternative sources of measurement and their correlation with students’ sense of belonging.

Despite low power, results of the present work offer a potential contribution to communities interested in the study of belonging, Self Determination Theory, and learning interventions in both online and offline environments. Results revealed limited evidence that a value affirmation intervention intended to boost students’ sense of belonging produced significantly stronger academic performance immediately following an online mathematics assignment. Due to low delayed treatment fidelity, it was not evident that this effect could be more powerful for students facing challenge. However, a significant interaction effect between treatment and knowledge level suggested that this intervention could be especially promising for low knowledge students, many of whom would qualify as being "at-risk." This suggests potential capacity for addressing the achievement gap through a moderately successful motivational intervention that requires little time and virtually no cost, which could potentially be repeated numerous times per year to produce salient positive recursive value.

CHAPTER 8

Aguilar, L., Walton, G., & Wieman, C. (2014). Psychological insights for improved physics teaching. Physics Today, 67(5): 43-49.

Alderfer, C. (1972). Existence, relatedness, & growth. New York: Free Press.

Baker, R. (2016) Stupid tutoring systems, intelligent humans. International Journal of Artificial Intelligence in Education. Volume 26, Issue 2, Springer. 600-614. Retrieved from http://www.columbia.edu/~rsb2162/STS-Baker-IJAIED-v15.pdf

Bandura, A. (1977). Self-efficacy: The exercise of control. New York: W.H. Freeman.

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.

Burnett, G., & Walz, G. (1994). Gangs in the schools (Rep. No. EDO-CG-94-28). Greensboro, NC: ERIC Clearinghouse on Counseling and Student Services.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). A self-affirmation intervention to reduce the racial achievement gap. Science. 313(5791): 1307-1310.

Cohen, G. L., & Garcia, J. (2008). Identity, Belonging, and Achievement: A Model, Interventions, Implications. Current Directions in Psychological Science. 17(6): 365-369.

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: intervening to close the minority achievement gap. Science. 324(5925): 400-3.

deCharms, R. (1968). Personal causation. New York: Academic Press.

Deci, E. L. (1975). Intrinsic Motivation. New York: Plenum.

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. New York: Plenum.

Finn, J. (1989). Withdrawing from school. Review of Educational Research, 59, 117-142.

Fredericks, J., McColskey, W., Meli, J., Mordica, J., Montrosse, B., & Mooney, K. (2011). Measuring student engagement in upper elementary through high school: A description of 21 instruments (Issues & Answers Report, REL 2011-098). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Goff, B. G. & Goddard, H.W. (1999). Terminal core values associated with adolescent problem behavior. Adolescence, 34, 47-60.

Goodenow, C. (1993a). Classroom belonging among early adolescent students: Relationships to motivation and achievement. Journal of Early Adolescence, 13(1), 21-43.

Goodenow, C. (1993b). The Psychological Sense of School Membership Among Adolescents: Scale Development and Educational Correlates. Psychology in the Schools. 30:79-90.

Hamre, B.K. & Pianta, R.C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. Child Development, 72, 625-638.

Harter, S. (1978). Effectance motivation reconsidered: Toward a developmental model. Human Development, 1, 661-669.

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. *International Journal of Artificial Intelligence in Education*. 24 (4): 470-497.

Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, G. & Soffer, D. (2013). Estimating the effect of web-based homework. In Lane, Yacef, Mostow & Pavlik (eds) Proceedings of the 16th International Conference on Artificial Intelligence in Education. Springer Berlin Heidelberg. 7926: 824-827.

Koedinger, K., McLaughlin, E. & Heffernan, N. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. Journal of Educational Computing Research. Baywood Publishing. 4: 489-510.

Lazowski, R.A. & Hulleman, C.S. (2016). Motivation Interventions in Education: A Meta-Analytic Review. Review of Educational Research. 86(2): 602-640.

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. Journal of Educational Psychology, 97, 184-196.

Ma, X. (2003). Sense of Belonging to School: Can Schools Make a Difference? The Journal of Educational Research. 96(6): 340-349.

Maehr, M. L., & Midgley, C. (1999). Creating optimum environments for students of diverse sociocultural backgrounds. In J. Block, S. T. Everson, & T. R. Guskey (Eds.), Comprehensive school reform: A program perspective (pp. 355-375). Dubuque, IA: Kendall/Hunt.

Maslow, A. (1943). A theory of human motivation. Psychological Review, 50, 370-396. Retrieved June 2001, from http://psychclassics.yorku.ca/Maslow/motivation.htm.

Mendicino, M., Razzaq, L. & Heffernan, N.T. (2009). Improving learning from homework using intelligent tutoring systems. Journal of Research on Technology in Education. 41 (3): 331-346.

National Governors Association Center for Best Practices (NGACBP) & Council of Chief State School Officers (CCSSO). (2010). Common Core State Standards. Washington, DC: Authors.

Niemiec, C. P. & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. Theory and Research in Education. 7: 133-144.

Nicholls J. G. (1979). Quality and equality in intellectual development: The role of motivation in education. American Psychologist, 34(11), 1071-1084.

Nohria, N., Lawrence, P., & Wilson, E. (2001). *Driven: How human nature shapes our choices*. San Francisco: Jossey-Bass.

Ostrow, K. S. (2018). Belonging within ASSISTments: OSF Registration Page. https://osf.io/wg2d6/, last accessed 2018/04/09.

Ostrow, K., Heffernan, N.T., Heffernan, C., & Peterson, Z. (2015). Blocking vs., Interleaving: Examining Single-Session Effects within Middle School Math Homework. In Conati, Heffernan, Mitrovic & Verdejo (eds.) Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015). Springer International Publishing. 388-347.

Ostrow, K. S., Heffernan, N. T., & Williams, J. J. (2017). Tomorrow's EdTech Today: Establishing a Learning Platform as a Collaborative Research Tool for Sound Science. Teachers College Record. 119(3): 1-36.

Ostrow, K., Selent, D., Wang, Y., Van Inwegen, E., Heffernan, N., & Williams, J.J. (2016). The Assessment of Learning Infrastructure (ALI): The Theory, Practice, and Scalability of Automated Assessment. In Proceedings of the 6th International Conference on Learning Analytics and Knowledge.

Pianta R, Belsky J, Vandergrift N, Houts R, Morrison F, the NICHD ECCRN Classroom Effects on Children's Achievement Trajectories in Elementary School. American Educational Research Journal. 2008; 45(2).

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS)*. Baltimore: Paul H. Brookes.

Pianta, R. C. & Hamre, B. K. (2009). Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. Educational Researcher. 38(2): 109-119.

Razzaq, L. & Heffernan, N. (2009). To Tutor or Not to Tutor: That is the Question. In Dimitrova, Mizoguchi, du Boulay & Graesser (Eds.) Proceedings of the 2009 Artificial Intelligence in Education Conference. IOS Press. 457-464.

Roschelle, J., Feng, M., Murphy, R. & Mason, C. (2016). Online Mathematics Homework Increases Student Achievement. AERA OPEN. October-December 2016, Vol. 2, No. 4, pp. 1–12. DOI: 10.1177/2332858416673968

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

Rotter, J.B. (1954). Social learning and clinical psychology. NY: Prentice-Hall.

Ryan, R. M., & Deci, E. L. (2009). Promoting self-determined school engagement: Motivation, learning, and well-being. In K. R. Wentzel & A. Wigfield (Eds.), Handbook on motivation at school (pp. 171-196). New York: Routledge.

Ryan, R., & Deci, E. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68-78. Retrieved from http://www.psych.rochester.edu/SDT/publications/documents/2000RyanDeciSDT.pdf.

Schwartz, D. L., Tsang, J. M., & Blair, K. P. (2016). The ABC's of How We Learn: 26 scientifically proven approaches, how they work, and when to use them. New York: W. W. Norton & Company.

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, Bull, Kay, & Mitrovic (eds) Proceedings of the 15th International Conference on Artificial Intelligence in Education Conference. Springer Berlin Heidelberg. 6738: 328–336.

Soffer, D., Das, V., Pellegrino, G., Goldman, S., Heffernan, N., Heffernan, C., & Dietz, K. (2014). Improving long-term retention of mathematical knowledge through automatic reassessment and relearning. American Educational Research Association Conference, Division C - Learning and Instruction / Section 1c: Mathematics.

Thompson, M., Grace, C., & Cohen, L. (2001). Best friends, worst enemies: Understanding the social lives of children. New York: Ballantine Books.

U.S. Department of Education. (2010). Transforming American education: Learning powered by technology, the national educational technology plan. Office of Educational Technology, Washington, D.C.

Vygotsky, L.S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. Science, 331(6023), 1447-1451.

Wahba, A., & Bridgewell, L. (1976). Maslow reconsidered: A review of research on the need hierarchy theory. Organizational Behavior and Human Performance, 15, 212-240.

Wehlage, G. (1989). Dropping out: Can schools be expected to prevent it? In L. Weis, E. Farrar, & H. Petrie (Eds.), Dropouts from school. Albany, NY: State University of New York Press.

White, R. W. (1963). Ego and reality in psychoanalytic theory. New York: International Universities Press.

You, S., Ritchey, K, Furlong, M., Shochet, I.M., & Boman, P. (2011). Examination of the latent structure of the psychological sense of school membership scale. Journal of Psychoeducational Assessment. 29(3): 225-237.

9 Competence

This chapter describes a randomized controlled trial designed to target the basic psychological need of competence. Supported by Self-Determination Theory, this work considers the inclusion and timing of a skill lesson as an intervention intended to enhance learning outcomes. The effect of the intervention is examined with regard to posttest performance and Intrinsic Motivation Inventory subscale scores. Moderators of student's initial confidence levels and prior knowledge are also considered. This chapter was prepared for conference submission to EDM 2018 and reviewers offered valuable feedback. Portions of this work will be revised and resubmitted to another venue following completion of my degree.

Proper citation of this chapter is as follows:

Ostrow, K. S. & Heffernan, N. T. (In Preparation). Confidence and Prior Knowledge as Moderators of Lesson Placement on Performance and Intrinsic Motivation.

9.1 Introduction

9.1.1 Motivation in Education

Motivational interventions in education consistently reveal benefits to performance and motivation with little cost or consequence. Lazowski & Hulleman (2016) conducted a meta-analysis of 74 motivational interventions within educational contexts that covered 15 commonly referenced theoretical frameworks, including achievement goals, expectancy-value, implicit theories of intelligence (mindsets), self-affirmation, self-determination, and social belongingness. These interventions produced a powerful average effect size ($d = 0.49$, 95% CI = [0.43, 0.56]), suggesting the practical importance of socioemotional interventions for education (Lazowski & Hulleman, 2016). Positive effects are further emphasized by the low costs and accessibility of most motivational interventions, requiring few resources and minimal time on task. Additively or recursively implementing motivational supports for students may offer a simple, empathetic approach toward stronger performance and academic enjoyment.

9.1.2 Self Determination Theory

Self Determination Theory (SDT), developed by (Deci & Ryan, 1985; Ryan & Deci, 2000; Ryan & Deci, 2002), is a macro-theory of human motivation that combines theories supporting students' feelings of competence (or, their ability to meet academic challenges: Harter, 1978; White 1963) and autonomy (or, participating of their own volition: deCharms, 1968), with the innate need for relatedness (or, having a sense of community: Maslow, 1943). SDT posits that motivation can progress along a continuum and grow increasingly intrinsic through supports for these basic psychological needs. Lazowski & Hulleman's work (2016) specifically considered 11 interventions rooted in self-determination theory, with a high average effect size (d = 0.70, 95% CI = [0.53, 0.87]).

SDT has practical importance for education, as learning tasks tend to undermine intrinsic motivation. However, the integration of extrinsic motivators (i.e., grades) promotes ownership and valuing of educational tasks, thereby increasing performance and enjoyment (Ryan & Deci, 2000; Niemiec & Ryan, 2009). Standardized testing and assessments with controlling orientations can instill anxiety, boredom, and alienation (Niemiec & Ryan, 2009). SDT posits that support for competence, especially in an autonomy supportive environment, can establish autonomous orientations that emphasize well-being. Support for competence can come in many forms, including the provision of optimal levels of challenge that fall within a student's zone of proximal development (Vygotsky, 1978) and instruction or feedback that emphasizes effectance rather than evaluation (Harter, 1978). The present work considers the efficacy of a worked example skill lesson as a motivational intervention targeting students' perceptions of competence, interest/enjoyment, and ultimately, their subsequent academic performance.

9.1.3 Online Learning Environments

Online learning environments provide a natural setting for examining student competence. Considering new-age capacity for connection, educational stakeholders have delegated many traditional learning tasks to online environments. With this influx of online learning, student errors can be corrected with immediacy, log files make student data more prevalent and actionable, and teachers can intervene with efficiency and efficacy. Many online learning environments were developed with the goal of artificial intelligence automating scholastic tasks or serving to personalize education. However, with the creation of these platforms came the creation of fields focused on their efficacy. Endeavors in educational data mining, learning analytics, and authentic educational research at scale have grown popular in recent years by unveiling the power of online learning environments: their data. By supporting data driven classroom discourse (Baker, 2016) these systems afford new, data-rich methodologies capable of strengthening learning processes.

The present work is conducted within ASSISTments (www.ASSISTments.org), an online learning platform provided as a free service of Worcester Polytechnic Institute and used by more than 50,00 students for mathematics classwork and homework. The platform has over 25,000 certified problems mapped to the Common Core State Standards (NGACBPC & CCSSO, 2010). The platform offers students with instructional *ASSIST*ance while serving as a powerful assess*MENT* tool for teachers, thus establishing the moniker "ASSISTments" (Heffernan & Heffernan, 2014). A recent efficacy trial revealed that use of the platform for homework over the course of a year resulted in large, reliable learning gains on standardized assessments (Roschelle, Feng, Murphy & Mason, 2016). ASSISTments also serves as a powerful tool for researchers, hosting randomized controlled trials featuring student-level randomization to efficiently investigate best practices in online education (ASSISTments TestBed Resource Guide, 2018). As such, it offered an optimal environment from which to examine how skill lesson placement can affect performance and motivation, as moderated by prior knowledge and initial confidence.

9.1.4 Framework & Research Questions

This study speculates that students' performance and ultimately their perceptions of competence, interest, and enjoyment, differ based on their prior knowledge (or global skill level) and initial confidence in the skill content. Thus, we examine the effects of lesson placement on posttest score and IMI score as moderated by prior knowledge or initial confidence using the following four research questions:

1. Does initial confidence predict later IMI Score?
2. Does lesson placement predict posttest score or IMI score?
3. Is the effect of lesson placement on posttest score or IMI score moderated by prior knowledge?
4. Is the effect of lesson placement on posttest score or IMI score moderated by initial confidence estimates?

The conceptual and statistical diagrams of these moderation effects are shown (condensed for brevity) in Figure 9.1. It is hypothesized that initial confidence will correlate with higher scores on the IMI and its subscales, and that students with, higher initial confidence and ultimately higher IMI scores, will exhibit stronger posttest performance. It is also hypothesized that students receiving a skill lesson (regardless of timing) will outperform students in the "Business as Usual" control. We suspect that students will gain more from a delayed skill lesson, presented when they are in need of assistance, than from an immediate lesson that can be more readily skimmed and ignored before students are aware that they may need the additional skill assistance. However, it is possible that delayed lesson placement and the opportunity to struggle will negatively affect students' perceptions of competence as well as their interest in and enjoyment of their assignment. It is also suspected that the effect of lesson placement will be moderated by prior knowledge. In theory, the effect should be more robust in low performing students. However, it is possible that when initial confidence is also low, students may simply resort to gaming the system. As such, it is also hypothesized that initial confidence in the math material will moderate students' posttest performance, with those reporting high confidence expected to outperform those reporting lower confidence.

9.2 Methods

9.2.1 Sample

Participants were collected using ASSISTments by embedding a randomized controlled intervention, posttest, and subsequent survey material into an ASSISTments Certified Skill Builder. A sample consisting of 425 students was collected between January and March of 2018. This sample was representative of 15 classes administered by 10 teachers at 7 schools in 3 States. Teachers had an average of 42.5 students (SD = 31.6, median = 37.5). A name-based predictive feature of ASSISTments' log files suggested that 32% of the sample was female, 40% was male,

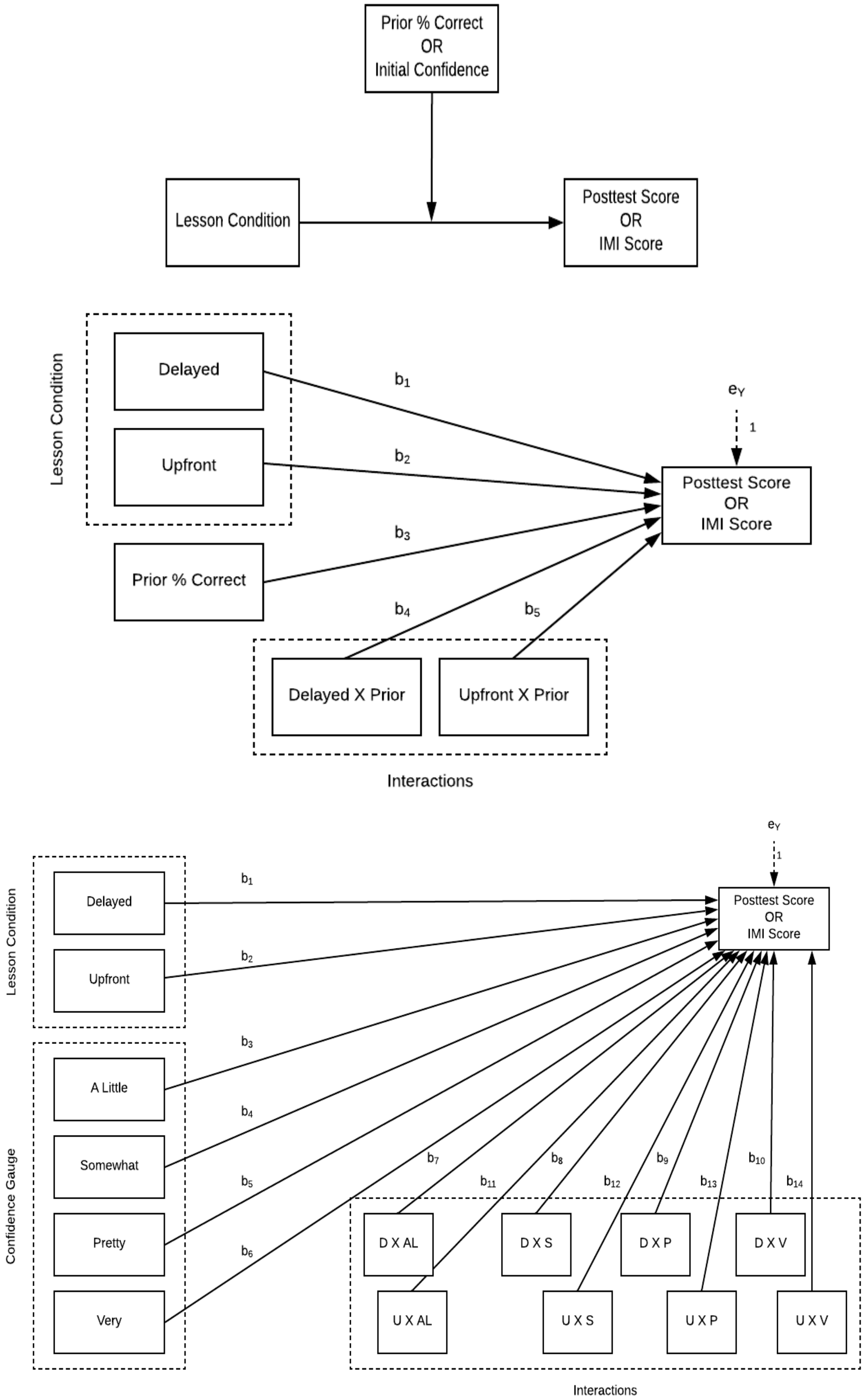


Figure 9.1. Conceptual and statistical diagrams of the moderation of the effect of lesson placement on posttest score or IMI score by prior knowledge or initial confidence. Models are combined where possible for brevity.

and 27% was undefined by name alone. The majority of students (77%) were in the 7th grade, while 3% were in the 6th grade, and the remaining 20% did not have a defined class grade in the log files.

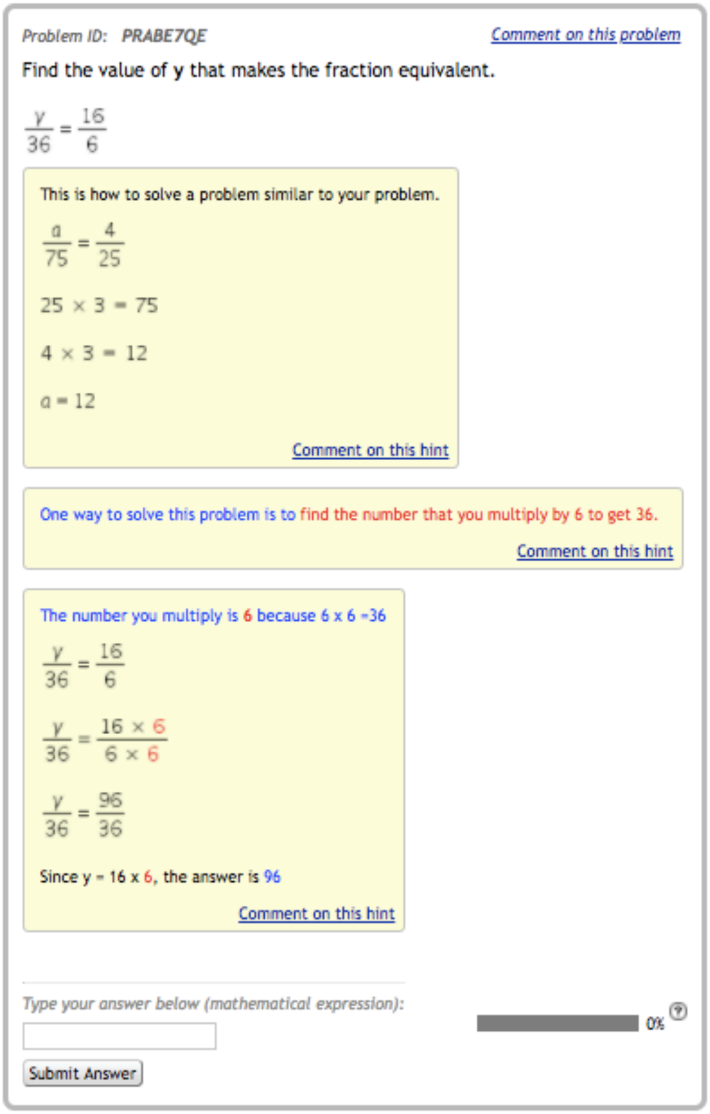
Within the sample, 330 students completed the entire assignment and 177 students opted-in to answer survey content. Of the 359 students that reached the opt-in (29 students showed attrition between posttest completion and final assignment submission), 214 opted to participate, while 145 opted out. Thus, 37 students who opted-in to answer survey items failed to complete and submit their survey and assignment. A total of 95 students had incomplete assignments. The analytic sample was reduced based on missing data using listwise deletion. As such, results are based on samples with complete response patterns for the dependent and independent variables of interest.

9.2.2 Math Material

The present study was embedded within an ASSISTments Certified Skill Builder with mathematics content mapped to the Common Core State Standard (NGACBP & CCSSO, 2010) of "Equivalent Ratios," a sixth grade skill in Ratios and Proportional Reasoning (6.RP.A.3a). Skill Builders are mastery-learning-based assignments that provide fluency practice, requiring that students accurately solve three consecutive problems to complete or "master" the assignment. Embedding research within Skill Builders allows researchers to collect data from authentic learning environments in minimally invasive ways using the ASSISTments TestBed (www.ASSISTmentsTestBed.org). Sample populations increase steadily over time as teachers assign study content to their students naturalistically. The present skill was chosen because of its popularity during the data collection period of interest based on usage data from the previous year. Problems within the assignment required students to solve for a missing value in a set of equivalent fractions, as shown in Figure 9.2.

9.2.3 Intrinsic Motivation Inventory

The IMI is a multidimensional scale intended to measure the subjective experiences of participants (Intrinsic Motivation Inventory, 2018; Ryan & Deci, 2000). Iterations of the IMI have been used for more than 30 years, with well-established validity and subscale reliability across tasks, conditions, and settings (Intrinsic Motivation Inventory, 2018). The IMI features seven subscales that can be mixed and matched to suit research needs: interest/enjoyment, perceived competence, effort, value/usefulness, felt pressure and tension, perceived choice, and relatedness/belonging. Past work has suggested that order effects are negligible and that subscales can be included or excluded as necessary (Intrinsic Motivation Inventory, 2018; Schell & Oswald, 2013). Subscales for perceived competence and interest/enjoyment are considered in the present work, and are combined to form an overall IMI Score, as described in the subsections below. All scale items were modified slightly to reflect an academic task or setting; such modifications are thought to be inconsequential to outcomes (Intrinsic Motivation Inventory, 2018). Students were asked to indicate how true each statement was for them using a Likert scale (1 = *Not at All True*, 7 = *Very True*).



Problem ID: PRABE7QE

Comment on this problem

Find the value of y that makes the fraction equivalent.

$\frac{y}{36} = \frac{16}{6}$

This is how to solve a problem similar to your problem.

$\frac{a}{75} = \frac{4}{25}$

$25 \times 3 = 75$

$4 \times 3 = 12$

$a = 12$

Comment on this hint

One way to solve this problem is to find the number that you multiply by 6 to get 36.

Comment on this hint

The number you multiply is 6 because 6 x 6 =36

$\frac{y}{36} = \frac{16}{6}$

$\frac{y}{36} = \frac{16 \times 6}{6 \times 6}$

$\frac{y}{36} = \frac{96}{36}$

Since y = 16 x 6, the answer is 96

Comment on this hint

Type your answer below (mathematical expression):

Submit Answer

0%

Figure 9.2. Math content.

9.2.3.1 Interest/Enjoyment

This subscale is the primary measure of intrinsic motivation. It has previously been shown that Autonomy and Competence scores in respective scales are predictive interest/enjoyment scores (Intrinsic Motivation Inventory, 2018), suggesting that these needs, when fulfilled, increase internalization, integration, and intrinsic motivation. In its full form, this subscale includes seven items regarding intrinsic motivation (i.e., “I enjoyed doing this assignment very much”). Higher values indicate a stronger sense of interest and enjoyment. Prior work validating this subscale within ASSISTments achieved high reliability by including five items, α = .92 (Ostrow & Heffernan, In Press). These five items were averaged in the present work to achieve a single Interest/Enjoyment subscale score.

9.2.3.2 Perceived Competence

This subscale is the primary measure of competence or feeling capable and confident. Scores from this subscale have previously been shown to predict Interest/Enjoyment (Intrinsic Motivation Inventory, 2018). In its full form, this subscale includes six items estimating perceived competence (i.e., "I am satisfied with my performance on this assignment"). Higher values indicate a stronger sense of perceived competence. Prior work validating this subscale within ASSISTments achieved high reliability by including five items, $\alpha = .92$ (Ostrow & Heffernan, In Press). These five items were averaged in the present work to achieve a single Perceived Competence subscale score.

9.2.3.3 IMI Score

The ten items from the Interest/Enjoyment and Perceived Competence subscales were averaged to form an aggregate IMI Score, following the scoring protocols traditionally implemented with use of the IMI (Intrinsic Motivation Inventory, 2018). Higher values indicate stronger motivation for the assignment as defined by respective subscales.

9.2.4 Other Student Measures

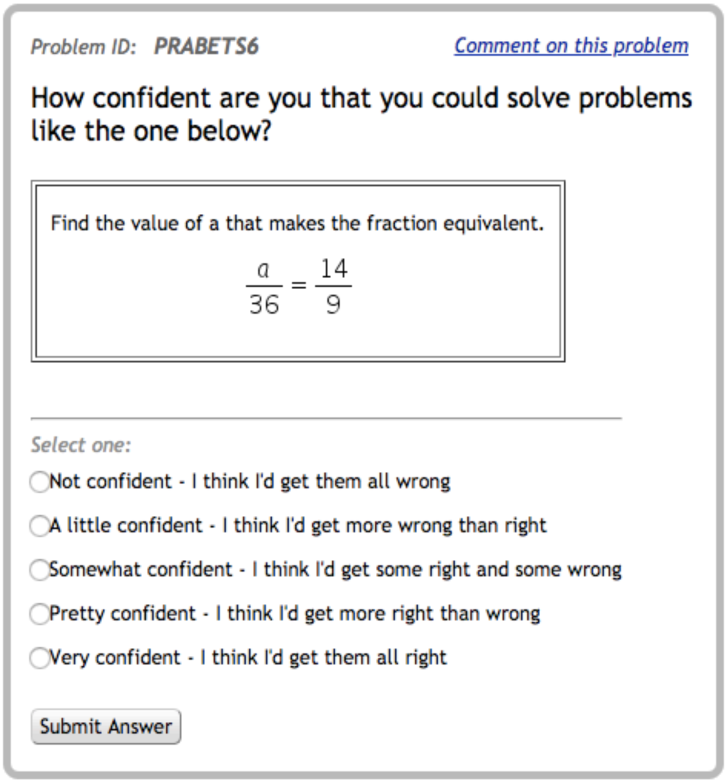
Student performance measures were logged by ASSISTments and describe students' mathematics abilities before and after their math assignment. Metrics considered in the present work are described in the subsections below.

9.2.4.1 Initial Confidence

At the start of the assignment, students were asked about their confidence in their ability to solve similar problems, as shown in Figure 9.3. Students responded by selecting from a multiple-choice selection of five options ranging from "Not confident – I think I'd get them all wrong" to "Very confident – I think I'd get them all right." This nominal variable was then dummy coded for use as a moderator in regression analyses predicting posttest and IMI scores.

9.2.4.2 Prior Knowledge

All problem accuracy data from student's work within ASSISTments prior to their participation in the present study was averaged into a single value meant to serve as a metric of global math ability. For instance, a student who has accurately solved 75 out of 100 attempted problems within various ASSISTments assignments would have a prior knowledge score of 0.75, or 75%. This metric does not control for the amount of work a student has completed (i.e., another student accurately solving 3 out of 4 attempted problems would share the score of 0.75), or the difficulty levels of problems solved, but has been shown to reflect significantly different types of learners when the value is dichotomized on its median within a dataset; resulting groups of "low prior knowledge" students have exhibited significantly more impactful reactions to learning



Problem ID: PRABETS6

How confident are you that you could solve problems like the one below?

Find the value of a that makes the fraction equivalent.

$$\frac{a}{36} = \frac{14}{9}$$

Select one:

○ Not confident - I think I'd get them all wrong

○ A little confident - I think I'd get more wrong than right

○ Somewhat confident - I think I'd get some right and some wrong

○ Pretty confident - I think I'd get more right than wrong

○ Very confident - I think I'd get them all right

Figure 9.3. Initial confidence gauge.

interventions than groups of "high prior knowledge" students, suggesting that such interventions may target the achievement gap (Razzaq & Heffernan, 2009; Ostrow, Heffernan, Heffernan, & Peterson, 2015). As such, prior percent correct is used as a quantitative scale measure in the present work, with simple effects graphs displaying regression effects at the mean and +/- 1SD.

9.2.4.3 Posttest Score

Following Skill Builder content, participants were subjected to a three-item posttest. The posttest consisted of transfer items meant to measure students' ability to solve equivalence problems of a higher difficulty level than those practiced within the assignment. Problems from the same skill area were sourced from 7th and 8th grade materials within ASSISTments and were presented to all students in the same order, progressing from "Easy" (i.e., near transfer) to "Hard" (i.e., far transfer). Problems were presented without hint feedback, but students were allowed to make multiple attempts and to request the correct answer before proceeding. ASSISTments logged the accuracy of each student's first responses on each problem and values were averaged to arrive at their overall posttest score.

9.2.5 Procedure

The experimental design of this randomized controlled trial is presented in Figure 9.4. Students were initially asked to complete a gauge of their initial confidence in the specific math skill that would be the subject of the Skill Builder (Figure 9.3). Students were then randomly assigned to either a traditional, "Business as Usual" (BAU) Skill Builder, or to a Skill Builder supplemented by a brief worked example lesson. Those assigned to receive the lesson were randomly split to receive the lesson immediately before starting to solve problems (Upfront), or following struggle defined as not being able to achieve mastery within the first six problems (Delayed). The lesson is presented in Figure 9.5.

Upon achieving mastery by accurately solving three consecutive problems (or ending the assignment by completing all available problems within their randomly assigned condition), students were routed into a three-item posttest comprised of near and far transfer items.

Following the posttest, students were notified that survey content was available. Ideally, aspects of personalization underlie the goal of research examining the efficacy of learning interventions like lesson placement on performance, perceived competence, and interest and enjoyment. As such, students were told that opting-in would help the developers of the system to better tailor their future experiences within ASSISTments. Those who opted-in were give the IMI subscales (n = 214), while those who opted-out were given an assignment completion screen (n = 145). IMI items were delivered within their scale constructs. All students were asked to respond to all items pertaining to Interest/Enjoyment before moving on to all items pertaining to Perceived Competence in a separate page view. Survey data was collected by integrating Qualtrics, a popular survey infrastructure, with ASSISTments via the ASSISTments Survey System available through the ASSISTments TestBed (ASSISTments TestBed Resource Guide, 2018). This system uses an iframe to establish a connection between the two platforms and pass information back and forth, linking survey responses to mathematics performance through student and assignment identifiers.

ASSISTments data was retrieved using the Assessment of Learning Infrastructure (ASSISTments TestBed Resource Guide, 2018; Ostrow, Selent, Wang, Van Inwegen, Heffernan & Williams, 2016). IMI data was retrieved and compiled from Qualtrics and merged with ASSISTments data before being preprocessed and analyzed in IBM SPSS Statistics. Variables were cleaned, and missing data was labelled for proper exclusion from analysis. While merging data, redundancies were removed (24 students from one class accessed the survey multiple times due to a system error; in these cases, first responses were retained, and redundant responses were discarded). The resulting data file contained posttest scores for 382 students and complete survey responses from 177 students. Items were reverse scored as necessary and reliability analyses were conducted. De-identified survey data is available at (Ostrow & Heffernan, 2018) for further reference.

9.3 Results

An initial assessment of the sampling distribution was conducted to assure that differential attrition did not occur across conditions for assignment completion or survey participation. Assignment completion was not significantly different by lesson placement, χ^2 (2) = 5.62, p = .060, and neither

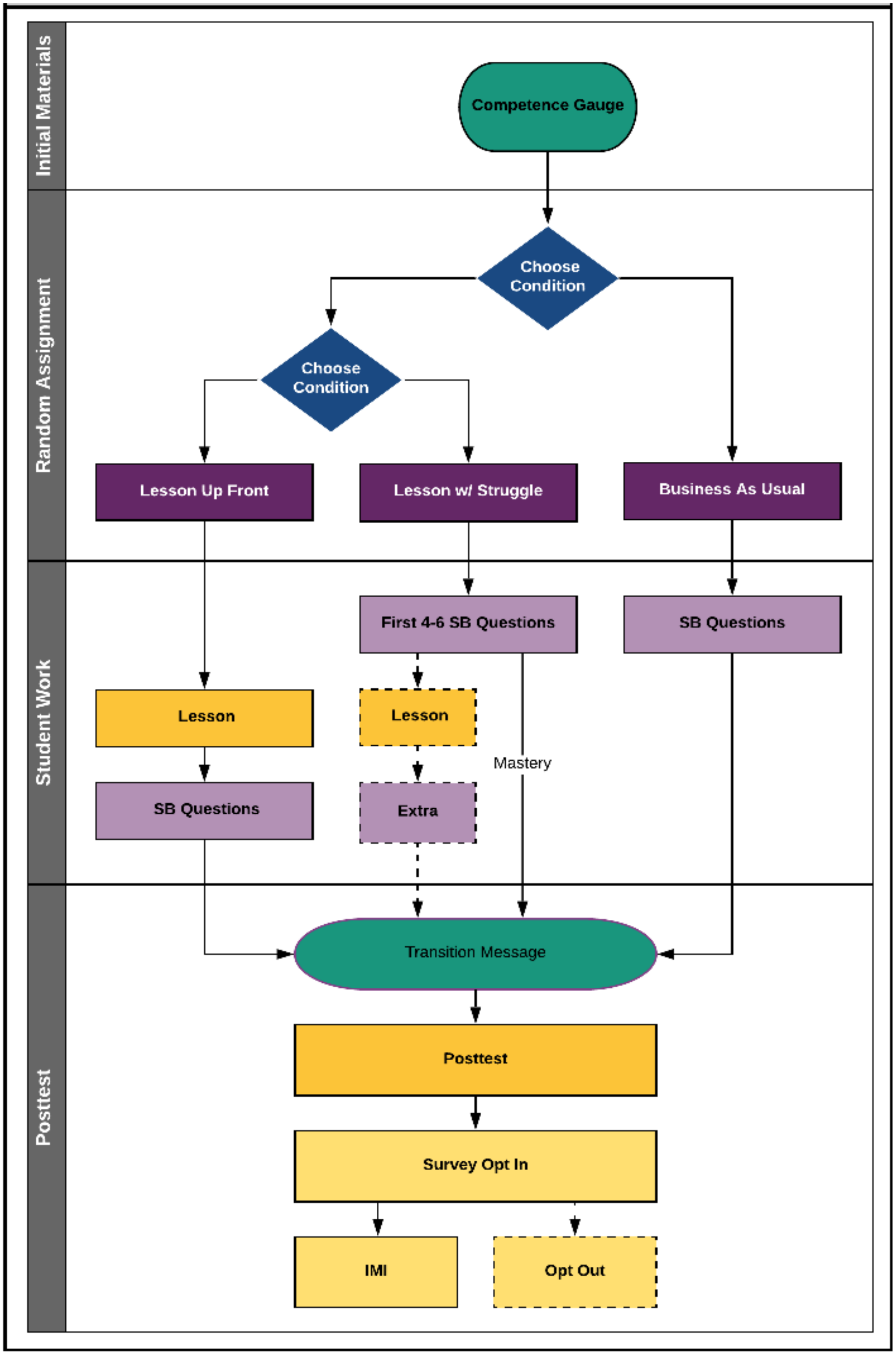
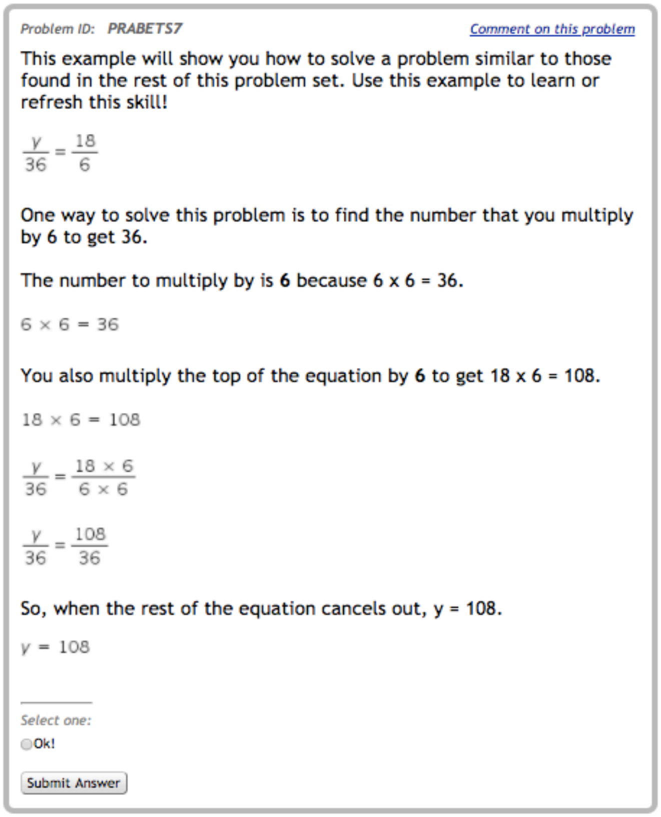


Figure 9.4. Experimental design.



Problem ID: PRABETS7

This example will show you how to solve a problem similar to those found in the rest of this problem set. Use this example to learn or refresh this skill!

$$\frac{y}{36} = \frac{18}{6}$$

One way to solve this problem is to find the number that you multiply by 6 to get 36.

The number to multiply by is **6** because 6 x 6 = 36.

$6 \times 6 = 36$

You also multiply the top of the equation by **6** to get 18 x 6 = 108.

$18 \times 6 = 108$

$$\frac{y}{36} = \frac{18 \times 6}{6 \times 6}$$

$$\frac{y}{36} = \frac{108}{36}$$

So, when the rest of the equation cancels out, y = 108.

$y = 108$

Select one:

○ Ok!

Figure 9.5. Worked example lesson.

was survey opt-in by lesson placement, $\chi^2(2) = 3.48$, p = .176. However, as shown in Table 9.1, students who were not confident had a completion rate of 60%, while those who were very confident had a completion rate of 89%. Discrepancies in completion rates specific to posttest and IMI scales are detailed by lesson placement condition in Table 9.2. The inconsistency in completion rate sample sizes between Table 9.1 and Table 9.2 comes from students who finished the posttest but failed to formally submit their assignment (n = 52). These differences should not necessarily deter the present analysis because the use of confidence as a moderator should capture this interaction. It should be kept in mind that the direction of this attrition likely results in inflated posttest and IMI scores within the analytic sample.

A pressing issue and potential limitation of the current study was the lack of intervention fidelity in the Delayed lesson condition. As shown in Table 9.3, although 116 students were randomly assigned to receive a lesson during struggle, as defined by not being able to master the Skill Builder within six questions, only three students actually “struggled.” This is likely based on a definition of struggle that was too lenient and allowed for self-correction or hint usage that would allow for strugglers to reach mastery before the lesson could be provided. Further, ceiling effects from the skill being too easy for students may have influenced this outcome. As shown in Figure 9.2,

Table 9.1. Completion rates based on initial confidence.

	Complete	**Incomplete**	**%**
Not confident - I think I'd get them all wrong	21	14	60
A little confident - I think I'd get more wrong than right	22	16	58
Somewhat confident - I think I'd get some right and some wrong	67	24	74
Pretty confident - I think I'd get more right than wrong	92	25	79
Very confident - I think I'd get them all right	128	16	89
Total	**330**	**95**	**78**

Table 9.2. Sample sizes for completed/assigned posttests and IMI scales by initial confidence and condition.

	Posttest					**IMI**				
	BAU	Delayed	Upfront	Total	%	BAU	Delayed	Upfront	Total	%
Not confident	16/19	8/11	2/5	26/35	74	3/19	3/11	2/5	8/35	23
A little confident	12/14	7/13	9/11	28/38	74	7/14	3/13	5/11	15/38	39
Somewhat confident	45/47	18/23	19/21	82/91	90	25/47	10/23	12/21	47/91	52
Pretty confident	52/56	31/33	24/28	107/117	91	23/56	19/33	10/28	52/117	44
Very confident	61/65	35/36	43/43	139/144	97	16/65	13/36	26/43	55/144	38
Total	186/201	99/116	97/108	382/425	90	74/201	48/116	55/108	177/425	42
%	93	85	90	90		37	41	51	42	

Table 9.3. Distribution of students receiving lesson or hints.

	No Lesson	**Lesson**	
	BAU (n = 201)	Delayed (n = 116)	Upfront (n = 108)
Lesson	0	3	108
Hints	37	24	18
Lesson + Hints	0	3	18

on demand hint feedback followed a format similar to the worked example provided in the lesson (Figure 9.5), just with slightly less structure. As such, the effect of lesson placement was also likely weakened in the present analysis by washout from hint usage.

9.3.1 Confidence & Competence

RQ1. Does initial confidence predict later IMI Score? Linear regression was used to predict IMI score from students' initial confidence estimates. Results suggested that initial confidence estimates significantly predict IMI score, $R^2 = .05$, $F(1, 175) = 9.28$, $p < .01$. Thus, although this significant correlation exists, confidence only explains 5% of the variance in IMI scores. Predicted

IMI score increases by .26 with each incremental increase in initial confidence (β = .26, p < .01). Thus, a very confident student exhibits a 1.04-point improvement on the combined IMI (1-7).

Initial confidence estimate was also a significant predictor of students' Perceived Competence subscale scores, $R^2 = .11$, F(1, 177) = 21.44, p < .001. Increases in initial confidence significantly predict higher perceived competence (β = .42, p < .001). However, it was not a significant predictor of students' Interest/Enjoyment subscale scores, $R^2 = .003$, F(1, 184) = 0.59, p = .444.

Interestingly, as shown in previous work with the IMI subscales (Intrinsic Motivation Inventory, 2018), students' perceived competence subscale scores were predictive of their Interest/Enjoyment subscale scores, $R^2 = .18$, F(1, 175) = 38.13, p < .001. The correlation table (Table 9.4) summarizes the significant relationships between confidence and IMI measures.

For reference, the reduced IMI subscales implemented based on prior validation and scale reduction (Ostrow & Heffernan, In Press) again retained high reliability with the current sample with five items in the Interest/Enjoyment subscale (α = 0.92), and five items in the Perceived Competence subscale (α = 0.901). This suggests generalization of previously established scale reliability across a new ASSISTments population, helping to further validate the scale.

9.3.2 Moderation Analyses

9.3.2.1 Lesson Placement Predicting Posttest Score

RQ2.1. Does lesson placement predict posttest score? Linear regression was used to predict posttest score from lesson placement condition. Results suggested that lesson placement did not significantly predict posttest score, $R^2 = .01$, F(2, 375) = 1.97, p = .141. This is possibly due in part to the extreme lack of intervention fidelity in the Delayed lesson condition, noted in Table 9.3. With the control group as the reference group (constant, β = .68, p < .001), assignment to the Delayed lesson condition did not produce significantly different posttest scores (β = .05, p = .218). Assignment to the Upfront lesson condition was suggestive of different posttest scores, but the difference did not meet conventional measures for statistical significance (β = .08, p = .060).

Regardless of the lack of a direct effect, moderation analyses were still conducted because they explore a different research question – does the effect of lesson change based on levels of a moderator? As Hayes notes, "Evidence of an association between X and Y is not required in order

Table 9.4. Correlations between Confidence and IMI scales.

Independent Variables	1	2	3	4
1. Confidence Gauge				
2. IMI Score	**.22**			
3. Perceived Competence	**.33**	**.82**		
4. Interest/Enjoyment		**.86**	**.42**	

Note. Bold correlations significant at p < .01. Suppressed correlations not statistically significant.

for X's effect to be moderated, just as the existence of such an association says nothing about whether that association is dependent on something else" (Hayes, 2018, p. 235). As such, the effect of lesson is examined as moderated by prior knowledge and initial confidence in the subsections below, and as depicted in Table 9.5 and Table 9.6.

9.3.2.1.1 Moderated by Prior Knowledge

RQ3.1. Is the effect of lesson placement on posttest score moderated by prior knowledge? Prior knowledge was a significant control when added to the regression model, $R^2 = .18$, $F(3, 374) = 26.60$, $p < .001$. When allowing the effect of lesson to vary across prior knowledge (i.e., the interaction effect), the model was significant, $R^2 = .18$, $F(3, 372) = 16.75$, $p < .001$, but significant differences were not observed by the addition of interaction effects to the model, as shown in Table 9.5. Thus, prior knowledge was not a significant moderator of the effect of lesson on posttest score, but it offered a useful control. This finding is depicted visually in Figure 9.6.

9.3.2.1.2 Moderated by Initial Confidence

RQ4.1. Is the effect of lesson placement on posttest score moderated by initial confidence estimates? Initial confidence was a significant control when added to the regression model, $R^2 = .12$, $F(6, 375) = 8.60$, $p < .001$. Specifically, students who were "Pretty Confident" or "Very Confident" had significantly greater posttest scores, $\beta = .22$, $p < .01$ and $\beta = .30$, $p < .001$, respectively. This is likely because students were portraying appropriate meta-cognitive skills and their performance aligned with their confidence in their abilities. When allowing the effect of lesson to vary across prior knowledge, the model was significant, $R^2 = .14$, $F(14, 367) = 4.25$, $p < .001$. Significant differences were observed for students who were "A Little" confident and who experience an Upfront lesson ($\beta = .58$, $p < .05$). Main effects for those who were "Pretty Confident" or "Very Confident" remained, as shown in Table 9.5. Thus, initial confidence was a significant moderator of the effect of lesson on posttest score. This finding is depicted visually in Figure 9.7.

9.3.2.2 Lesson Placement Predicting IMI Score

RQ2.2. Does lesson placement predict IMI score? A Linear regression predicting IMI score from lesson placement condition suggested that lesson placement did not significantly predict IMI score, $R^2 = .02$, $F(2, 173) = 2.04$, $p = .133$. Again, the lack of a direct effect here may be due in part to low intervention fidelity in the Delayed lesson condition (Table 9.3). With the control group as the reference group (constant, $\beta = 4.06$, $p < .001$), assignment to the Delayed lesson condition trended toward predicting significantly different IMI scores ($\beta = .46$, $p = .055$). However, assignment to the Upfront lesson condition did not predict significantly different IMI scores, ($\beta = .31$, $p = .184$). Moderation analyses were conducted to explore whether the effect of lesson condition changes by levels of each moderator, as described in the subsections below and in Table 9.5 and Table 9.6.

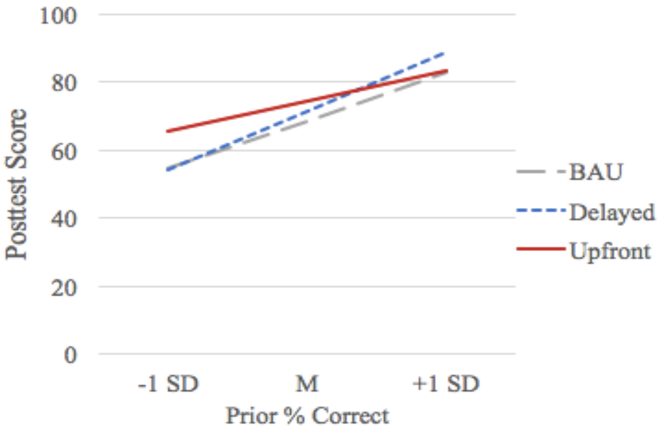


Figure 9.6. Posttest Score as moderated by prior knowledge.

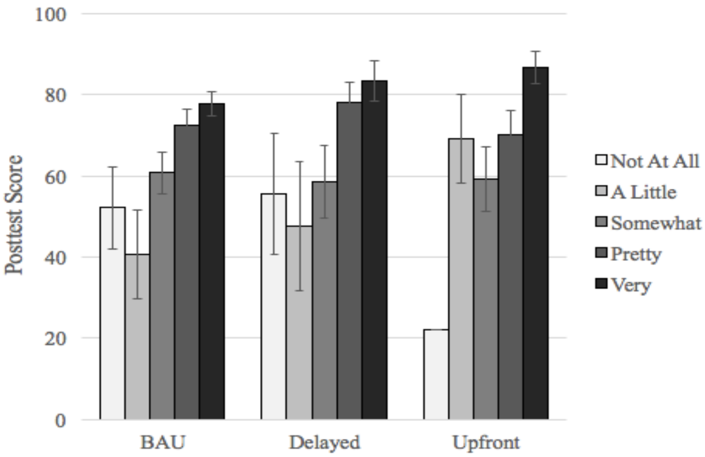


Figure 9.7. Posttest Score as moderated by initial confidence.

9.3.2.2.1 Moderated by Prior Knowledge

RQ3.2. Is the effect of lesson placement on IMI score moderated by prior knowledge? Prior knowledge was not a significant control when added to the regression model, $R^2 = .03$, F(3, 172) = 1.91, p = .13. When allowing the effect of lesson to vary across prior knowledge, the model was not significant, $R^2 = .04$, F(5, 170) = 1.24, p = .292. The lack of significant differences for interactions is noted in Table 9.5. Prior knowledge was not a significant moderator of the effect of lesson on IMI score. This finding is depicted visually in Figure 9.8.

9.3.2.2.2 Moderated by Initial Confidence

RQ4.2. Is the effect of lesson placement on IMI score moderated by initial confidence estimates? Initial confidence was a significant control when added to the regression model, $R^2 = .13$, F(6,

Table 9.5. Lesson placement, prior knowledge, and interactions on posttest score and IMI score.

		Posttest Score				IMI Score			
	Path	Coeff.	SE	t	p	Coeff.	SE	t	p
Model 1									
Constant	i_Y	.68	.02	27.89	< .001	4.06	.15	27.11	< .001
Delayed Lesson	b_1	.05	.04	1.24	.218	.46	.24	1.94	.055
Upfront Lesson	b_2	.08	.04	1.89	.060	.31	.23	1.33	.184
Model 2									
Constant	i_Y	.02	.08	.22	.825	3.31	.60	5.51	< .001
Delayed Lesson	b_1	.03	.04	.80	.427	.45	.24	1.90	.059
Upfront Lesson	b_2	.05	.04	1.38	.168	.27	.23	1.20	.233
Prior % Correct	b_3	.01	.01	8.67	< .001	.01	.01	1.28	.203
Model 3									
Constant	i_Y	-.01	.11	-.07	.947	2.91	.97	2.99	.003
Delayed Lesson	b_1	-.13	.20	-.68	.500	1.51	1.50	1.01	.315
Upfront Lesson	b_2	.31	.19	1.59	.112	.60	1.42	.42	.673
Prior % Correct	b_3	.01	.01	6.39	< .001	.02	.01	1.20	.232
Delayed X Prior	b_4	.01	.01	.83	.406	-.02	.02	-.72	.475
Upfront X Prior	b_5	-.01	.01	-1.34	.183	-.01	.02	-.24	.810
Model 1		R^2 = 0.010, MSE = 0.108				R^2 = 0.023, MSE = 1.636			
Model 2		R^2 = 0.176, MSE = 0.090				R^2 = 0.032, MSE = 1.630			
Model 3		R^2 = 0.184, MSE = 0.090				R^2 = 0.035, MSE = 1.644			

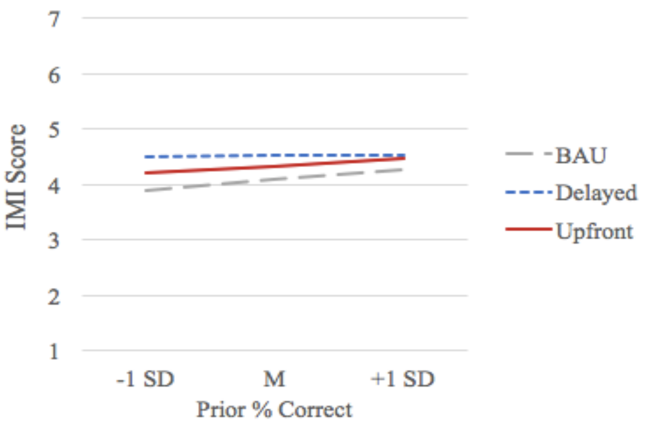


Figure 9.8. IMI score as moderated by prior knowledge.

170) = 4.15, p < .001. However, only “Very Confident” trended toward significantly greater posttest scores, β = .83, p = .078. This could suggest that although initial confidence and later IMI score were correlated (r = .329, p < .01), the construct was not a clear a moderator. When allowing the effect of lesson to vary across prior knowledge, the model remained significant, R^2 = .146 F(14, 162) = 2.28, p < .01. Significant differences were not observed, as shown in Table 9.5. Thus, initial confidence was not a significant moderator of the effect of lesson on IMI score. This finding is depicted visually in Figure 9.9.

Table 9.6. Lesson placement, initial confidence, and interactions predicting posttest and IMI Score.

		Posttest Score				IMI Score			
	Path	Coeff.	SE	t	p	Coeff.	SE	t	p
Model 1									
Constant	i_Y	.68	.02	27.99	< .001	4.08	.15	27.40	< .001
Delayed Lesson	b_1	.05	.04	1.21	.226	.44	.24	1.85	.066
Upfront Lesson	b_2	.07	.04	1.61	.108	.28	.23	1.25	.215
Model 2									
Constant	i_Y	.49	.06	7.89	< .001	3.96	.45	8.82	< .001
Delayed Lesson	b_1	.04	.04	1.03	.306	.42	.23	1.81	.072
Upfront Lesson	b_2	.04	.04	1.08	.281	.05	.22	.24	.812
A Little Confident	b_3	-.01	.09	-.01	.993	-.16	.54	-.29	.770
Somewhat Confident	b_4	.09	.07	1.24	.215	.07	.47	.14	.886
Pretty Confident	b_5	.22	.07	3.24	.001	-.20	.47	-.44	.663
Very Confident	b_6	.30	.07	4.52	< .001	.83	.47	1.78	.078
Model 3									
Constant	i_Y	.52	.08	6.69	< .001	3.90	.71	5.51	< .001
Delayed Lesson	b_1	.04	.14	.26	.797	.57	1.00	.57	.572
Upfront Lesson	b_2	-.30	.23	-1.28	.202	.05	1.12	.05	.964
A Little Confident	b_3	-.11	.12	-.95	.341	-.67	.85	-.79	.429
Somewhat Confident	b_4	.09	.09	.96	.340	.25	.75	.33	.741
Pretty Confident	b_5	.20	.09	2.27	.024	-.10	.75	-.13	.894
Very Confident	b_6	.26	.09	2.94	.004	.88	.77	1.14	.255
Delayed X A Little	b_7	.03	.20	.17	.865	1.91	1.31	1.45	.148
Delayed X Somewhat	b_8	-.06	.16	-.35	.729	-.61	1.10	-.55	.584
Delayed X Pretty	b_9	.02	.15	.15	.878	-.28	1.07	-.26	.792
Delayed X Very	b_{10}	.02	.15	.15	.881	-.14	1.10	-.13	.899
Upfront X A Little	b_{11}	.58	.27	2.15	.032	.48	1.33	.36	.718
Upfront X Somewhat	b_{12}	.28	.25	1.13	.258	-.11	1.20	-.09	.929
Upfront X Pretty	b_{13}	.28	.25	1.13	.261	.01	1.21	.01	.993
Upfront X Very	b_{14}	.39	.24	1.60	.111	.00	1.19	.00	1.000
Model 1		$R^2 = 0.008$, MSE = 0.108				$R^2 = 0.021$, MSE = 1.640			
Model 2		$R^2 = 0.121$, MSE = 0.097				$R^2 = 0.128$, MSE = 1.495			
Model 3		$R^2 = 0.139$, MSE = 0.097				$R^2 = 0.164$, MSE = 1.503			

9.4 Discussion

The present work considered the efficacy of a worked example skill lesson as a motivational intervention targeting students' perceptions of competence, interest/enjoyment, and subsequent performance on a near and far transfer posttest. It was hypothesized that skill lesson placement would affect posttest and IMI scores, which was not proven. Whether students received a lesson upfront before practicing fluency problems, during struggle with their assignment, or not at all, did not ultimately influence posttest or IMI scores. Further, initial confidence served as a significant moderator of the effect of lesson on posttest score. Students with lower initial confidence performed poorly with an Upfront lesson, which is a counterintuitive finding that may have been

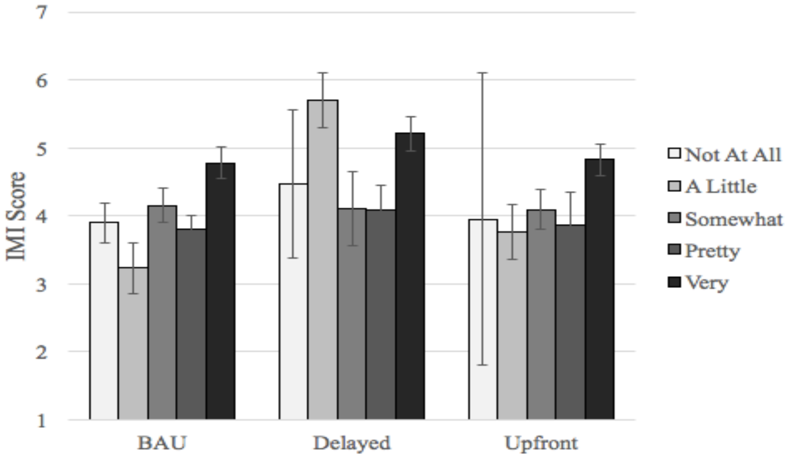


Figure 9.9. IMI Score as moderated by initial confidence.

observed from low subsample sizes. Alternatively, this finding may be explained by motivation-less approaches to assignment completion like gaming the system. It is possible that students with low confidence in their abilities gave up early and got through the assignment by any means necessary. Although initial confidence was a significant predictor of later IMI scores, it was not a significant moderator of the effect of lesson on IMI score. This finding is also counterintuitive, but likely suggests that although a relationship exists between confidence and IMI score (the effect differs from zero), confidence levels did not differ significantly (group slopes did not differ significantly).

Contrary to previous work on the efficacy of prior percent correct as a measure of ‘global’ prior math knowledge, prior knowledge was not a significant moderator of the effect of lesson on posttest score. However, it offered a useful control for the regression analysis, again suggesting the difference between the research questions underlying a main effect and moderation. Visualizations of simple effects suggested that although prior knowledge was not a significant moderator, low knowledge students with upfront lessons showed stronger performance on posttest. In addition, prior knowledge was not a significant moderator of the effect of lesson on IMI score.

The limitations of this work were plentiful. The most pressing issue was the lack of experienced condition for the Delayed lesson. Although 116 students were randomly assigned to receive a lesson in the presence of struggle, the definition of struggle applied was too flexible. Students were ultimately able to master the Skill Builder within 6 questions, and only three students actually “struggled” and received the delayed lesson. Given that much of sample was representative of 7th grade, while the skill in question was representative of 6th grade, it is also possible that ceiling effects were to blame for the null effects of lesson placement.

In addition, on demand hint feedback likely added noise to the effect of lesson placement. Lesson feedback was modeled on hint feedback, which followed the same worked example format first before showing students a worked solution to the current problem. As such, an effect of lesson placement was not observed in the present analyses. Combined with the influence of low sample size in the delayed lesson condition, it is possible that skill lessons may actually influence subsequent performance and measures of intrinsic motivation, but that the present study was ill defined to observe such effects.

The ultimate goal of future work will be to redefine the concept of struggle and to find a more difficult skill to focus the effects of a skill lesson intervention. If the effect of a lesson can be observed using another iteration of this work, it will be useful to examine how moderators, and perhaps mediators, influence and explain that effect. With largeer sample sizes and stronger fidelity, it may also be possible to examine a moderated moderation in which prior knowledge moderates initial confidence, which then moderates the effect of lesson placement on posttest or IMI scores.

References

ASSISTments TestBed Resource Guide. (2018). Retrieved: https://www.assistmentstestbed.org.

Baker, R. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*. 26 (2): 600-614.

deCharms, R. (1968). Personal causation. New York: Academic Press.

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. New York: Plenum.

Harter, S. (1978). Effectance motivation reconsidered: Toward a developmental model. *Human Development*. 1: 661-669.

Hayes, A.F. (2018). Introduction to Mediation, Moderation, and Conditional Process Analysis. The Guilford Press. New York, NY.

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. *International Journal of AIED*. 24(4): 470-497.

Intrinsic Motivation Inventory. (2018). From: http://selfdeterminationtheory.org/questionnaires/.

Lazowski, R. A. & Hulleman, C. S. (2016). Motivation Interventions in Education: A Meta-Analytic Review. *Review of Educational Research*. 86 (2): 602-640.

Maslow, A. (1943). A theory of human motivation. *Psychological Review*. 50: 370-396.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). Common Core State Standards. Washington, DC: Authors.

Niemiec, C. P. & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. *Theory and Research in Education*. 7 (2): 133-144.

Ostrow, K. S. & Heffernan, N. T. (2018). Confidence and Prior Knowledge as Moderators of Lesson Placement on Performance and Intrinsic Motivation: Submission Data Page. www.tiny.cc/EDM2018submission, last accessed 2018/03/07.

Ostrow, K. S. & Heffernan, N. T. (In Press). Testing the Validity and Reliability of Intrinsic Motivation Inventory Subscales within ASSISTments. To be included in the Proceedings of the 2018 International Conference of AIED.

Ostrow, K., Heffernan, N. T., Heffernan, C., & Peterson, Z. (2015). Blocking vs., Interleaving: Examining Single-Session Effects within Middle School Math Homework. In Conati, Heffernan, Mitrovic & Verdejo (eds.) Proceedings of the 17th International Conference on Artificial Intelligence in Education. Springer International Publishing. 388-347.

Ostrow, K., Selent, D., Wang, Y., Van Inwegen, E., Heffernan, N., & Williams, J. J. (2016). The Assessment of Learning Infrastructure (ALI): The Theory, Practice, and Scalability of Automated Assessment. In Proceedings of the 6th International Conference on Learning Analytics and Knowledge.

Razzaq, L. & Heffernan, N. (2009). To Tutor or Not to Tutor: That is the Question. In Dimitrova, Mizoguchi, du Boulay & Graesser (Eds.) Proceedings of the 2009 Artificial Intelligence in Education Conference. IOS Press. 457-464.

Roschelle, J., Feng, M., Murphy, R. & Mason, C. (2016). Online Mathematics Homework Increases Student Achievement. *AERA OPEN*. October-December 2016. 2 (4): 1–12.

Ryan, R. M., Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist.* 55: 68-78.

Ryan, R. M., & Deci, E. L. (2002). An overview of self-determination theory. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research*. Rochester, NY: University of Rochester Press.

Schell, K. L., Oswald, F. L. (2013). Item grouping and item randomization in personality measurement. *Personality and Individual Differences*. 55: 317-321.

Vygotsky, L. (1978). Interaction Between Learning and Development. In Gauvain & Cole (Eds.) Readings on the Development of Children. New York: Scientific American Books. 34-40.

White, R. W. (1963). *Ego and reality in psychoanalytic theory*. New York: International Universities Press.

Synthesis

This synthesis is intended to summarize Parts I & II to reiterate how my work has helped the realization of research at scale within ASSISTments. Commentary is provided for each chapter in Parts I & II, with summaries and explanations of how lessons learned led to personal growth, growth for ASSISTments Research, and broader contributions to the Learning Sciences.

Evolution

Part I of this work described the evolution of two of the primary research tools underlying ASSISTments Research: the ASSISTments TestBed and the Assessment of Learning Infrastructure (ALI).

Chapter 1 presented the seminal journal article detailing the early stages of the ASSISTments TestBed and advocating for other online learning platforms to harness their data and to serve as tools for transparent educational research at scale. The article was essential to the progress of the NSF SI2 grant (SI2-SSE&SSI: 1440753, Adding Research Accounts to the ASSISTments' Platform: Helping Researchers Do Randomized Controlled Studies with Thousands of Students). This work will likely continue to hold its weight as the idea of pushing the boundaries of educational research gains visibility in the Learning Sciences community. As one of my first journal articles, this paper was an exceptional learning opportunity for my personal growth. It offered experience in writing about past work as part of a larger framework, through the conceptualization of the cycle of perpetual evolution. This exercise was critical in stepping back and assessing the potential power of ASSISTments Research when scaled, allowing me to better understand how rapid, iterative research ebbs and flows in an open system. By assessing the longitudinal effects of the platform as a shared scientific tool, this work also helped establish the team's goals for the future.

Chapter 2 presented the seminal conference paper on ALI that I presented at the 6th International Conference on Learning Analytics and Knowledge (LAK 2016). This work focused on the theory and practice of automated and anonymized data reporting for research conducted using the ASSISTments TestBed. This was also the first work that used ALI analytics to simultaneously consider the data of multiple experiments, paving the way for later datamining endeavors in ASSISTments. This work also advocated support for the NSF SI2 grant (NSF SI2-SSE&SSI: 1440753). Personally, this work provided an excellent opportunity to strengthen my presentation and networking skills in Scotland. It was also a heavily collaborative effort that allowed me to serve more of an editorial role, offering experience in unifying the work of others through a single cohesive tone. The paper also provided excellent experience in writing about a moving target; ALI was evolving under our noses and rapid iteration through least viable products made it difficult to present a concise picture of the tool's value. With regard to lessons learned for ALI, this work was important in helping the ASSISTments Research team to define the important features of the tool and agree upon decisions for its future dissemination.

Chapter 3 featured the abstracts of a set of published work that originated from significant collaborations during my Ph.D. candidacy. Five randomized controlled trials and three data mining explorations were presented with commentary on lessons learned and how each piece of work afforded iterative improvements to the tools underlying ASSISTments Research. Ultimately, the improvements brought about by these exemplary works involved strengthening the accessibility and universality of the team's research and reporting efforts. Common themes included providing more data through covariates while making sure that the data supplied is accurate and interpretable, using stronger organization and supplemental programs to structure and restructure the data for optimized versatility, and cutting back on the types of research that are supported given the strain on internal resources and the common tendency for researchers to push the system's limits. As such, these themes should be maintained when looking to the future. Personally, this chapter was surprisingly difficult to write. Although it was pleasing to look back over past work and watch the quality of my contributions improve over time, it was difficult to separate precise improvements linked to each piece of work. Much of the work occurred within the same time period, with considerable overlap in observed influences. For instance, the results of a randomized controlled trial grew into explorations on partial credit that then became valuable data mining research about the consequences of data flexibility. This was a nice reminder that even though the balance of my work has felt somewhat disconnected during my candidacy (as the 'odd egg' spread across a wide variety of projects), an underlying flow of progress has linked my theories and observations.

Chapter 4 offered my future goals for ASSISTments Research illustrated by the NSF funded Data Infrastructure Building Blocks grant that I organized with Neil Heffernan. The grant (1724889 - CIF21 DIBBs PD: Enhancing and Personalizing Educational Resources through Tools for Experimentation, $494,644) was awarded to PI Heffernan for work slated to span three years (2017-2020). Goals of the proposed work included restructuring the TestBed workflow and ALI output to streamline the research process using experimental design templates. Ideally, researchers will be able to select from a set of experimental designs, easily implement their intervention content, and receive detailed reports featuring automated default analyses. Given slow uptake of proposed goals in the 2017-2018 year as the team awaits completion of the new version of ASSISTments (being released in late 2018), it is likely that this work will carry over into 2021 through a no-cost extension. I anticipate remaining with the ASSISTments team upon completion of my Ph.D. as a Research Scientist, serving as a Co-PI on this grant and driving the immediate future of ASSISTments Research in other capacities through continued advocacy, research publication, and evolution. As such, I suspect that the broader impact of this work as well my personal lessons learned will continue to evolve well into the future.

Application

Part II of this work described my personal application of ASSISTments Research by taking readers on a journey through the framework of Self-Determination Theory (SDT). Chapter 5 gathered pertinent background information on SDT and related theories of motivation to preface Part II with a broad context and framing. Chapters 6-9 presented a scale validation exercise and three randomized controlled trials that were all proposed, run, and reported on within a single calendar year. Summarized goals, contributions, and lessons learned are presented in the following subsections.

Scale Validation

Research Goals & Contributions

Chapter 6 described a scale validation exercise intended to assess the reliability of pertinent Intrinsic Motivation Inventory subscales within ASSISTments in support of a series of randomized controlled trials being conducted to assess the efficacy of learning interventions aligned to the framework of SDT. The work considered four of the seven subscales: Interest/Enjoyment, Perceived Autonomy, Perceived Relatedness (Belonging), and Perceived Competence. High reliability was achieved for the Interest/Enjoyment, Perceived Autonomy, and Perceived Competence scales. Reliability was increased through iterative item reduction, providing support for use of these scales in the randomized controlled trials presented in Chapters 7 and 9. The Perceived Relatedness (Belonging) subscale did not achieve high enough reliability for its later use; instead, the PSSM was validated and applied in Chapter 8.

Lessons Learned

This work was intended to remind the Learning Sciences community that validating a measure in a learning environment before its formal use strengthens the validity and reliability of resulting claims. The paper was accepted to the 19th International Conference on Artificial Intelligence in Education (AIED 2018). Although scale validation practices are commonplace in experimental psychology fields, they are less prevalent in the new age of educational research involving big data and online contexts. As such, advocating the use of these methods among the Learning Analytics and Learning Sciences communities promotes rigor and improves the quality of resulting claims.

For ASSISTments Research, this work was a useful example in the potential benefits of scale reduction exercises. The reduction of nine items across subscales will make future data collection using the IMI less obtrusive, thereby allowing survey efforts to scale to new contexts more efficiently. Similar approaches should be used to address the psychometric properties of other socio-emotional scales prior to their implementation within ASSISTments. Future work should also consider the validity of such scales by assessing how response scores correlate with other measures of performance.

Personally, this work felt like a big risk but led to a big reward. The project was completed within ten days, from conceptualization through to data collection, analysis, and reporting. The work applied the ASSISTments Research tools, integrated with Qualtrics through the ASSISTments Survey Service, to access assignment and survey data. This work was somewhat of a whim, intended to support the remainder of Part II. It involved learning and applying new methodologies with the goal of peer review; I did not anticipate that it would be so well received. Ultimately, this experience taught me to trust myself and to believe in my abilities.

Autonomy

Research Goals & Contributions

Chapter 7 described a randomized controlled trial designed to target the basic psychological need of autonomy, thought to be the most critical need for proper adjustment and well-being within the context of Self-Determination Theory. The work was rooted in prior research conducted in ASSISTments on the effects of supporting autonomy through choice and affecting learning outcomes by altering hint medium. Results suggested that video hints were associated with stronger performance and greater motivation among low knowledge students, especially in the presence of support for autonomy. As such, providing the option of video hints offers a low-cost, low-effort learning intervention within ASSISTments that appears to target the achievement gap. Future work will be necessary to assess if the observed effect can be compounded across assignments and to better understand mechanism underlying observed efficacy.

Lessons Learned

This work offers the Learning Sciences community an example of strong experimental research conducted within an online learning environment that isolated what appears to be an effective learning intervention for low knowledge students. These students are more likely to be labelled as "at risk," and are commonly the focus of research in the community. Related fields are rife with research addressing the achievement gap through cognitive, socio-emotional, and behavioral interventions, often requiring parental involvement or expensive, school-wide reform. This work may have a strong impact by instead offering evidence that small interventions tailored to individual assignments can be just as effective. As such, this chapter will be submitted to a journal venue following completion of my candidacy.

I see this work, in part, as a vision of the potential future of ASSISTments Research. The experimental design was strong but common, employing a three-way independent factors ANOVA that could easily be developed into a design template under DIBBs funding. The analyses were somewhat rote, suggesting that partial automation of output organization would be feasible for the team. Any planned automation should be confirmed by researchers in advance of data collection in order to reduce Type I error. This extra step would relate to efforts for pre-registration and data transparency. In addition, raw data files should always accompany any automation by default.

Personally, this article was frustrating but rewarding. It was easily one of the strongest studies I have designed during my candidacy, with a great deal of effort applied to the delivery of matched content hints and the development of the assignments' structure. With each new iteration of video vs. text work I have learned new audiovisual techniques for video implementation, with this work representing the cleanest version yet. I suspect that the effect of video may be mediated by the video's style and accessibility, as past work has shown inconsistency in video design, hosting server, and ultimately, effect. As such, I have learned that replication is a very complex business, but well worth the efforts. In many ways, this work also epitomizes my growth over the last five years, with direct comparisons to my first published study in 2014.

Belonging

Research Goals & Contributions

Chapter 8 described a randomized controlled trial designed to target the basic psychological need of relatedness or belonging as defined by the context of Self-Determination Theory. The experiment manipulated the presence and timing of a value affirmation intervention and examined effects across levels of students' baseline belonging and prior knowledge. Despite collecting data for a full academic year, sample sizes were relatively small, and analyses were underpowered. However, results revealed limited evidence that the value affirmation was associated with significantly stronger academic performance at posttest. This intervention was especially promising for low knowledge students, suggesting another low-cost possibility for addressing the achievement gap. Future work will be necessary to assess the potential for recursive value following repeated intervention.

Lessons Learned

This work offers the Learning Sciences community another example of an effective learning intervention in an online context with disproportionate effects on low knowledge students. Coupled with the positive effects of support for autonomy observed in Chapter 7, the positive effects of support for belonging suggest that the framework of Self-Determination Theory should be seriously considered in the design and development of online learning environments. When properly devised, these environments can serve as valuable resources in supporting the basic psychological needs of students.

For ASSISTments Research, this study highlighted an interesting predicament regarding content advertisement and sample accessibility. This work was developed by aggregating usage data from the previous year to select popular content with characteristics of high difficulty (i.e., a high prevalence of hint usage and a lack of ceiling effects in mastery rates). Based on past usage, it was theorized that this study would collect 400-500 students during the spring of 2017. However, the study ran for a full calendar year and only achieved a sample size of 378. Although this was possibly due to embedded survey content altering how teachers perceived and therefore assigned content, reductions in uptake also potentially suggest that using past performance data to influence content selection decisions may be poor practice. Researchers should instead be advised to select content based on their theoretical framework and they should be notified that samples may take anywhere from 3-12+ months to accrue when experiments are embedded in Skill Builders. In addition, this study implemented opt-in to a long survey (18 items) at the start of the assignment, causing higher attrition than normal. Scale reduction techniques such as that featured in Chapter 6 will likely prove valuable in retaining users in studies implementing similar survey content.

Personally, this study was my gateway into the framework of Self-Determination Theory. Originally developed for a class project in Spring 2017, this portion of Part II took the longest to develop and experienced the most revision. Research evolved during data collection, and I learned to trust my own instincts for data analysis and interpretation. While working on this chapter I learned how to use G*Power to conduct a priori and post hoc power analyses, strengthening my analytic skills. This work also reminded me that overly complex research designs can show ignorance; simple designs afford higher power, make for easier interpretation, and isolate effects more readily than convoluted designs.

Competence

Research Goals & Contributions

Chapter 9 described a randomized controlled trial designed to target the basic psychological need of competence within the context of Self-Determination Theory. This work considered the efficacy of a worked example skill lesson as a motivational intervention targeting students' perceptions of competence, interest and enjoyment, and subsequent performance on a near and far transfer posttest. Results suggested that the skill lesson was an ineffective intervention, regardless of its placement within the assignment (immediate or delayed). Interestingly, students' confidence significantly moderated their later posttest and IMI scores, but significant moderation was not observed based on prior knowledge levels. Coupled with low sample sizes and ceiling effects, results suggested that the intervention was not an effective means of improving perceived competence and therefore had little effect on resulting performance or motivation.

Lessons Learned

This chapter provided a valuable warning for the Learning Sciences community in the form of a missed opportunity for the practice of pre-registration. My proposed goal was to preregister all studies in the application portion of this dissertation with the Open Science Foundation (osf.io). This was much more difficult in practice than one would expect, and preregistration status varied significantly across studies. This variation was a factor of both my personal confidence in study design (i.e., I seemed less likely to follow through with all steps of preregistration when studies used methodologies I was less familiar with as it made the stakes seem high) and the timing and structure of authoring four large projects simultaneously (i.e., I seemed less likely to follow through with preregistration when data processing and analysis began before formally structuring an introduction and methods section). My research tendencies have been fairly exploratory, as is often the case among data miners using the big data environments afforded by online learning platforms. If this work had been successfully preregistered, it may have had a chance in publication. However, low sample sizes, lack of fidelity, and ultimately a failed intervention will likely cause this project to end up in the "file drawer."

For ASSISTments Research this work offered insight into risks inherent to common experimental designs. Targeting struggling students caused low power from low treatment fidelity.

Recommending a lower threshold for ‘struggle’ may help resolve this issue in future work. In addition, the skill lesson intended to improve performance was closely related to traditional hint content accessible by all students, allowing a high potential for washout. This was a poor design and a lesson learned but emphasized the importance of rapid iterative research made possible by the ASSISTments TestBed. Another issue for the ASSISTments Research team to consider is potential discrepancy in material used for new learning and material used for review. Ceiling effects may be observed when material is used for review (i.e., in this work, the majority of the sample was labelled as 7th grade but the skill targeted 6th grade learning). Including flags in ALI to parse new learning from review would likely serve as a powerful covariate in future analyses.

Personally, this chapter represents a step out of my comfort zone with regard to data analysis. I took the risk of learning a new method of analysis just a few days before submitting the chapter for consideration at EDM 2018. My research questions drove me to learn PROCESS, an SPSS plug-in created by Andrew Hayes to conduct complex mediation, moderation, and conditional process analyses. My initial goal was to assess a moderated moderation in which prior knowledge moderated initial confidence, which then moderated the effect of lesson placement on posttest or IMI score. This would have condensed presented models but increased the complexity of model interpretation beyond what the sample size allowed. The chapter is not my strongest work, but I am happy to have started exploration into a new method and tool and I gained valuable feedback from EDM review process.

Broader Limitations & Contributions

It is my hope that this dissertation has provided the Learning Sciences community with strong evidence for both the powers and perils of conducting educational research at scale within an online learning platform. By leveraging the ASSISTments user population, the ASSISTments TestBed and the Assessment of Learning Infrastructure are leading the community as significant resources for changing how we learn about learning. While I have spent years advocating for these tools as part of the ASSISTments Research team, my personal application and paired commentary suggest that the method is not without its own flaws. Natural learning environments are far less controlled than laboratory settings and although the tutor collects a wealth of data, it can leave much to be desired when it comes to interpreting the complexities of a student’s performance. Further, while the approach reduces the cost of rapid, iterative research, issues like slow sample acquisition, ceiling effects, and attrition can be frustrating, demanding time and patience. Still, after putting these tools through their paces, I would argue that while research at scale may not suit the needs of all educational researchers, it definitely offers a powerful alternative in the digital age. Lessons learned from this work will undoubtedly continue to strengthen ASSISTments Research and refine my message of advocacy for educational research at scale.

APPENDIX

A.1. Python code for original program created to transform problem level data from the ASSISTments Data Dumper (prototype of the Assessment of Learning Infrastructure) into a student level file, or a file where each row represents the performance attributes of a single student with features laid out in columns and presented in opportunity order.

```
# ASSISTments.csv reader
# Version 2.2
# Author: Korinn Ostrow

import csv

def extractAllRecords(fileName):
    """
    Extract all CSV records and return as single LIST-of-RECORDS

    Note that the first element in this list contains the
    description of the columns as defined in the CSV file
    """

    file = open (fileName, 'rU')
    reader = csv.reader(file)
    results = []
    for row in reader:
        results.append(row)
    file.close()
    return (results)

def finder(key, row, columns):
    if not key in columns:
        return ('')
    else:
        return (row[columns[key]])


def studentLevel(fileName):
```

```
"""
Transforms database csv file into studentLevel data.
"""

logs = extractAllRecords(fileName)

columns = {}
students = {}
featureList = []

header = logs[0]
for col in range(len(header)):
    if not header[col] in columns:
        columns[header[col]] = col
        featureList.append(header[col])

for idx in range(1, len(logs)):
    row = logs[idx]
    user_id = finder('problem_logs_user_id', row, columns)
    assignment_id = finder('problem_logs_assignment_id', row, columns)
    if assignment_id == '':
        assignment_id = finder('assignment_logs_assignment_id', row, columns)

    if assignment_id != '' and user_id != '':
        UUID = (str(user_id) + '_' + str(assignment_id))
    else:
        UUID = user_id

    if UUID != '':
        if not UUID in students:
            features = {}
            students[UUID] = features

        for key in columns.keys():
            if not key in students[UUID]:
                students[UUID][key] = []
            if finder(key, row, columns) is '':
                students[UUID][key].append('-')
            else:
                students[UUID][key].append(finder(key,row, columns))
    return students, featureList
def transform(fileName):
    """
    Transforms the data pulled using studentLevel(fileName).
    """

    outFile = open(fileName[:-4] + '.processed.txt', 'w+')

    students, featureList = studentLevel(fileName)

    featureDict = {}
    collapsible = {}

    overallMax = 0

    for i in range(len(featureList)):
        featureDict[featureList[i]] = 0

    for entry in students:
        for feature in featureList:
            maxfeature = len(students[entry][feature])
            if featureDict[feature] <= maxfeature:
                featureDict[feature] = maxfeature

            if featureDict[feature] >= overallMax:
                overallMax = featureDict[feature]

            data = students[entry][feature]
            for index in range(1,len(data)):
                first_val = data[0]
                if feature not in collapsible:
                    collapsible[feature] = True
                collapsible[feature] = (collapsible[feature] and (data[index] == first_val))

                if collapsible[feature] == False:
                    break

    true_block = True
    latest_true = 0
```

```
for index in range(len(featureList)):
    feature = featureList[index]
    if true_block == True and collapsible[feature] == False:
        latest_true = index
        true_block = False
    if true_block == False and collapsible[feature] == True:
        featureList.remove(feature)
        featureList.insert(latest_true, feature)
        latest_true += 1


for feature in featureList:
    if collapsible[feature] == False:
        for val in range(1, (overallMax + 1)):
            outFile.write((str(feature) + ' Problem ' + str(val) + ','))
    else:
        featureDict[feature] = 1
        outFile.write((str(feature) + ','))

outFile.write('\n')

for entry in students:
    for feature in featureList:
        cellcount = 0
        max_len = len(students[entry][feature])
        while cellcount < featureDict[feature]:
            if cellcount < max_len:
                outFile.write(str(students[entry][feature][cellcount]))
            else:
                outFile.write('-')
            outFile.write(',')
            cellcount += 1
    outFile.write('\n')

outFile.close()
```

A.2. Python code for original program created to combine (left to right) two files of student level data produced by ALI, where each row represents the performance attributes of a single student with features laid out in columns and presented in opportunity order, with multiple assignments concatenated together across columns.

```
# Combine LtoR Program - Ostrow 2015

import csv

def MasterCombine(ProblemSets):
      """
      Takes a LIST OF STRINGS of file names and returns one master file
      in the form MasterCombine(["part1.csv", "part2.csv", "part3.csv"])
      """

      final = ProblemSets[0]
      ProblemSets.remove(ProblemSets[0])
      for i in range(len(ProblemSets)):
            print (final, ProblemSets[i])
      final = Combine(final, ProblemSets[i])

def Combine(File1, File2):
      """
      Extracts all rows from two csv files as List of Lists (LoL). Reads 1st file as master;
      when 2nd file matches on Column A (user ID), row data from 2nd file is tacked on to row
      data from 1st file (i.e., placing posttest data after practice session data). New file is
      written as combination "File1+File2.csv".
      """

      file = open(File1, 'rU')
      reader = csv.reader(file)
      LoL1 = []
      for row in reader:
            LoL1.append(row)
      file.close()

      file = open(File2, 'rU')
      reader = csv.reader(file)
      LoL2 = []
      for row in reader:
            LoL2.append(row)
      file.close()

      """
      for each row in file 1 if number of columns copied is not equal to max,
      add buffer cells until you reach number. add buffer between files as well?
      """

      LoL1Length = 0
      for row in range(len(LoL1)):
            if len(LoL1[row]) > LoL1Length:
      LoL1Length = len(LoL1[row])

      LoL2Length = 0
      for row in range(len(LoL2)):
            if len(LoL2[row]) > LoL2Length:
      LoL2Length = len(LoL2[row])

      for row in range(len(LoL1)):
            for row2 in range(len(LoL2)):
                  if LoL1[row][0] == LoL2[row2][0]:
            LoL1[row].extend(LoL2[row2])
            break

      for row in range(len(LoL1)):
      while len(LoL1[row]) < (LoL1Length + LoL2Length):
      LoL1[row].extend('-')

      fileName = File1[:-4] + "+" + File2[:-4] + ".csv"
      outFile = open(fileName, 'w+')
      writer = csv.writer(outFile, delimiter = ',', quotechar='"')
      for row in range(len(LoL1)):
            writer.writerow(LoL1[row])
      outFile.close()

return filename
```

A.3. Python code for original program created to combine (top to bottom) two files of student level data produced by ALI, where each row represents the performance attributes of a single student with features laid out in columns and presented in opportunity order, with multiple assignments concatenated together across rows.

```
# Combine TtoB Program - Ostrow 2015

import csv

def MasterCombine(ProblemSets):
      final = ProblemSets[0]
      ProblemSets.remove(ProblemSets[0])
      for i in range(len(ProblemSets)):
            print (final, ProblemSets[i])
      final = Combine(final, ProblemSets[i])

      def Combine(File1, File2):

      """
      Extracts all rows from two csv files as List of Lists (LoL). Reads 1st file as master;
      when 2nd file matches on Column A (user ID), row data from 2nd file is tacked on to row
      data from 1st file (i.e., placing posttest data after practice session data). New file is
      written as combination "File1+File2.csv".
      """

      file = open(File1, 'rU')
      reader = csv.reader(file)
      LoL1 = []
      for row in reader:
            LoL1.append(row)
      file.close()

      file = open(File2, 'rU')
      reader = csv.reader(file)
      LoL2 = []
      for row in reader:
            LoL2.append(row)
      file.close()
      """
      For each row in file 1 if number of columns copied is not equal to max,
      add buffer cells until you reach number. add buffer between files as well?
      """

      LoL1Length = 0
      for row in range(len(LoL1)):
            if len(LoL1[row]) > LoL1Length:
      LoL1Length = len(LoL1[row])

      LoL2Length = 0
      for row in range(len(LoL2)):
            if len(LoL2[row]) > LoL2Length:
      LoL2Length = len(LoL2[row])

      for row in range(len(LoL1)):
            for row2 in range(len(LoL2)):
                  if LoL1[row][0] == LoL2[row2][0]:
                        LoL1[row].extend(LoL2[row2])
      break

      for row in range(len(LoL1)):
            while len(LoL1[row]) < (LoL1Length + LoL2Length):
      LoL1[row].extend('-')

      fileName = File1[:-4] + "+" + File2[:-4] + ".csv"
      outFile = open(fileName, 'w+')
      writer = csv.writer(outFile, delimiter = ',', quotechar='"')
      for row in range(len(LoL1)):
            writer.writerow(LoL1[row])
      outFile.close()

return filename
```