Graspable Math K-12: Perspectives and Design for Formative Assessment of Mathematical Proficiency with Learning Technologies

by

Taylyn Hulse
(trhulse@wpi.edu)

A thesis

Submitted to the faculty

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

In

Learning Sciences & Technologies

February 2019

APPROVED:

Erin Ottmar, Ph.D., Advisor

Ivon Arroyo, Ed. D., Reader

Neil Heffernan, Ph. D., Program Director

Abstract

This thesis grounds the design of learning technologies in cognitive learning theory to explore deeper formative measurement of the learning process. This work implements *Graspable Math (GM;* Ottmar, Landy, Weitnauer, Goldstone, 2015), a dynamic learning technology that has been designed using perceptual-motor learning theory, which has been shown to have a strong connection to mathematical reasoning (Kirshner, 1989; Kellman, Massey, & Son, 2010; Goldstone, Landy, & Son, 2010). With this dynamic mathematics learning technology, we can measure the algebraic problem solving process in ways that are not possible with pencil and paper or other more traditional learning technologies. By collecting this data, this research will explore how to move beyond traditional correctness-based assessment and design more formative measures of the learning process. This work provides a rich perspective on the evolution of research on mathematical proficiency, how this research is applied in practice, and an in-depth example of how one technology-based learning environment has been developed to measure mathematical proficiency. This work has three main objectives: *1) develop a theoretical framework to assess mathematical proficiency within GM, 2) explore GM-based measures of mathematical proficiency across K-12 populations, and 3) design GM-based tools that are grounded in theory on mathematical proficiency.*

This work first presents a conceptual model that maps student behavior data measured through *GM* onto the five theoretical strands of mathematical proficiency as defined by the National Research Council's 2001 publication, *Adding it Up*. The first study reveals underlying constructs in Elementary student data and suggests there is an added benefit of including these formative measures within predictive models. Above and beyond background characteristics and summative measures of knowledge, formative measures of the learning process revealed subtle interactions based on student behaviors and prior knowledge. These constructs also show potential in mapping onto certain strands of mathematical proficiency. The second study compares underlying constructs within Elementary data to High School data using exploratory factor analysis and finds similar factors across both populations. These results suggest that certain constructs may underlie different age groups and have the potential to be used as measures of mathematical proficiency. While the first two sections describe the definition and measurement of mathematical proficiency within GM, the final section explores the implementation of these measures within the design process of new GM-based activities for students and tools for teachers. Ultimately, the goal of this work is to serve as an example method for other researchers, educators, and designers to move beyond summative measures of assessment and enhance the formative assessment capabilities of learning technologies by grounding measures in theories of learning**.**

Acknowledgements

This work would not have been possible without the continued support of my family, friends, and colleagues. I first want to acknowledge my parents, who have always been there to help me reach my dreams. No matter how crazy the idea, they have always found ways to help me thrive. I would also like to thank my grandparents, who have always been supportive of my education and instilled a passion for lifelong learning that has lead me to the learning sciences field. As for Logan, from every corner of my heart, thank you for supporting me each step of the way. You might already be a Master Mechanic, but you're also a Master listener, inspirer, and probably learning scientist from all the conversations you've had with me about education, teaching and learning. Also, thank you for introducing me to the wonderful world of skating. Without this new passion and quirky and welcoming community, I'm not sure I could have stayed sane. We have been exploring all over the world together and I so greatly appreciate your smile and cool attitude through this Master's adventure in New England. We did it!

This experience has given me a new WPI family through the Learning Sciences and Technologies program. Dan, thank you for being my first friend at WPI. You keep us all laughing and grounded in the bigger picture of life. Avery, every day you become more of a soul sister to me. I think the universe has been trying to bring us together and it finally happened. Anthony and Korinn, you have both given me so much wisdom about how to be successful in the field of learning sciences. Hannah, you brought a breath of fresh air to the lab. I love how open and honest you are. Katie, thank you for helping me every step of the way, continually giving me career and life advice, and sharing all of your stories. Jenny, it has been great to have another lover of cognitive research and sunny personality in the lab. Maiya, thank you for sharing laughs and a love for skating. You are such a bright light in the world. Jackson, thank you for always being ready to jump in and help anyone. The wolf pack better prepare itself for you. Maria and Grace, I am so glad that you two were there to be a calming presence and keep us connected to the greater WPI campus.

Thank you to the entire learning sciences faculty and staff for creating a program that is truly interdisciplinary and incredibly collaborative. Tricia has been the backbone of the program and can find a solution to any problem you send her way. Joe always keeps students in mind and puts great thought into his teaching. Neil has been the biggest advocate and an absolute pleasure to work with. Ivon brightens every room she enters and is a true bridge between the psychology and computer science components of the learning sciences. I have been so fortunate to work and present with her, as well as experience her adventurous and playful spirit. Erin has been a mentor to me not only for the Master's, but since I met her many years ago. She naturally fell into the mentorship role and has been there for me through my undergraduate research, traveling in Japan, working at the University of Richmond, and now at WPI. Thank you so much for your continued support and friendship. I am so grateful for the many opportunities you have introduced to me.

Thanks to this community, I feel confident that I belong in the learning sciences with such great people and impactful research. It has been an inspiring journey and I am excited for whatever the future brings.

Table of Contents

Part 1: Theoretical and Conceptual Framework

Integrating learning technologies into schools has become increasingly popular since technologies have become faster, more portable, and more accessible for the classroom. The US Department of Education has spent $6 billion in K-12 and $5.9 billion on Information and Communication Technologies (ICT) in 2009 (Nut, 2010). While government funding is a key component of giving teachers and students access to technologies, it is just the first step. Giving classrooms technologies like chromebooks has not yet proven to be successful in terms of learning gains or increasing motivation and affect (Darling-Hammond, Zielezinski, & Goldman, 2014; Gülbahar, 2007). This is due in part to implementation issues in terms of usability, content relevance, and theoretical grounding. Once we have a better understanding of how students learn the content and how to support and measure that learning, we can then design more effective technologies that meet student and teacher needs in the classroom

There are many potential benefits to learning technologies for students and teachers alike (Figure 1). Learning technologies can make learning more personalized, connected, and mobile for students (Johnson, Pavleas & Chang, 2013). These dynamic environments have the potential to increase engagement for students who might not otherwise be interested by providing them with adaptive scaffolding, timely feedback, and more adaptive content not offered by traditional, paper-and-pencil-based summative assessments (Romero & Ventura, 2010; Foster, 2008). For teachers, these technologies instantly collect student data and can automatically grade and display student progress reports, which can then be used to inform instruction. Frequently using formative assessments to adapt teacher instruction has been shown to improve student achievement (Bergan, Sladeczek, Schwarz, & Smith, 1991; Speece, Molloy, & Case, 2003). Using learning technologies can amplify teacher abilities (Baker, 2016), such as increase on-on-one time while other students are working on their devices (Schofeld, 1995; Holstein, McLaren, & Aleven, 2018). Learning technologies can ultimately enhance classroom management by individualizing student learning and improving teacher knowledge about their students' learning. In turn, teachers can use this knowledge to inform instruction and feedback to students. In order to transform education with learning technologies, their design needs to be grounded in learning theory and take into account the relationship between the student, teacher, and the device.

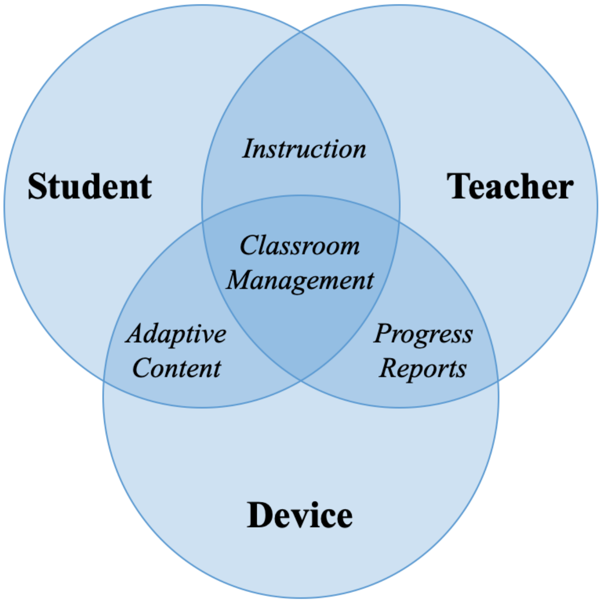


Figure 1-1. *Benefits of learning technologies described through the relationships between the student, teacher, and device.*

This thesis grounds the design of learning technologies in cognitive learning theory to explore deeper formative measurement of the learning process. This work implements *Graspable Math (GM;* Ottmar, Landy, Weitnauer, Goldstone, 2015), a dynamic learning technology that has been designed using perceptual-motor learning theory, which has been shown to have a strong connection to mathematical reasoning (Kirshner, 1989; Kellman, Massey, & Son, 2010; Goldstone, Landy, & Son, 2010). With this dynamic mathematics learning technology, we can measure the algebraic problem solving process in ways that are not possible with pencil and paper or other learning technologies that give simple multiple choice or type-in answers. By collecting this data, this research will explore how to move beyond traditional correctness-based assessment and design more formative measures of the learning process.

Part 1 of this thesis reviews the importance of using learning technologies in mathematics education to foster algebraic understanding, in particular, as it is the foundation for higher mathematics. The review also discusses mathematical proficiency, how it has been defined, and how it has evolved into the current common core standards for mathematics. It culminates in a definition of mathematical proficiency situated within a mathematics education and cognitive psychology perspective that focuses on specific skills that students develop over time across problem types and processes. Lastly, this section explains the theoretical framework behind *Graspable Math (GM)*, the sole technology being researched in this thesis. Together, the first section set up the stage for the current work, which aims to take a technology that is already grounded in theory for teaching and learning mathematics, and develop a theoretical framework to assess that learning so it can be utilized by teachers to inform instruction.

Part 2 of this work explores how to utilize *GM* as a tool for mathematics practice and for the measurement of mathematical proficiency. Rather than focusing on learning outcomes, this section explore the benefits of using formative measurement to analyze the learning process. Two studies test the feasibility of *GM* in an Elementary and High School population. These studies also analyze and compare latent constructs of problem solving behavior in both populations as measured by *GM*. The benefits and pitfalls of this method of formative assessment is discussed in terms of the relationship between the revealed constructs and theoretical framework of mathematical proficiency.

Part 3 of this work utilizes the theoretical framework and study results from sections 1 and 2 to inform the development of new *GM*-based activities. These new activities serve as example method of how to design technology-based learning environments with a goal of measuring mathematical proficiency. This section describes the iterative design process to create technology-based activities that have the potential to measure the five components of mathematical proficiency. This highlights the twists and turns of the development process all the way through to a final pilot study which implements the technology into high school classrooms. The goal of this section is to give practitioners and researchers a concrete example of applying research into practice via learning technologies. The final chapter discusses related and future work that applies *GM* in the context of formative assessment and discusses its potential in terms of implementation in the classroom.

Together, these three sections provide a rich perspective on the evolution of research on mathematical proficiency, how this research is applied in practice, and an in-depth example of how one technology-based learning environment has been developed to measure mathematical proficiency. Ultimately, the main goal of this work is to explore student problem solving behavior within *GM* in order to tease apart components of mathematical proficiency more efficiently and at a deeper level than possible with traditional summative assessment. In order to accomplish this, this work has three objectives:

1) develop a theoretical framework to assess mathematical proficiency within GM
2) explore GM-based measures of mathematical proficiency across K-12 populations
3) design GM-based tools that are grounded in theory on mathematical proficiency

Chapter 1: Literature Review

This work is situated in the intersection of mathematics education, learning sciences, and educational technologies. The first objective of this work, to ***develop a theoretical framework to assess mathematical proficiency within GM***, is addressed in the first two chapters. This review first highlights the importance of algebraic reasoning in the context of mathematics education and the evolution of defining mathematical proficiency. Then, it explores the benefits of learning technologies in K-12 populations, including their potential role in formative assessment. Lastly, the review introduces the theoretical framework of *Graspable Math* (*GM*), the primary learning technology used throughout this work. Ultimately, this first section develops a conceptual framework to ground *GM* in theory on the assessment of mathematical proficiency. This method of theoretical and conceptual framing could serve as an approach for grounding any learning technology within teaching and learning theory.

Mathematics Education and Mathematical Proficiency

Only 33% of 8th grade and 25% of 12th grade students in the US are proficient in mathematics (McFarland et al., 2017). In an aim to define and better understand student proficiency in mathematics, the primary aim of this research is to explore the use of dynamic learning technologies in K-12 populations to formatively measure and assess mathematical proficiency. More specifically, this work focuses on algebraic learning as it is seen as the foundation of higher mathematics and the stepping stone from arithmetic to algebraic generalizations (Carraher, Schliemann, Brizuela, & Earnest, 2006). This switch from concrete numerical symbols to abstract variables, typically faced in middle school, is notoriously difficult, likely due to increased abstractness and reliance on a deep conceptual knowledge of symbolic structure and equivalency (Booth, Barbieri, Eyer, & Pare-Blagoev, 2014; Koedinger & MacLaren, 2002; Kaput, 1998). However, Many students demonstrate inflexibility with algebraic structure and struggle to understand which strategies are appropriate when presented with non-standard equations (Sfard & Linchevski, 1994; Jiang, Cooper, & Alibali, 2014). Students also often become frustrated and disengaged with mathematics and never master fundamental algebraic skills (Stein, Kaufman, Sherman, & Hillen, 2011). Algebra is a precursor and a strong predictor of success in advanced mathematics courses, so it is crucial for students to have access to algebra courses, to acquire the fundamental skills, and to have a positive experience for future motivation, engagement and performance in higher mathematics (Adelmann, 1999; Nord et al, 2011).

Research posits that the deficit in mathematical and algebraic understanding begins to arise as students enter the transitional shift between concrete representation of numbers and abstract conceptualization. This may occur due to a lack of understanding of number sense and the ability to see the flexibility and fluidity of expressions through operations (Kalchman et al., 2011). One of the most important developments in children's mathematical thinking is number sense, or flexibility in thinking about numbers (NCTM, 2000). This involves being able to understand how to represent numbers in different ways, understand the size of numbers, and understand how different operations will impact the transformation of numbers (Sowder, 1992). There are specific misconceptions and difficulties that students struggle with, namely, the overall understanding of order of operations, the use of parentheses within an expression, and the concept of equivalence (use of the equal sign) (Knuth et al., 2006; Welder, 2012; author). For example, children often do not understand that parentheses also function as a multiplicative indicator as well as an organizational tool. As an example, the value of 18 may be written as 3x6 or 3(6). Importantly, 18 can also be written using multiple combinations of operations and symbols, like this: (3+17) - 2. Children who do not have a solid understanding of the order of operations would likely struggle to determine the appropriate order in which they could solve the expression, making complex math expressions that require multiple operations nearly unsolvable. These difficulties continue throughout schooling, with order of operations being noted as a major area of confusion for students learning algebra (Welder, 2012).

Students also struggle with decomposition, or the ability to recognize and that any number can be broken down many combinations of other numbers (Clements & Sarama, 2007). Decomposition as a math tactic is defined as the understanding that numbers are made of many different components, and may be rearranged in a way that makes the most sense to the student (Clements & Sarama, 2007). When considering decomposition, students begin with a single number and are asked to explore its properties, for example, "*what two numbers can make 10?*" Inclination and tactics of decomposition are taught as early as kindergarten, as teachers see the meaningful action behind children understanding grouping, relationships, and patterns. Acting as a springboard for children's math understanding at an early age, decomposition is imperative to understanding progressively more formal mathematical learning such as algebra. According to the NCTM Principles and Standards (2000), "students should be able to compose and decompose two- and three-digit numbers" by the second grade. When students have a solid base in number sense and decomposition, they are more likely to be successful with algebra (VanDerHeyden & Burns, 2009). Decomposition allows students to see various ways for them to approach problems (ex. 4 x 6 = 24 replace 6 to 4+2 to see 4 x (4 + 2) = 24, maintaining the same value about the equal sign). However, many times decomposition tasks only involve one operation and often this skill is not explicitly taught in relation to algebra despite its position as a fundamental algebraic concept (Clements, 2000).

Equivalence in mathematics is also noted as a rudimentary foundation of algebra, and relies on strong quantitative skills fostered in early elementary mathematics teachings (Knuth et al., 2006). For instance, children when presented with 3 + 3 = 4 + 2 instead of 3 + 3 = 6 and 4 + 2 = 6, may begin to understand the flexibility of the equivalence notation rather than view it as a rigid obstacle (i.e. the number to the right of the equal sign does not need to be the expressions definitive answer). The notation of equality and its role, is fostered in students understanding of the symbolism of equivalence, rather than as a directional symbol or one that separates problem from answer (Welder, 2012). This understanding becomes critical in algebraic understanding as students must be able to correctly interpret equal sign and view its relation and equivalence (Knuth et al., 2006; Welder, 2012). If provided early, exposure to not only decomposition of the expressions numbers, but also flexibility about the equal sign, may help increase mathematical understanding. Through the introduction and exposure of critical algebraic reasoning and fundamental concepts at earlier ages, children are provided the necessary tools to succeed in algebraic and future math conceptualization. Students who are successful in learning algebra progress through a series of conceptual steps that can be more precisely defined as number sense, representation, fact families, and (most importantly) decomposition (VanDerHeyden & Burns, 2009).

It is upon this foundation that the learning of algebraic ideas is built in middle and high school. By following the natural development of number sense and cardinality, interventions that begin with building a solid foundation of number sense and the concrete properties of numbers may result in improved mathematical understanding. However, it is likely that the deficit in algebraic performance and formal math understanding stems from both lack of exposure to algebraic concepts and misconceptions that develop early, in critical windows where students form the foundations of math understanding. To better prepare students for future algebraic learning, some researchers suggest introducing algebraic concepts in early elementary school (Blanton, Stephens, Knuth, Gardiner, Isler, & Kim, 2015; National Council of Teachers of Mathematics (NCTM), 2000). Children begin to develop the ability to reason algebraically even before they begin formal schooling (Doig & Ompok, 2010); developmentally, many students are certainly capable of learning algebraic ideas early, provided that the topic is scaled down to meet their skill level (Bay-Williams, 2001; Carpenter, Levi, Franke, & Zeringue, 2005; Carraher, Schliemann, Brizuela & Earnest, 2006). By exposing children to algebraic ideas earlier, as students progress in their mathematical thinking, they may be better prepared to learn more difficult concepts down the road (Koedinger, Alibali, & Nathan, 2008; Bransford & Schwartz, 1999; NCTM, 2000).

To understand algebraic thinking within a younger age group, it would be beneficial, from a research perspective, to develop a theoretical framework to define proficiency in early mathematics. At a basic level, mathematical proficiency are the skills and practices needed to be "learn mathematics successfully" (NRC, 2001, p. 5). These skills and practices, however, are defined differently based on the goals and perspective of the person who is measuring them. While teachers and students may want to identify what students know and do not know in order to identify areas of need, mathematicians may want to know more about the broader spectrum of math content and processes students engage in, policy makers may want indicators of how well the school system is doing, while test developers may want to know about the validity of the psychometric properties of the assessment itself (Schoenfeld, 2007). All of these goals are important and make an impact on each student's education. A key issue with high stakes testing is that it has an different overarching goal compared to classroom teaching. This often plays out in the difference between measuring students' ability to perform calculations on demand, compared to measuring their deeper, conceptual knowledge (Ramaley, 2007). The first measurement has its place as it often has high test-retest reliability, meaning that students will perform similarly on the test one week compared to the next. This is great for making assessments to compare students across the nation. However, the latter measurement is necessary for identifying valid constructs of deeper learning and ensuring that students learn mathematics for application in the real world. This work uses a definition of assessment from Niss and colleagues (1998), which "*refers to the identification and appraisal of students' knowledge, insight, understanding, skills, achievement, performance, and capability in Mathematics*". In typical US classrooms, however, assessment in mathematics is defined with content, skills and ability to reproduce these on demand (Royer, 2003).

The early 2000s was a time where mathematical proficiency was in the spotlight for both the National Council of Teachers of Mathematics (NCTM) and the National Research Council (NRC), as this time period followed the height of the Math Wars in the United States (Klein, 2003). In 2000, NCTM defined five process standards that describe mathematical practices students should engage in to reach proficiency in mathematics. These standards were problem solving, reasoning and proof, communication, connections, and representation. *Problem solving* was defined as applying and adapting a variety of strategies to solve problems as well as monitor and reflect on the process of problem solving. *Reasoning and proof* means that students can develop and evaluate mathematical arguments and use various types of reasoning and methods of proof. *Communication* refers to using mathematical language to express mathematical ideas precisely and the ability to analyze and evaluate the mathematical thinking of others. Making *Connections* is defined as understanding how mathematical ideas build upon each other to produce a coherent whole, as well as applying mathematics to other contexts. Lastly, *Representation* means that students can create, analyze, connect, and apply multiple representations model and reason about physical, social, and mathematical phenomena. These five standards were created to guide teachers in the type of mathematical processes students should have the opportunity to practice in order to learn mathematics content.

Around the same period of time, the NRC published one of the most widely cited definitions of "mathematical proficiency" today. *Adding it Up* (NRC, 2001), highlights that *mathematical proficiency* in problem solving requires a set of interwoven and interdependent skills that can be represented by 5 strands: conceptual understanding, adaptive reasoning, procedural fluency, strategic competence, and productive disposition. *Conceptual understanding* is the comprehension of mathematical concepts, while *adaptive reasoning* is the capacity for explanation and reflection on problem solving. *Procedural fluency* can be defined as the skill of carrying out procedures accurately, efficiently and flexibly whereas *strategic competence* is the ability to represent and solve math problems. The last strand, *productive disposition*, is not based on procedural or knowledge-based components, but rather is defined as the self-efficacy, motivation, and the ability to see the utility of mathematics. While these strands are defined separately, they are inherently connected and intertwined (Figure 2).

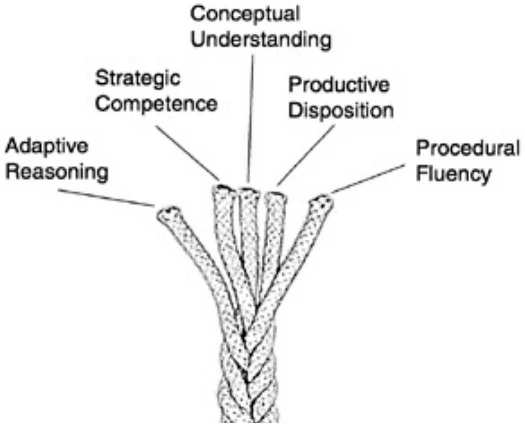
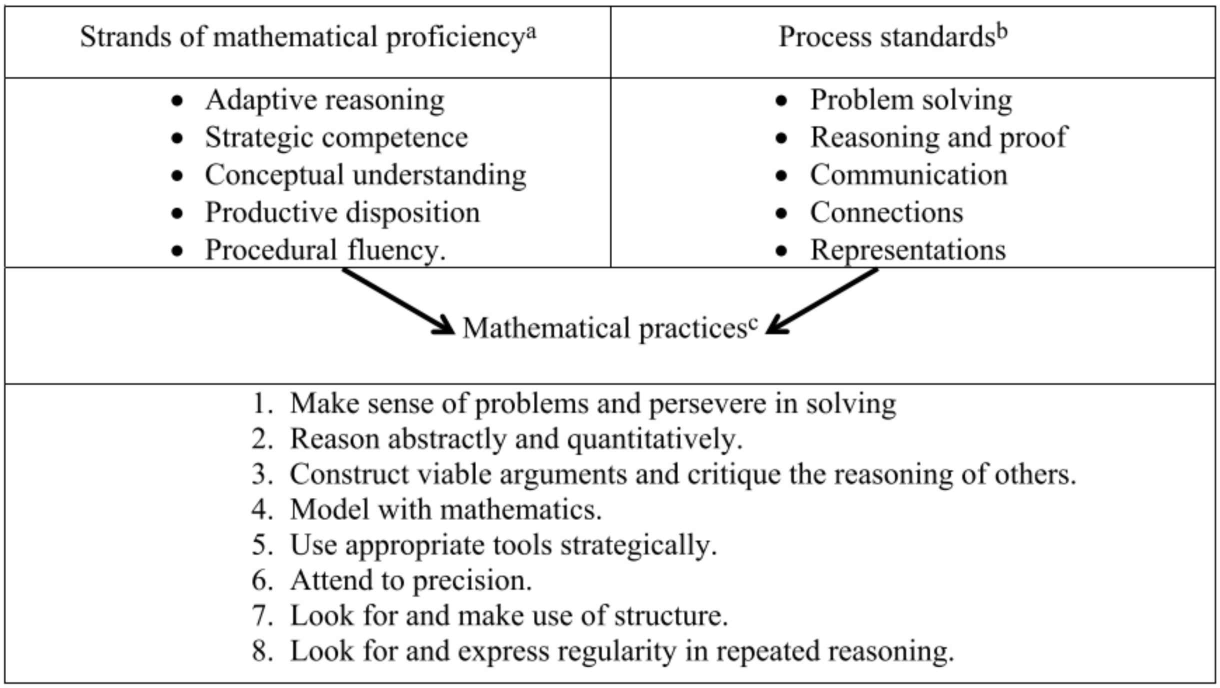


Figure 1-2. *Five strands of mathematical proficiency as described by the NRC publication Adding it Up (2001)*

Together, the 5 strands of Mathematics Proficiency and Process Standards became the foundation for the eight mathematical practices that are used in the Common Core State Standards for Mathematics (CCSSM) in current classrooms (National Governors Association Center & Council of Chief State School Officers, NGA & CSSO, 2010). Based on two pillars of process and proficiency, the CCSSM identify eight mathematical practices that reflect practices of students who are mathematically proficient (Table 1) *Making sense of problems and persevering through them* refers to students' ability to think mathematically and persist through challenges. 2) *Reason abstractly and quantitatively* means that students can both contextualize and de-contextualize the elements of the problem at hand to better understand quantities and their properties. 3) *Construct viable arguments and critique the reasoning of others* refers to the ability to understand, evaluate, and justify mathematical arguments. 4) *Model with mathematics* means that students can apply their mathematical knowledge to solve problems in the real world. This also involves reflecting on their solutions and asking if they make sense. 5) *Use appropriate tools strategically* describes students' ability to understand the benefits and limitations of different tools and choose tools appropriately for the situation. 6) *Attend to precision* means that mathematically proficient students clearly define their processes and communicate clearly about mathematics. 7) *Look for and make use of structure* refers to the practice of analyzing mathematical structures, decomposing problems into parts, and noticing mathematical patterns that are useful to sense making. 8) *Look for and express regularity in repeated reasoning* means that students can maintain the big picture of the process, while also discovering the details including repeated calculations that could serve as shortcuts for problem solving. These standards have moved far beyond the procedural vs. conceptual debate of the Math Wars to try and highlight the best mathematical processes, proficiencies, and practices. Today these standards have been adopted by 45 of the 50 United States and serve as a starting point for providing the right opportunities for students to become mathematically proficient.



Strands of mathematical proficiency[a]	Process standards[b]
• Adaptive reasoning • Strategic competence • Conceptual understanding • Productive disposition • Procedural fluency.	• Problem solving • Reasoning and proof • Communication • Connections • Representations

Mathematical practices[c]

1. Make sense of problems and persevere in solving
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.
8. Look for and express regularity in repeated reasoning.

Table 1-1. *NRC Strands of Mathematical Proficiency (2001)[a], NCTM Process Standards (2000)[b], and Common Core State Standards for Mathematics (2010)[c]. Original figure used in Allsopp, Lovin, & van Ingen (2017)*

In conclusion, this section has overviewed multiple perspectives on mathematical proficiency in the 20th and 21st century and how they have evolved into the current Common Core Standards for Mathematics today. The current work will utilize the modern definition of Mathematical Proficiency as the 5 Strands as defined by the NRC in 2001. While the Process Standards (NCTM, 2000) contain important practices and contexts for developing mathematical proficiency, this work argues that the 5 Strands are more descriptive of core skills that contribute to mathematical proficiency. While the research shows that these strands are important, there is less conclusive work on applying these measures in classrooms. While some summative assessments and qualitative analyses have been applied to evaluate students' mathematical proficiency (Samuelsson, 2008; Gotwals, Philhower, Cisterna, & Bennett, 2015), there is a lack of measuring all five strands during the process of learning. Learning technologies might be the right tool to provide formative assessment of mathematical proficiency in one context that could be used to inform further instruction and learning.

Mathematics Education and Assessment Facilitated by Learning Technologies

Learning technologies can provide students with immediate feedback, more individualized and self-paced learning, and more engagement through interactive content that rarely exists in more traditional forms of summative assessment (Cayton-Hodges, Feng, & Pan, 2015; Kiili, Devlin, Perttula, Tuomi, & Lindstedt, 2015). Compared to traditional pencil and paper tasks, these dynamic environments can provide students with scaffolding, adaptive content, and timely feedback such as hints or motivational messages (Romero & Ventura, 2010). Together, these features have been shown to enhance student learning compared to other learning technologies or traditional instruction (Kulik & Fletcher, 2016) and be nearly as effective as one-on-one in-person tutors (VanLehn, 2011). Learning technologies can also benefit students by giving them control over their own learning by allowing them to choose the pace of learning, the hints requested and whether to skip problems. Such choices produce higher learning gains and more positive affect (Ostrow & Heffernan, 2015; Aleven & Koedinger, 2000).These technologies turn learning into a self-paced and more individualized experience and can bring life to learning theories that can be experienced in ways not possible with traditional pencil-and-paper tasks.

Much work has been done in the field of educational data mining in terms of utilizing artificial intelligence and machine learning-based models to adapt content and features to best fit student needs (Conati, Gertner, & VanLehn, 2002). Substantial attention has been made to using student behaviors within the adaptive features of intelligent tutoring systems (ITS) to predict lower level student features such as student knowledge and next problem correctness (Botelho, Adjei, & Heffernan, 2016; Ferguson, Arroyo, Mahadevan, Woolf, & Barto, 2006; Pardos & Heffernan, 2011; Roll, Baker, Aleven, & Koedinger, 2014). ITS-based detectors can determine student knowledge level by recording their performance behaviors within the tutor, such as the number of attempts made and the time it takes to solve a problem. When it detects that a student is too challenged, the ITS can adapt feedback and scaffolding on an individual student level.

Not only have detectors been used to respond to student knowledge level, but also higher-level behaviors such as detecting student affect (Arroyo et al., 2014; Wixon et al., 2014) gaming the system (Baker, Corbett, Koedinger, & Wagner, 2004) and help seeking behavior (Roll, Baker, Aleven, & Koedinger, 2014). Student emotion, i.e., attitudes, interests, and values that students exhibit and acquire in school, can play a profoundly important role in students' post-school lives, possibly an even more significant role than that played by students' cognitive achievements (Popham, 2009). Research shows that student emotion (e.g., boredom, confusion, and frustration) while involved in online problem solving can impact learning and performance. Affect is recognized as a key indicator of student engagement and is crucial to learning (Pekrun, Vogl, Muis, & Sinatra, 2017; Bieg, Goetz, & Lipnevich, 2014; Goleman, 1996). Online feedback and support is crucial for student’s negative emotion, such as disengagement and boredom (DiMello et al., 2008; Pekrun, Goetz, Daniels, Stupnisky, & Perry, 2010). While traditional ITS feedback tends to be in the form of identifying errors and providing extra hints and examples to support students, other ITS have implemented feedback that targets supporting student affect. MathSpring (formally called Wayang Outpost), for example, adapts problems to individual students using a student model that assesses both cognition and effort (Arroyo et al., 2014) MathSpring also provides affective support in the form of an avatar that instills Growth Mindset, worked-out examples, multimedia hints tailored to each problem, and tutorial videos (Arroyo, Woolf, Cooper, Burleson, & Muldner, 2011). Embodying a Growth Mindset has been shown to be hugely effective in increasing student motivation to solve more problems, as well as improving student attitudes towards challenging content (Dweck, 2002). MathSpring invites students to grow a garden, where each problem set is represented by a growing plant; in doing so, it engages students in a game-like activity where the reward is to grow these plants that represent their sustained effort in a problem set, as well as mastery charts that represent a students’ achievement. MathSpring has been found to increase student learning (Arroyo, Beal, Murray, Walles, & Woolf, 2004), as well as increase motivation and engagement (Arroyo & Woolf, 2005; Arroyo et al., 2011).

Recently, a few projects have emerged that apply ITS-based detectors in learning platforms to inform instruction and learning. One project, Lumilo, has implemented multiple detectors into a single platform that serves both as an intelligent tutoring system for learners and a classroom management system for teachers (Holstein, Hong, Tegene, McLaren, B. & Aleven, et al., 2018).

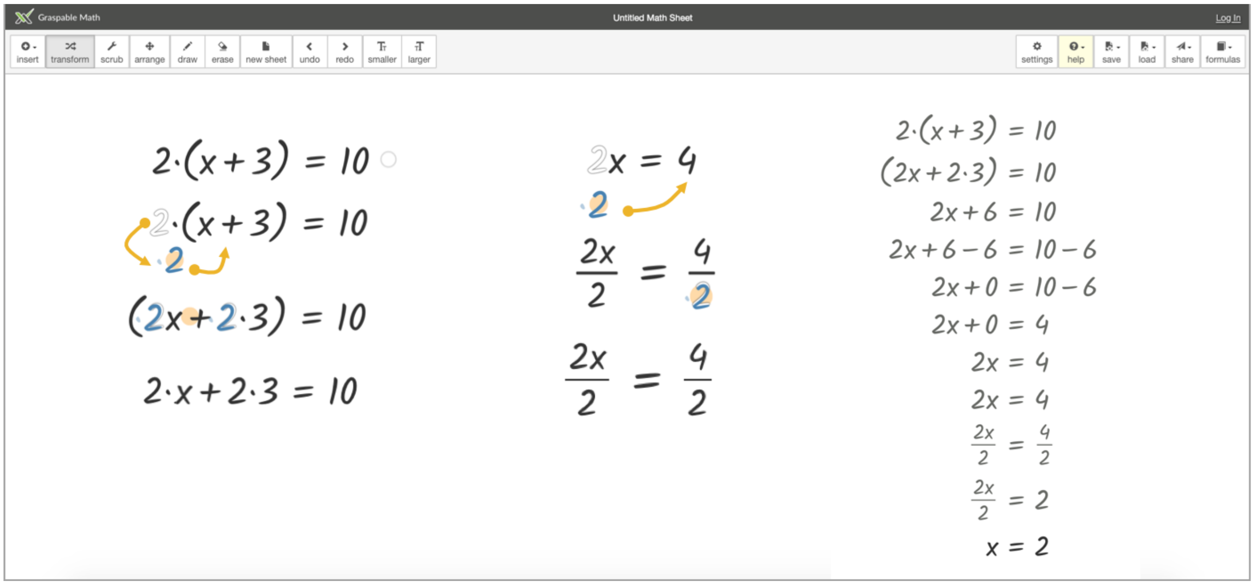
While ITS have been shown to be effective for student learning in many contexts (Kulik & Fletcher, 2016), often these systems focus on the student experience and leave out the role of teacher facilitation and orchestration of the classroom (Baker, 2016; Yacef, 2002). Lumilo, however, provides a platform to transform both the learning experience for students and the classroom management experience for teachers. Lumilo utilizes ITS-based detectors to provide teachers with real time analytics on their students through a mixed-reality smart glasses (Holstein et al., 2018). By incorporating multiple measures of learning through multiple detectors, this system is able to synthesize more about the learning process than traditional summative tests, which focus on the learning outcomes. Lumilo can measure how much effort a student puts into their work in the form of help-seeking behavior, wheel-spinning, and gaming. Not only does this technology give a better view of student knowledge and the learning process than traditional classroom assessments, but it also displays this information in a way that helps teachers transform the learning experience. Lumilo restructures the learning environment by allowing teachers to see live analytics in a mixed-reality format. In a randomized-controlled experiment, researchers found that adding Lumilo to the classroom can affect both student and teacher behaviors to improve learning gains (Holstein, McLaren, & Aleven, 2018). Giving teachers the ability to monitor students with the glasses changed student behavior as they became aware that they could be monitored and produced more desirable behaviors. Adding the real-time analytics, changed teacher behavior to spend more time with struggling students, which in turn, narrowed the achievement gap between low and high knowledge students from pre to post test. This research exemplifies how technology can be used to measure the learning process more deeply and can be applied to restructure the learning environment and teaching and learning practices to enhance the learning experience.

These are state-of-the-art examples of how user modeling of student interactions with learning technologies can be applied to student knowledge, emotion, and affect. While some detectors are used to prevent or respond to negative behaviors, such as gaming and boredom, most detectors are implemented with the ultimate goal of creating effective learning environments that can identify and encourage strong predictors of positive behaviors, specifically learning gains. Learning technologies, like ITS provide ample data to recreate student problem solving activities, as well as build models of optimal problem solving behavior, learning curves, moments of learning, and even student affect (Arroyo et al., 2016; Baker, Goldstein, & Heffernan, 2011; Koedinger & Mathan, 2004). One limitation in detecting knowledge, however, is that prediction models of learning often focus on correctness of solution to determine learning. For example, identifying learning curves relies on students mastering concepts, which often is defined as answering a certain number of problems correctly in row. However, log files from intelligent tutors can potentially predict so much more, such as more formative measures of learning (number of hints chosen, number of problem attempts, time between steps, and error types) or assessments during the process of learning. Whereas correctness is summative and can only be assessed once a problem is finished, formative measures could be better indicators of learning during the process of problem solving.

This work acknowledges the research that has been done within the mathematics education community in terms of breaking down components of mathematical learning, as well as the learning sciences community in terms of modeling that student knowledge and learning. The overarching goal is to integrate the theoretical framework of the five strands of mathematical proficiency into the assessment and measurement capabilities of learning technologies. May this work serve as one method to develop a theoretical and conceptual framework that grounds the formative assessment of learning through technology within the context of mathematics.

Chapter 2: Graspable Math as a Tool to Measure and Develop Mathematical Proficiency

The current work focuses on *Graspable Math* (*GM*), an innovative dynamic learning technology that utilizes motion to teach algebraic structure and problem solving (Ottmar, Landy, Goldstone, & Weitnauer, 2015). During problem solving, users transform and solve equations by physically moving the terms of an equation or expression on the screen. Figure 2-1 shows a few step-by-step images of how algebraic transformations work in *GM*. For example, if users want to distribute the *2* in *2*(x+3)=10*, they need to click and drag the two from the outside to the inside of the parentheses. Then the system automatically multiplies the *2* across to the *x* and the *3*. In another example, if users try to incorrectly add two unlike terms, such tapping the plus sign between *2* and *3x*, the system will shake the entire expression from left to right, like a shaking head, to indicate that this action is invalid.



$$2\cdot(x+3)=10$$
$$2\cdot(x+3)=10$$
$$2$$
$$(2x+2\cdot3)=10$$
$$2\cdot x+2\cdot3=10$$

$$2x=4$$
$$2$$
$$\frac{2x}{2}=\frac{4}{2}$$
$$\frac{2x}{2}=\frac{4}{2}$$

$$2\cdot(x+3)=10$$
$$(2x+2\cdot3)=10$$
$$2x+6=10$$
$$2x+6-6=10-6$$
$$2x+0=10-6$$
$$2x+0=4$$
$$2x=4$$
$$2x=4$$
$$\frac{2x}{2}=\frac{4}{2}$$
$$\frac{2x}{2}=2$$
$$x=2$$

Figure 2-1*. Step-by-step algebra transformations in Graspable Math*

In one sense, the system supports users in fully performing algebraic transformations, such as distributing out across all terms in the parentheses. In another sense, it allows users to attempt to make algebraic mistakes, like adding unlike terms, but provides immediate feedback that prevents users from committing to mathematical errors. Compared to traditional problem solving on pencil and paper, GM allows students to dynamically transform equations, making the problem solving process more fluid. *GM* also provides users with immediate feedback that reinforces mathematical rules. *GM* has been shown to increase student performance and engagement compared to non-motion based methods of instruction (Landy & Goldstone, 2007; 2010; Ottmar, Landy, & Weitnauer, 2015; Weitnauer et al., 2016; Ottmar, Landy, & Manzo, 2017; Ottmar, et al., 2017).

Theoretical Framework

Algebra pedagogy is progressing from memorizing rules and static line-by-line problem solving to using alternative modalities including learning technologies and physical movement. Algebra has been found to have a strong element of perceptual-motor learning (Kirshner, 1989; Kellman, Massey, & Son, 2010; Goldstone, Landy, & Son, 2010). Being able to manipulate the terms of an equation as objects allows students to better understand and manipulate algebraic notation. Experts in mathematics have shown to rely on perceptual motor cues in order to solve problems quickly and efficiently (Braithwaite et al, 2016; Rumelhart et al, 1986). Teachers can facilitate the learning and practice of these perceptual skills through the use of manipulatives, like tiles, that students can physically move. This taps into students' perceptual learning systems, allowing them to explore the innate structure of algebra physically and visually, and has shown to improve student engagement and algebraic reasoning (Ottmar, Landy, & Goldstone, 2012). They key to successful perceptual practice and manipulation of algebraic structures relies on systems that correctly embody mathematical rules.

One pitfall of using physical manipulatives to represent algebraic equations is that these manipulatives are used to replace the mathematical system rather than explain it (Uttal, Scudder, & DeLoache, 1997). Though many argue that giving students concrete objects that represent mathematical expressions can increase understanding, this only happens when students can also relate the manipulatives to their underlying mathematical concepts (Ball, 1992; Gentner & Ratterman, 1991). Physical manipulatives do not guarantee that students will understand the underlying concept. Physical manipulatives can also fall apart when students try to use them in unintended ways. Computer-based manipulatives, however, can be more specifically designed and can provide students with flexibility in thinking in ways that physical manipulatives cannot (Clements, 2000). In addition to more flexible design features, computer-based manipulatives can provide students with immediate and individual feedback. Rather than competing with other students for teacher feedback, designing a system with technology can provide students with scaffolding and feedback that informs them when they are on the right or wrong path.

GM is based on a perceptual-motor framework (Goldstone, Landy, & Son, 2010) that allows students to physically move algebraic terms around the screen as if they were literal objects. Prior research has shown the benefits of perceptual-motor learning in algebraic understanding by enabling students to "see" terms in an equation as objects to manipulate (Kellman, Masey, & Son, 2010; Kirshner, 1989). By embedding the rules of algebra within the structure and environment of the gesture-based system and providing instantaneous feedback so that students know exactly what moves are mathematically correct and incorrect, *GM* encourages students and users to explore and 'play' with the structure of algebra in every move. Using this approach, prior work has shown that use of *GM* can increase students' engagement and knowledge of early algebraic ideas, compared to non-motion based instructional activities (Landy & Goldstone, 2007; 2010; Ottmar, Landy, & Weitnauer, 2015; Weitnauer, Landy, & Ottmar, 2016; Ottmar, Landy, & Manzo, 2017; Ottmar, et al., 2017; Braith, Daigle, Manzo, & Ottmar, 2017).

GM-based Tools

One instantiation of *GM* is a game-based activity called *From Here to There! (FH2T)*. The primary focus of *FH2T* is to practice foundational algebraic concepts and create a solid basis of understanding surrounding the decomposed properties and flexibility of numbers. The intended audience for *FH2T* is middle school, where algebra is traditionally introduced, but has the potential to support students through early high school as algebra is the foundation of higher mathematics.

The game's universe-like module progression aligns with the Common Core standards and allows students to 'play' the game by slowly increasing in complexity through levels (i.e. subtraction, addition, order of operation).

Each module presents a series of puzzles. Rather than simply solving for "x", students are asked to make the given expression look like an equivalent expression that was specified in the goal (Figure 2-2. Both gamified and plain versions). To achieve this goal, students perform a series of dynamic interactions, including rearranging terms to apply the commutative and associative properties, decomposing numbers, combining terms (like 1+1=2 or 2*1=2), and performing the same operation to both sides of an equation (like multiplying by 3). This promotes the essential algebraic skills of number sense and decomposition, as students must understand how to break apart and recombine numbers in order to progress. The unique features of *FH2T* such as its goal-state 'solution', provide a suitable environment for students to engage in trial-and-error decomposition while remaining within the confines of natural math law.

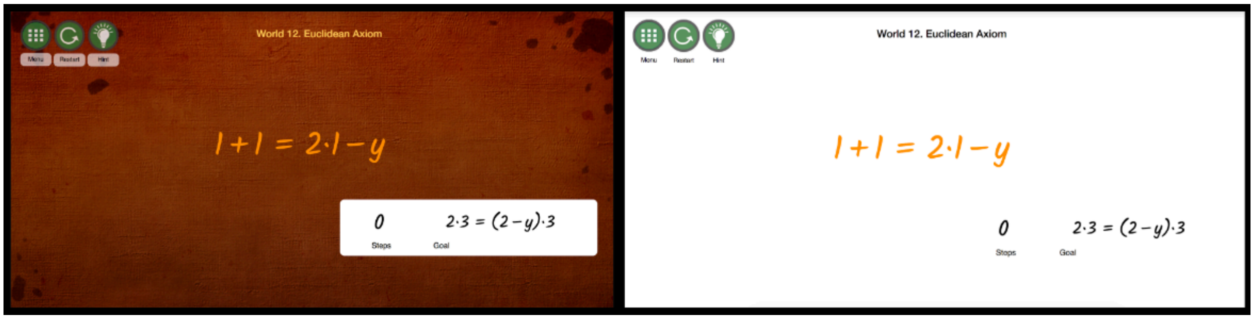


***Figure 2-2.** Gamified and plain versions of FH2T.*

Current Work: Measuring Mathematical Proficiency with Graspable Math

The idea of better measurement of mathematical proficiency is not new. NCTM's publication, *An Agenda for Action (1980)*, called for a new direction for mathematics and wider range of measures than conventional testing, however, even today, most assessments used in schools are summative, test procedural knowledge, and emphasize the correctness of the answers (Darling-Hammond et al., 2013; Pellegrino et al., 2016; Pellegrino, 2012; Chudowsky & Pellegrino, 2003; Schoenfeld, 2007; Rittle-Johnson, 2017). Not only do correctness-based assessments exclude any measures of process, but they also ignore measures of student disposition, motivation, and persistence, which have been shown to be key elements in reforming understanding, preparing for future learning, and developing expertise (Hiebert & Grouws, 2007; Kapur, 2010; Skemp, 1971). Though each of these elements of mathematical proficiency are comprised of separate skills, understanding the relationships between them will provide insight into what types of students are successful and how mathematical proficiency is developed. However, without appropriate measures of these skills in a single problem solving context, it is difficult to examine these relationships (Schoenfeld, 2007). As a field, we need to break away from correctness-based standardized testing and design more innovative assessment features that help students learn and succeed by measuring the entire learning process in real time (Chudowsky, & Pellegrino, 2003; NRC, 2001). Utilizing the data collection capabilities of dynamic learning technologies could be the solution, providing a single platform to analyze many aspects of mathematical problem solving.

Many learning technologies have the ability to collect process data, but much research and classroom practices still focus on correctness and other external factors such as prior knowledge, gender, and SES. Though many of these external factors are strong predictors of learning, this work also proposes that in-app behaviors can add a lot in terms of explaining students' misconceptions and their learning process. *GM* technology shows great potential as a tool to research mathematical proficiency compared to other learning technologies due to its unique gesture-based system. These gestures allow the system to record all student actions, including intentional steps and errors, during algebraic problem solving. *GM* is the perfect platform to explore student problem solving behavior and to tease apart the strands of mathematical proficiency. By adding in-app behaviors into our models, we may be able to see trends that explain more about the learning process. This thesis is a first step to understanding how to leverage the data collected within the app to identify different aspects of mathematics proficiency.

In order to do this, the *GM* team spent a considerable amount of time and effort on feature design using *GM* measures. This included deciding what variables should be recorded for every *GM*-based study and defining exactly what each variable measures in relation to user interactions with the system. Based on these definitions, the conceptual model below (Figure 2-3) situates the formative measures of *GM* within the theoretical framework of mathematical proficiency developed in Chapter 1. In the innermost layer, the model depicts the five strands of mathematical proficiency (NRC, 2001). The middle layer then maps those strands onto constructs of learning found in the literature on educational research. Finally, the outermost layer maps the specific measures collected in *GM* onto both the learning constructs and strands of mathematical proficiency. This conceptual model will serve as a guide on how to ground the measurement of mathematical proficiency within the theoretical framework and in the context of *GM*. This work proposes that building a conceptual model on theory is a key component of designing effective learning technologies both in learning theory and assessment. This conceptual framework will continue to be used throughout this work as a guide to test the hypothesized mappings of *GM*-based measures onto the five strands in Section 2. This framework will also guide work in Section 3, which describes the iterative development of *GM*-based activities designed for supporting and potentially measuring mathematical proficiency.

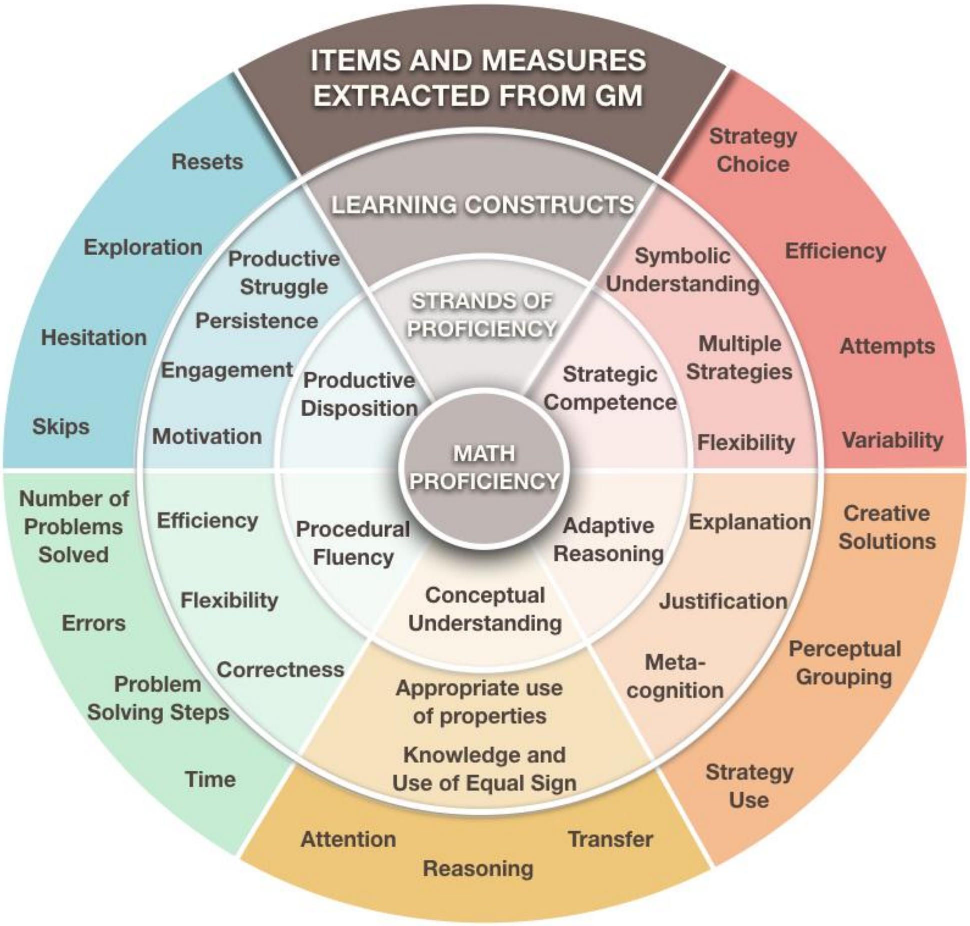


Figure 2-3. A conceptual model that maps GM measures of problem solving process (indicators) onto learning constructs and the 5 strands of mathematical proficiency.

This work utilizes *Graspable Math* to tease apart components of mathematical proficiency more efficiently and at a deeper level than possible with traditional summative assessment. These first two chapters contextualized the current work within the fields of mathematics education, learning sciences, and educational data mining. The major contribution of this work is developing a theoretical and conceptual framework to assess mathematical proficiency within *GM* specifically, but could be a method applied to any learning technology to assess components of learning. The chapters in Part 2 will compare constructs of mathematical proficiency across elementary and high school populations and attempt to map those constructs onto the conceptual framework. These studies lead into Part 3 and inform the development process of new *GM*-based activities and tools which aim to support the learning of all strands of mathematical proficiency, validate measures of mathematical proficiency, and design tools to support the implementation of these *GM*-based resources into K-12 classrooms.

Part 2: *Graspable Math* in K-12 Classrooms for Learning and Assessment

Part 2 addresses the second objective of this work, to ***explore GM-based measures of mathematical proficiency across K-12 populations***. Two studies evaluate the feasibility of *GM* in elementary and high school populations, as well as the benefits of using *GM* as a formative measurement tool to analyze the learning process. Previous research has shown the potential of the middle school version to increase student performance and engagement compared to non-motion based methods of instruction (Landy & Goldstone, 2007; 2010; Ottmar, Landy, & Weitnauer, 2015; Weitnauer et al., 2016; Ottmar, Landy, & Manzo, 2017; Ottmar, et al., 2017). Though the GM approach had been tested extensively at the middle school level, it had not yet been tested extensively at the elementary or high school levels. The studies in this section aim to fill that gap. While the first study implements *GM* into elementary school classrooms, the second study brings *GM* to high school classrooms where many students are considerably below grade level. Together, these two studies analyze 1) the added benefit of utilizing formative in-app measures in predicting learning, and 2) how these formative measures map onto the theoretical strands of mathematical proficiency in both populations.

Chapter 3: Feasibility and Formative Assessment: *Graspable Math* in Elementary School

This study has been published in the Journal of Educational Technology Research and Development (2019). This chapter extends the work previously established by Lindsay Braith for her Master's Qualifying Project (2017), as well as her, Dan Manzo, Maria Daigle, Erin Ottmar, and Jeanine Skorinko's poster presented at the American Psychological Society Conference (2017).

Citations:

Hulse, T., Daigle, M., Manzo, D., Braith, L., Harrison, A., & Ottmar, E. (2019). From here to there! Elementary: a game-based approach to developing number sense and early algebraic understanding. *Educational Technology Research and Development*, *67*(2), 423-441.

Braith, L., Ottmar, E., & Skorinko, J. (2017). *Even Elementary Students Can Explore Algebra!* (Undergraduate Major Qualifying Project No. E-project-042717-103336). Retrieved from Worcester Polytechnic Institute Electronic Projects Collection: https://web.wpi.edu/Pubs/E-project/Available/E-project-042717-103336/

Braith, L, Daigle, M, Manzo, D, & Ottmar, E. (2017). *Even Elementary Students Can Explore Algebra!: Testing the Feasibility of from Here to There!, a Game-Based Perceptual Learning Intervention.* Poster Presented at the American Psychological Society Conference, Boston, MA.

This first study was inspired by work that started as a Major Qualifying Project (MQP) at Worcester Polytechnic Institute by Lindsay Braith (2017). It is important to note that this was an incredible team effort by many of our lab members that spanned several years. Lindsay, along with our team sought out to 1) test the feasibility of scaling *GM* to the elementary level and 2) determine if there were any learning differences based on three conditions: traditional instruction, a plain version of the *FH2T:E* intervention, and a gamified version of *FH2T:E* (Braith, Daigle, Manzo, & Ottmar, 2017). First, changes were made to the middle school version of *FH2T!* so that the content was developmentally appropriate. This included focusing on the decomposition of numbers, scaling down the *GM*-based gesture tutorial to only introduce the four operations and decomposition, and limiting the vocabulary in written instructions and hints. In that study, there were nine second-grade classrooms (106 female, 113 male). Preliminary results showed that *FH2T:E* is feasible for elementary-aged students, as there was significant learning improvement from pre to post tests in relation to a traditional teaching control. However, there were no differences between the plain and gamified conditions of *FH2T:E*, suggesting that gamified elements in the intervention did not contribute to explaining learning gains.

That project used an established method of comparing experimental conditions to traditional classroom instruction. This was the first step in making *GM* accessible to students in elementary school, where the first ideas of pre algebra are introduced. Though the study found that students learned more if they engaged with the *FH2T:E* intervention, it is unclear what components of the program relate to those learning gains. This inspired the method for the study in Chapter 3, which uses the in-app data to examine the learning process at a more fine-grained level. Perhaps there is more to be explained when introducing in-app problem solving behavior into predictive models. There is natural variation in how students engage with the intervention, so it is logical to include this variance when explaining differences in learning. This study was designed to serve as an example method for learning technology-based interventions that includes both formative (in-app data) and summative (pre and post) assessments of learning.

With feasibility already assessed in that project, the main objectives of this chapter are to dig deeper into these findings to explore possible predictors and moderators. Using the student log data created during mathematical problem solving, we reveal latent constructs of mathematical proficiency within the context of FH2T. This study addresses three research questions:

1. Are there differences in learning between the gamified and non-gamified versions of FH2T:E?
2. Do in-app measures of student problem solving process predict learning gains?
3. Do certain student behaviors within *FH2T*:E differentially predict learning for high or low-knowledge students?

Participants, Experimental Conditions, and Procedures

The study included 185 second grade students from ten classrooms in three different elementary schools in Massachusetts (116 female, 78 male) participated in this study. The study spanned 3 weeks and followed a pretest, intervention, posttest structure (Figure 3-1). During week 1, students were given a 15-item pretest that assessed Common Core Standards for second grade mathematics. Students were then randomized into one of two experimental conditions: gamified versus non-gamified (see figure 2-2). During week 2, all students interacted with the app in their mathematics classrooms across 4 days in 20-minute sessions, for a combined total of 80 minutes of play. As part of the gamified condition, students played through the version of the game that possessed game-like features. Gamification elements included the presence of levels, color, prizes, bonuses, stars, etc. The non-gamified version of *FH2T:E* was stripped down to display only the 18 math problems within each level. As students played through this plain version, there was no recognition of level completion or rewarded points for accuracy and efficiency. This lack of reward-based prizes was intended to assess the degree to which the learning gains stemmed from the gamification features or the goal-state dynamic approach that the *FH2T:E* game provided. The math content in each version was exactly the same. The only differences between conditions were the presence or absence of gamified visual material. Therefore, if differences in learning between conditions are statistically significant, results may highlight possible mechanisms by which *FH2T:E* leads to gains. Finally, in week 3, all students were given the post-test assessment that matched the pre-test assessment with slight modifications.

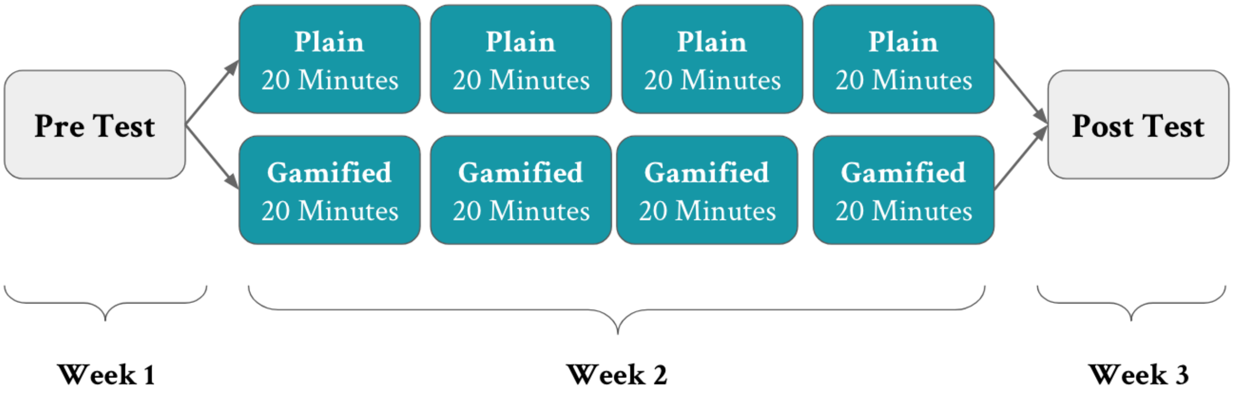


Figure 3-1. *Pretest, Intervention, Posttest Study Design*

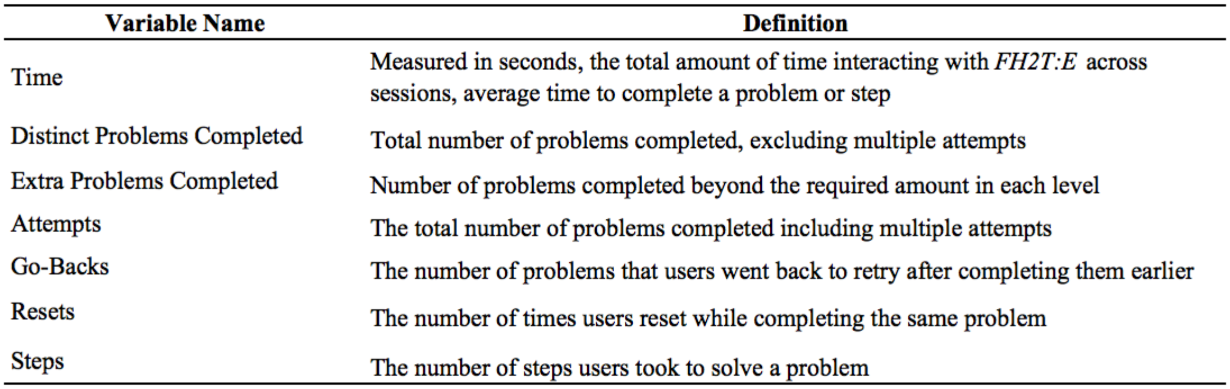
Measures

Data collected for this project included a combination of student scores on pre- and post- study worksheets and in-app data logs of the students interacting with the game.

Pre and post assessments: Prior to the introduction to the game, students completed a 15-item pre-study worksheet to assess prior math knowledge. These questions mirrored second grade math standards set forth by the Common Core (Common Core State Standards of Mathematics (CCSSM), 2010) and tested baseline understanding of decomposition, operational strategies, and basic notation. Completion of the pre-assessment was done one week before interaction with the game. A week after the four sessions were completed, students completed the post-study worksheet. The problems and expressions on the posttest were similar to those found on the pretest. To ensure baseline equivalence, an independent-samples t-test was conducted to compare pretest scores for gamified and plain conditions. There was not a significant difference in pretest scores for the gamified (M=9.85, SD=3.89) and plain (M=9.95, SD=3.60) conditions; t(183)=0.17, p = .865.

In app process data: As mentioned, *FH2T:E* has a data logging system that records all student actions, mouse clicks and trajectories, errors, and moment-by-moment problem solving steps while interacting with the system. The recorded data for this study were compiled and aggregated across problems, levels, sessions and overall to create a series of variables that described problem solving processes. This paper uses the 19 overall variables to represent composite measures of student action and problem solving process in *FH2T:E* over the duration of the study. A summary of the main measures are described in Table 3-1.

Table 3-1. *Labels and definitions of primary variables measured in FH2T:E.*



Variable Name	Definition
Time	Measured in seconds, the total amount of time interacting with *FH2T:E* across sessions, average time to complete a problem or step
Distinct Problems Completed	Total number of problems completed, excluding multiple attempts
Extra Problems Completed	Number of problems completed beyond the required amount in each level
Attempts	The total number of problems completed including multiple attempts
Go-Backs	The number of problems that users went back to retry after completing them earlier
Resets	The number of times users reset while completing the same problem
Steps	The number of steps users took to solve a problem

An exploratory factor analysis, using Principal Axis Factoring, was then conducted to identify the number and structure of the factors underlying the overall data variables that were recorded within *FH2T:E* as students solved problems. Before conducting the factor analysis, all 34 of the initial variables were examined in a correlation matrix to test a few assumptions. It is recommended that all variables should be significantly correlated with at least one other variable (Tabachnick & Fiddell, 2007). It is also recommended that factors should not be correlated above .9, as that would violate assumptions of multicollinearity (Field, 2009). There was only one factor that did not correlate with any others, *Star Score.* However, three groups of variables with correlations above .9: 1) *Extra Problems Completed* and 2) *Percentage Extra Completed, Average Time Per Problem, Problems Per Minute,* and *Average of Best Time*, as well as 3) *Distinct Problems Completed, Distinct Problems Unlocked, Percentage Problems Completed, Extra Problems Completed, Completed Stars* and *User Stars*. We chose to remove *Star Score* from analyses as it was an engineered measure from multiple other measures and did not correlate with any others. As for the groups of multicollinear variables, we decided to choose one variable from each group to represent the rest. We chose *Extra Problems Completed* to represent group 1 (about extra completed problems), *Average Time Per Problem* to represent group 2 (about time), and *Distinct Problems Completed* to represent group 3 (about overall completed problems). This left us with a total of 19 variables in the final analyses.

The KMO test values above .5 can be considered for EFA, with values above .9 considered as excellent (Hutcheson & Sofroniou, 1999). Our KMO resulted in .751, which means our sample is adequate for producing reliable factors. The Bartlett test was significant ^{2}(171) = 5877.65, p < .001, which means our correlations are significantly different from zero. With these considerations met, our sample was determined suitable for EFA.

Using SPSS 22, Principal Axis Factoring was conducted using a Promax rotation. Promax was chosen as it is an oblique rotation that assumes the factors are correlated. Communalities describe the proportion of variance explained by the underlying factors and values above .5 are considered adequate for factor analysis (MacCallum, Widaman, Zhang, & Hong, 1999) Communalities in our sample all resulted in values above .5. In fact, all variables except *Extra Problems* resulted in values above .9. Next, 5 factors were extracted using Kaiser's criterion (1958) criterion: that eigenvalues are greater than 1.00, that each factor be comprised of at least two factor loadings of > 0.40, and that the resulting components demonstrate good internal consistency.

Five factors with eigenvalues greater than 1.00 and sufficiently large loadings were extracted and they explained 29.74%, 24.70%, 21.29%, 8.82%, and 6.22% respectively, explaining a total of 90.78% of the variance (Table 3-2). The five factors are described in Table 3-2 and have been classified based on how variables loaded onto each factor. Factor 1, which included *Total Go-Backs, Percentage of Attempts, Percent of Go-Backs, Number of attempts,* and *Overall Time Interaction* has been classified as ***Engagement in Problem Solving***. This factor represents a measure of the number of problems solve: however, this measure does not represent greater progression through the app. students with higher scores on the go-backs factor were more likely to attempt and complete the same problems multiple times. Factor 2, which included *Distinct Problems Completed, Completed Best Step*, and *Extra Problems Completed* has been classified as ***Progression***. For example, students with higher scores on this progression factor solved more distinct problems and progressed through the app more quickly. The distinction between factor 1 and factor 2 is important as it allows us to test whether it is simply practicing problems (attempting and completing the same problem more than once) or progression through the app (moving through the app and completing more unique problems) that is more beneficial for students. Factor 3, which included *Percentage of Resets, Average Attempts Completed, Total Resets,* and *Average Resets* has been classified as ***Strategic Flexibility***. This represents a measure of how often students reset problems to try different approaches before successfully completing the puzzles. Factor 4, which included *Average Time Per Step, User First Step, Percentage Stars, User Total Step,* and *First Efficiency* has been classified as ***Strategic Efficiency***. Higher scores for Factor 4 (strategic efficiency) represents using a minimal number of steps while solving problems. Finally, Factor 5, which included *Average Time Per Problem* and *Best Time* has been classified as ***Speed***, a measure of student rate of solving problems. Correlations indicated that the 5 factors were also sufficiently independent of one another, indicating that they measure separate latent constructs.

Approach to Analysis

First, descriptive statistics and correlations were calculated for each factor and variables. Next, four multiple regressions were conducted to examine relations between predictors and outcomes. The first model examined whether there were differences in performance between students in the gamified and non-gamified condition. Next, in model 2, the 5 latent in-app process measures were added into the analysis to explore which game behaviors contributed to learning. Our next step was to examine whether certain behaviors within *FH2T:E* mattered more for high or low performing students. In this study, we hypothesized that the two indicators of problem solving practice within the app (progression and engagement with problem solving) may vary depending on students prior knowledge levels. In model 3, we examined the interaction between progression and prior knowledge, while in model 4, we examined the interaction between engagement with problem solving and prior knowledge.

Table 3-2. *Structure coefficients from principal axis factor*

Item	Engagement	Progress	Strategic Flexibility	Strategic Efficiency	Speed	Mean	SD
Total Go-Backs	0.935					25.41	49.42
Percentage of Attempts	0.908					2.00	0.91
Percent of Go-Backs	0.873					0.19	0.29
Number of Attempts	0.778					151.61	66.03
Overall Time Interaction	0.677					2601.70	868.83
Problems Completed		1.049				78.53	21.53
Completed Best Step		1.038				122.43	44.34
Extras Completed		0.774				18.06	7.24
Percentage of Resets			0.880			0.20	0.11
Ave Attempts Completed			-0.864			0.70	0.15
Total Resets			0.821			15.50	9.58
Average Resets			0.784			0.10	0.05
Ave Time-Step				0.834		7.84	2.65
User First Step				-0.687		190.56	80.29
Percentage Stars				0.666		0.88	0.08
User Total Step				-0.635		366.22	189.77
First Efficiency				0.554		2.67	1.81
Ave Time-Problem					0.999	27.48	9.96
Best Time					0.585	2077.05	1002.77
Eigenvalues	5.65	4.69	4.05	1.68	1.18		
Percent of Variance (%)	29.74	24.70	21.29	8.82	6.22	Total:	90.78

Results

Means, standard deviations, and correlations among the pretest, posttest, and latent factors are presented in Table 3-4. Pretest scores were correlated with higher completion (r=0.27), higher go-backs (r=0.24), and higher post-test scores (r=0.70). Solving problems more quickly (time) was related to greater completion (r=0.37) and fewer go-backs (r=-0.25). Results from all models are presented in Table 3-5.

Research Question 1: Our first aim was to determine whether there were differences in math posttest performance between students who received the gamified and non-gamified conditions. Results suggest that there were no differences in post test performance between the gamified and non-gamified conditions (p>0.05), when only condition, gender and pretest performance were used to predict posttest performance.

Table 3-4. *Descriptive Statistics and Correlations*

Factor Correlations	1	2	3	4	5	6	7	8	9
1. Posttest Score	--								
2. Pretest Score	.70**	--							
3. Gender	.09	.14	--						
4. Condition	-.05	.01	.01	--					
5. Factor 1- Engagement	.21*	.23	.05	.33**	--				
6. Factor 2- Progress	.27**	.32**	.05	.26**	.04	--			
7. Factor 3- Strategic Flexibility	.12	.06	.07	-.22**	.16**	.21**	--		
8. Factor 4- Strategic Efficiency	.12	.11	-.07	-.09	.01	-.23**	.23**	--	
9. Factor 5 Speed	-.07	-.08	-.18*	.04	-.22**	.51**	.25**	-.18*	--
Mean	74.23	65.91	.53	0.61	0	0	0	0	0
Standard Deviation	23.96	25.14	0.50	0.49	1	1	1	1	1

*p < .05; ** p < .01

Research Question 2: After including in-app student interaction components, a significant effect of condition emerged (p<0.05). Students in the gamified condition performed, on average, 6.58 points higher on the posttest than students in the non-gamified condition. Further, progress (factor 2) was approaching significance (*p*=.056), suggesting that students who progressed faster and completed more unique problems in the app may demonstrate higher posttest scores. More specifically, for every one standard deviation increase in completion, students performed approximately 3.07 points higher on the posttest. No other in-app measures predicted learning.

Research Questions 3 and 4: As displayed in figure 3, a significant interaction was present for Progress (factor 2) and prior knowledge. Students with lower initial pretest scores who completed more problems in the *FH2T:E* game demonstrated increased learning gains compared to students who completed less problems. However, posttest achievement for initially high knowledge students was similar, regardless of the amount of problems students completed (Figure 3-2). A similar interaction and pattern emerged for Engagement with Problem Solving (Figure 3-3), with low knowledge students who engaged more with problems gained more than students who did not go-back and solve problems more than once. Engagement with Problem Solving did not seem to relate to achievement for high knowledge students.

***Table 3-5**. Model results predicting post-test achievement*

Parameter	**Model 1**	**Model 2**	**Model 3**	**Model 4**
(Constant)	27.95 (4.04)**	31.24 (4.34)**	33.91 (4.41)**	34.25 (4.36)**
Gender	-0.03(2.61)	-0.37 (2.72)	0.52 (2.70)	0.54 (2.68)
PreTest % Correct	0.68 (0.05)**	0.60 (0.06)**	0.58 (0.06)**	0.55 (0.06)**
Gamified	3.86 (2.61)	6.58 (3.16)*	5.51 (3.14)$^{+}$	8.39 (3.14)**
Engagement		1.90 (1.56)	15.17 (5.60)**	2.30 (1.53)
Progress		3.49 (1.81)$^{+}$	3.07 (1.80)$^{+}$	14.55 (4.09)**
Strategic Flexibility		0.63 (1.58)	-0.74 (1.59)	0.22 (1.55)
Strategic Efficiency		1.38 (1.48)	1.39 (1.46)	1.60 (1.45)
Speed		-1.48 (1.73)	-0.80 (1.72)	-1.28 (1.69)
Engagement x Pretest			-0.17 (0.07)**	
Progress x Pretest				-0.16 (0.05)**
F	57.54	22.45	21.25	21.93
R^2	0.50	0.52	0.54	0.55

Standard errors in parentheses. $^{+}$p <. 10; *p < 0.05; **p < 0.01

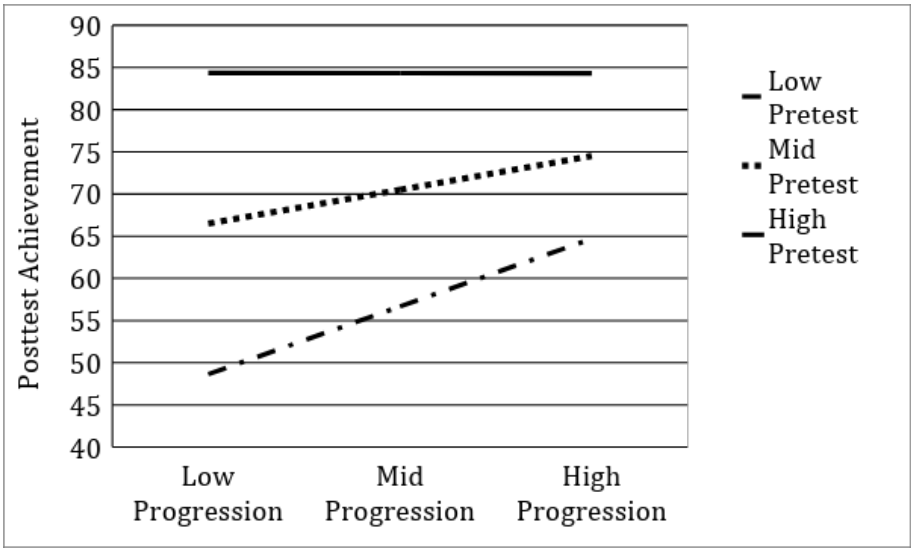


***Figure 3-2**. Interaction of Progression and Prior Knowledge on Posttest Achievement*

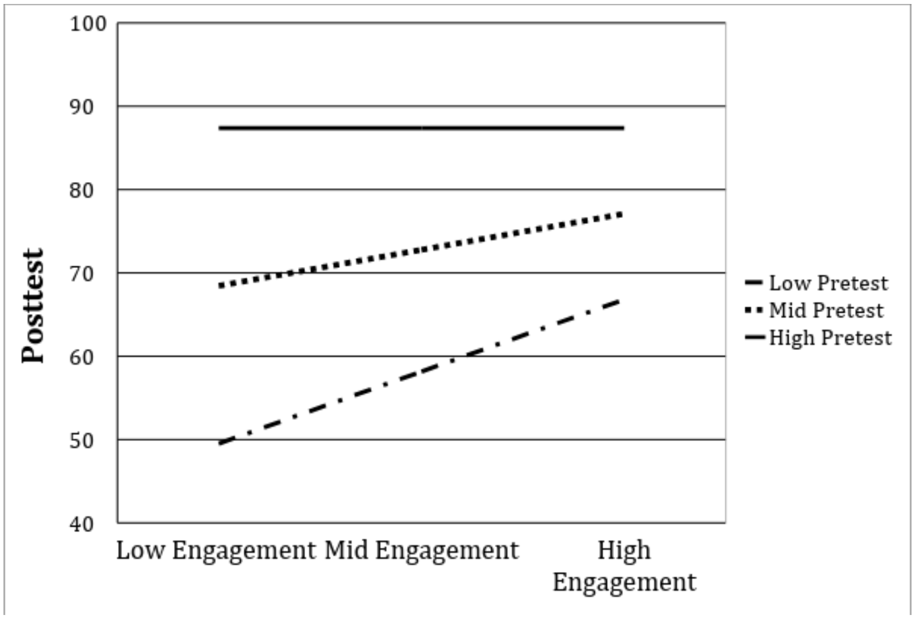


Figure 3-3. *Interaction of Engagement and Prior Knowledge on Posttest Achievement*

Discussion

This study examined several factors related to student behavior and math learning within *From Here to There!:Elementary*. Several main findings emerged from this study of second grade students. First, upon first examination, there did not appear to be significant differences in learning between gamified and non-gamified conditions. However, after accounting for in-app problem solving interactions, significant differences emerged, with students in the gamified condition being more likely to have larger gains on the posttest than students in the non-gamified condition. Next, solving more problems within the app could be related to higher achievement. Third, two significant interactions emerged, suggesting that solving problems within *FH2T:E* may be especially beneficial for low performing students: low performing students who solved more problems in the app and engaged in more behaviors in problem solving, including more attempts and going back to retry problems, were more likely to have larger learning gains than students who had initial higher levels of achievement.

After accounting for in-app behaviors, there is an advantage for gamification over non-gamification. Adding the support of gamified features may motivate students to engage with more difficult content that they have never learned before in a non-threatening environment. Furthermore, gamification may motivate these children to improve their problem solving strategies in order to receive rewards for the most efficient solution. However, it is important to note that efficiency and time were not significant predictors of mathematics learning. This is consistent with other work in mathematics education that values flexible problem solving process and thinking over speed and efficiency, even from the early years of mathematics instruction (Baroody, 2003). Completing more unique problems and progressing further through the app was related to improved learning, providing additional evidence of learning benefits by engaging with and using the app. It may be that completing more problems provided more opportunities for learning by increasing exposure to different types of content and problems that young children may have never seen before, such as more complex opportunities for decomposition with multiple operations.

The significant interaction effects identify differences in the more subtle aspects of interaction with the program and addresses the question, *Who does FH2T:E help most?* Results suggest that playing with and completing more problems in *FH2T:E* appears to be more beneficial for low performing students compared to high performing students whose learning did not significantly change. This may be due to the fact that low performing students have more to gain in terms of learning and *FH2T:E* can give low performing students a valuable learning opportunity. One benefit of online math games is that students can progress through the app at their own pace, allowing lower performing students to continue to practice mathematical concepts and problems within a safe environment. Interestingly, it does not seem to matter if low performing students complete more unique problems that continue to progress them through the app or if students practice the same easier problems multiple times (repeated practice). Similar patterns of gains in achievement are observed for both types of problem solving practice for low and average performing students. These findings are consistent with other work examining the benefits of math apps that allow for differentiation of learners with varying achievement levels (Moyer-Packenhaum & Suh, 2012), pointing to the importance of allowing students to re-do problems (attempts, go-backs) and solve math problems at their own pace. These findings are promising for using perceptually-guided puzzle-based problem solving as a means of decreasing the achievement gap between high knowledge and struggling students. Future studies should also address whether *FH2T:E* will benefit students with different demographic characteristics (racial, ethnic, linguistic, cultural, etc.) than those in the study population.

While we cannot definitively say why *FH2T:E* especially helped struggling students, it may be that the perceptual feedback, hints, and ability to reset and retry problems created new affordances for students that typically paper and pencil assessments does not provide. One plausible explanation may be that the puzzle-based design of the game was more motivating and engaging and less threatening for struggling students that the emphasis on correctness. Although we did not specifically measure math anxiety in this sample, these patterns are consistent with prior work in *FH2T* which suggested that students with higher levels of math anxiety and lower prior knowledge who engaged with *FH2T* solved more problems and did not experience detrimental effects of math anxiety on achievement compared to students who received more traditional instruction (Ottmar et al., 2015). Future studies should more closely examine the in-app data to compare the behaviors and relations between low and high performing students.

Implications for Math Teaching, Research, and Practice

These results suggest that it is feasible and productive to use games to support young students algebraic thinking through practicing early algebraic content, such as decomposition and order of operations. All students, regardless of their prior knowledge, were able to easily progress through the game. The flexible and accessible nature of the *FH2T:E* program supports the creation of new games in the future that can introduce physical interaction with content via a technological interface. From a preparation for future learning perspective, games might be especially effective because they can provide both motivation and learning gains while gradually exposing students to more difficult content and feedback within a supportive learning environment. The accessibility that web-based games provide may not only provide affordable opportunities for students to continue their math practice during the school year, but it may also serve as a promising intervention to bridge the gap over summer break when students often lose ground in content understanding.

Game-based learning technologies also have the potential to measure and assess student learning *during* the problem solving process. Though many instructional technologies have the ability to record all student interactions, there is little research on how these data can be used and mapped onto learning constructs of mathematical practice during instruction. This study is the first time that we have explored the predictability of new measures of in app interactions to assess mathematical learning within the *FH2T:E* game context. The additional information provided by this data revealed previously hidden effects of game-based components on learning Following these findings, future research directions should include studies to expand and generalize the *FH2T-E* approach within this age range and to develop additional versions of gamed-based perceptual learning algebra interventions designed for even younger students (Clements & Sarama, 2007; Lins & Kaput, 2004).

Now that we have identified five factors that seem to reflect student interaction with the game, the next step is to validate these factors within a different data set. Once validated, we can more generally use these composite scores to predict learning, as well as create profiles of student behavior to better understand which students succeed and fail. This could begin to tease apart differences in age, prior knowledge, and engagement with the app and shed light on how students, despite differing starting points, could utilize *FH2T:E* to increase mathematical performance. Future studies could include outcome measures reflecting differences in student engagement, motivation or strategy obtained from the in-app data logged for each individual student's "game-session." Finally, within this in-app data, the *FH2T:E* program has the capability to analyze errant attempts made by students as they approach solving various items. Thus, it enables researchers to visualize both the effective strategies used by students and the errors and maladaptive approaches. This sort of data could be used to examine questions of mathematical flexibility and intervene earlier by providing immediate feedback and additional practice more effectively.

Conclusion

Overall, this study provides further evidence of efficacy for *From Here to There!: Elementary* on improving student mathematical understanding. By providing activities that embed developmentally appropriate content and activities may make the introduction of early algebraic concepts into school classrooms more feasible and impactful. This study also shows the power of formative assessment in addition to summative assessment in understanding student problem solving behavior as they engage with the technology. It was only when in-app measures were added to the regression models that interaction effects were revealed between student prior knowledge and problem solving constructs (progression and engagement in problem solving).

Chapter 4: GM Problem Solving Behavior in Elementary and High School

This study has been presented as a poster at the 2019 Worcester Polytechnic Institute Graduate Research Innovation Exchange. This poster was awarded 1st place in the Social Sciences & Business category.

Citation:

Hulse, T., Manzo, D., Harrison, A., & Ottmar, E. (2019). Measuring Mathematical Proficiency Within a Dynamic Learning Technology. Poster presented as a finalist at the *Graduate Research Innovation Exchange*. Worcester Polytechnic Institute.

Chapter 3 explored the latent constructs of mathematical problem solving behavior in an Elementary population. That study showed the benefit of adding the formative measures within GM to the predictive models compared to only including summative measures alone. However, little is known whether the constructs measured by GM are consistent across grade levels. This chapter explores the latent constructs of mathematical problem solving behavior in a High school population using the same method as Chapter 3. Then the results from the Elementary and High school populations are compared in terms of the latent constructs that emerged from the interaction data.

The current study takes the first step in comparing problem solving behavior between two populations within the context of GM. While GM was initially designed to target pre-algebra and algebra content, which is typically introduced in middle school, the previous study already confirmed the feasibility of using GM as a formative assessment tool within elementary populations. This study will first replicate this in a high school population. Then it will compare problem solving behaviors of high school students to elementary students. This study serves as an example of how other learning technologies can explore measuring mathematical proficiency across populations while also grounding assessments in theory. In order to accomplish these research goals, this study answers the following research questions:

1. For a high school population, are there latent constructs of mathematical proficiency within the context of *GM*?
2. How do the latent constructions of high school students compare to the constructs revealed within the elementary student data?

To do this, an Exploratory Factor Analysis (EFA) will be conducted to reveal latent constructs of mathematical problem solving behavior within the data. Then the results of the High School EFA will be compared to the results from the Elementary School EFA.

Participants, Experimental Conditions, and Procedures

The high school population included 94 9th grade students from an urban high school. In the study, all participants were asked to engage with *GM* and solve algebraic problems in a puzzle-based activity. All problems in the intervention were goal-based and asked students to transform an expression or equation to reach a certain goal state, rather than asking students to find “x”. After learning how to use the system in a tutorial level, students moved on to a series of levels that targeted specific skills such as basic operations, distribution and factoring, and applying operations to both sides of the equation. All students engaged with these problems for at least 30 minutes and at most 50 minutes. In the last 5 minutes of the study, students were asked to solve a final problem.

This problem was intentionally difficult to test students' persistence. After 2 minutes of working on the final problem, students were given the option to skip the problem or continue working.

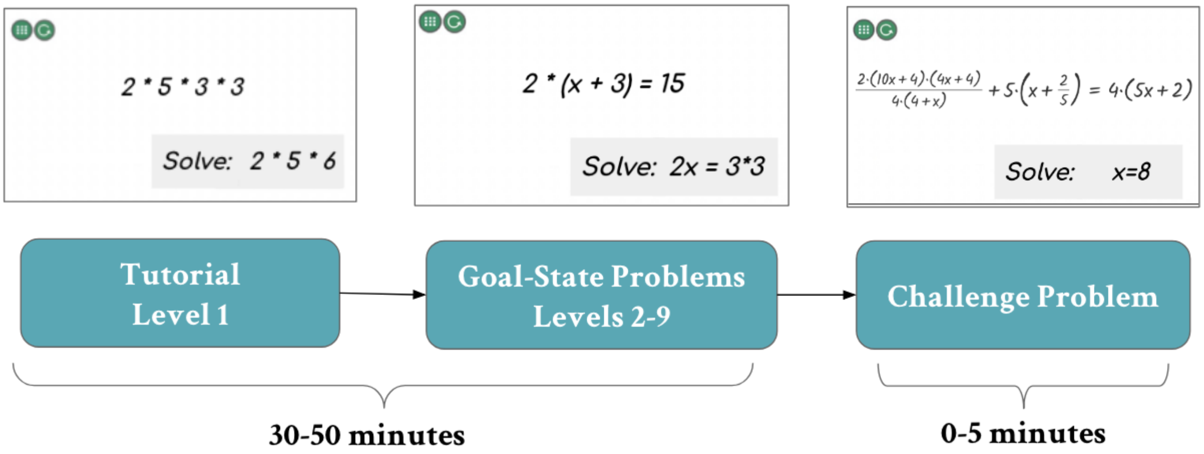


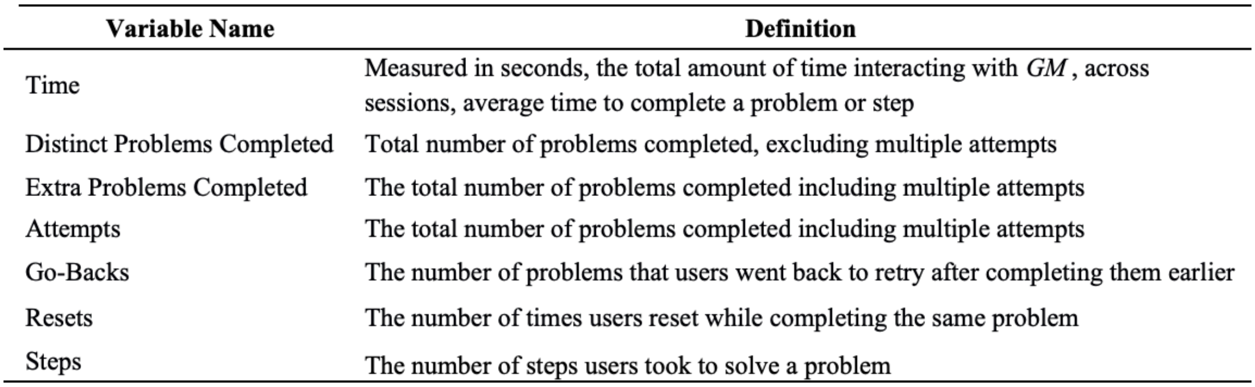
Figure 4-1. *High School Study Design*

The second population, elementary school students, is the same population from the study in Chapter 3. They included 185 second grade students from ten classrooms in three different elementary schools in Massachusetts (116 female, 78 male). The main difference between the study procedure in Elementary and High School populations was the complexity of the problems and the number of sessions. While Elementary students were given a set of problems that focused on expressions, decomposition and basic arithmetic, high school students were given problems that reached higher levels of algebra knowledge, such as distribution, factoring, and equations. In the study procedure, Elementary students interacted with *GM* across three sessions (*m*=78 total problems), while High School students interacted with *GM* in only one session (*m*=36 total problems). Student interactions logged within GM, however, were exactly the same.

Measures

In app process data: GM logs student clickstream data including all actions, mouse clicks, and problem solving steps. These data were aggregated across problems, levels, sessions, and overall. Identical to the method used for the Elementary study, analyzing the High School data started with 31 variables to represent composite measures of student problem solving behaviors. Table 4-1 shows the category types for variables logged in GM.

Table 4-1. *Labels and definitions of primary variables measured in GM*



Variable Name	Definition
Time	Measured in seconds, the total amount of time interacting with *GM* , across sessions, average time to complete a problem or step
Distinct Problems Completed	Total number of problems completed, excluding multiple attempts
Extra Problems Completed	The total number of problems completed including multiple attempts
Attempts	The total number of problems completed including multiple attempts
Go-Backs	The number of problems that users went back to retry after completing them earlier
Resets	The number of times users reset while completing the same problem
Steps	The number of steps users took to solve a problem

Before conducting the EFA, all variables were examined in a correlation matrix to test a few assumptions. It is recommended that all variables should be significantly correlated with at least one other variable (Tabachnick & Fiddell, 2007). It is also recommended that factors should not be correlated above .9, as that would violate assumptions of multicollinearity (Field, 2009). There were a few variables that did not have any recorded data during this study including *Number of Sessions*, *Problems Unlocked, Extra Problems Completed, and Percentage of Extra Problems Completed.* There were two variables that did not correlate with any others in this study, *Percentage of Go-backs* and *Average Resets*. However, three groups of variables with correlations above .9: 1) *Overall Time, Best Time* and *Average Time per Session* 2) *Average Time per Problem* and *Average Time per Minute* as well as 3) *Distinct Problems Completed, Percentage Problems Completed, Percentage of Problems Completed, Total Problems Completed, Completed Stars*, *User Stars, User First Step,* and *Completed Best Step.* We chose to remove the three variables without recorded data, as well as *Percentage of Go-backs* and *Average Resets* as it did not correlate with any other variables. As for the groups of multicollinear variables, we decided to choose one variable from each group to represent the rest. We chose *Overall Time* to represent group 1 (about larger measures of time), *Average Time Per Problem* to represent group 2 (about smaller measures of time), and *Distinct Problems Completed* to represent group 3 (about overall completion rates). This resulted in a total of 11 variables in the final analyses.

Analyses

RQ1 uses EFA to explore underlying constructs in student problem solving behavior in *GM*. To do this, we used 11 student problem solving behavior measures recorded in *GM* and used them as features in the EFA. Descriptive statistics and correlations were calculated for each factor and variables. These descriptive statistics were then analyzed to identify common threads and to explore their potential of mapping onto the five strands of mathematical proficiency.

RQ2 compares the results of the EFA between the elementary and high school students. This includes examining both the factors and the individual items within each factor. While some constructs might be similar, they may be made up of slightly different items. This analysis will also relate the constructs back to the theoretical framework and the five strands of mathematical proficiency to determine if the five strands are present in both populations. This question aims to explore the differences between problem solving behaviors in elementary and high school populations within the context of *GM*, while also grounding *GM* measurement in a theoretical framework.

RESULTS

RQ1: For a high school population, are there latent constructs of mathematical proficiency within the context of GM? An exploratory factor analysis was conducted using a total of 11 variables. To conduct an EFA, multiple criteria need to be reviewed. First, is a sample size of 300 is recommended for factor analyses (Tabachnick & Fiddell, 2007), so the current sample of 87 students is considered somewhat low. Second, KMO test values above .5 are considered appropriate for EFA, with values above .9 considered as excellent (Hutcheson & Sofroniou, 1999). Our KMO was .630, which means that our sample is adequate for producing reliable factors. The Bartlett's test of sphericity was significant $X^2(91) = 1076.53$, $p < .001$, which means that our correlations are significantly different from zero. With these considerations met, our sample was determined suitable for EFA.

Using SPSS 22, Principal Axis Factoring was conducted using a Promax rotation. Promax was chosen as it is an oblique rotation that assumes the factors are correlated. Communalities describe the proportion of variance explained by the underlying factors and values above .5 are considered adequate for factor analysis (MacCallum, Widaman, Zhang, & Hong, 1999) Communalities resulted below .5 for three variables, *Star Score (.263), Total Go-backs (.165),* and *Percentage of Resets (.165)*. These variables were removed and the EFA rerun. After removing those three variables all variables resulted in communalities above .5, except for two, *Percentage of Stars (.427)* and *First Efficiency (.425)*. Since they were above .4, they were included in analyses.

Next, factors were extracted using Kaiser's criterion and by examining the scree plot. Three factors with eigenvalues greater than 1.00 and sufficiently large loadings were extracted and they explained 40.31%, 21.83%, and 18.81% respectively, explaining cumulatively 80.95% of the variance (Table 4-3). The three factors have been classified based on how variables loaded onto each factor. Factor 1 (40.31% of the variance), which included *Total Steps, Distinct Problems Completed, Attempts, Resets, Efficiency, First Efficiency* has been classified as ***Engagement in Problem Solving*** as it incorporates a wide range of problems solving behaviors. Factor 2 (21.83% of the variance), which included *Percentage of Attempts, Average Attempts Completed* has been classified as ***Strategic Flexibility*** as it incorporates only attempt-related items. Factor 3 (18.81% of the variance), which included *Total Time, Time Per Problem,* and *Time Per Step* has been classified as ***Speed*****,** as it includes all time-related items.

Table 4-2. *Principal axis factor loading with Promax rotation and Kaiser normalization*

Item	Engagement	Strategic Flexibility	Speed
Total Steps	0.910		
Distinct Problems Completed	0.847		
Attempts	0.806		
Resets	0.795		
Efficiency	-0.615		
First Efficiency	-0.551		
Percentage Attempts		0.895	
Average Attempts Completed		-0.833	
Total Time			0.799
Time Per Problem			0.642
Time Per Step			0.624
Eigenvalues	4.434	2.402	2.069
Percent of Variance (%)	40.31	21.83	18.81

RQ2: How do the latent constructs of high school students compare to the constructs revealed within the elementary student data? Using the same data from Elementary students as Chapter 3, this analysis compares the similarities and differences between the EFA results in high school and elementary problem solving behaviors within *GM*. Table 4-4. Displays the results from both EFAs side-by-side. Text in black reflects variables and factors included only in the Elementary analysis. Text in Red reflects variables and factors included in the High School analysis, which also overlaps with data in the Elementary data.

While the elementary school data revealed five constructs, the high school data revealed only three. However, those three factors, *Engagement, Strategic Flexibility,* and *Speed*, were defined as factors in both populations. *Engagement* in Elementary school looked like attempting problems, going back to re-attempt problems, and time spent using *GM* overall. *Engagement* in High School, however, was much broader, including attempts, completion, resets, and efficiency. Since there were only three factors defined in the High School data, the *Engagement* factor seemed to incorporate variables from the *Progress* and *Strategic Efficiency* factors that only appeared in the Elementary School data.

Strategic Flexibility in the Elementary School data included four variables, while the High School data only included two variables. Some of the variables in the Elementary School factor of *Strategic Flexibility* were not included in the High School analysis, accounting for some of this difference. In both the Elementary and High School factors, Average attempts completed had a negative relationship with the factor. This suggests that while students were attempting and resetting problems, they were not completing those attempts. This might reflect more of a factor of exploration, rather than strategic flexibility as the number of attempts increased, the number of attempts decreased.

Table 4-3. *EFA comparison between elementary and high school student problem solving behavior within GM*

	Factor							
Variables	Engagement		Progress	Strategic Flexibility		Strategic Efficiency	Speed	
Total Go-Backs	0.935							
Percentage of Attempts	0.908				0.895			
Percent of Go-Backs	0.873							
Number of Attempts	0.778	0.806						
Overall Time Interaction	0.677						0.651	0.799
Distinct Problems Completed		0.847	1.049					
Completed Best Step			1.038					
Extra Problems Completed			0.774					
Percentage of Resets				0.880				
Average Attempts Completed				-0.864	-0.833			
Total Resets		0.795		0.821				
Average Resets				0.784				
Average Time Per Step						0.834	0.647	0.624
User First Step						-0.687		
Percentage Stars		-0.615				0.666		
User Total Step	0.625	0.910				-0.635		
First Efficiency		-0.551	-0.525			0.554		
Average Time Per Problem			-0.417				0.999	0.642
Best Time							0.585	
Eigenvalues	5.65		4.690	4.05		1.68	1.180	
Percent of Variance (%)	29.74		24.700	21.29		8.82	6.220	
Eigenvalues	4.434			2.402			2.069	
Percent of Variance (%)	40.31			21.83			18.810	

Speed was another factor that appeared in both the Elementary and High School data. In the Elementary factor, only Average time per problem and Best time were included, as Overall Time and Average time per step loaded onto other factors more strongly. However, all of these factors (except Best Time, which was not included in the High School analysis) overlap with the High School factor of *Speed*. This suggests that measures of time carry through as a single factor to account for variance in both Elementary and High School populations.

In terms of fit, the Elementary School EFA clearly models the data better than the High School EFA. First, the Kaiser-Meyer-Olkin measure of sampling adequacy for the Elementary school analysis (KMO=.751) was higher than the High School analysis (KMO=.630). This suggests that the Elementary sample has a higher proportion of variance that could be explained by underlying factors. The Bartlett's test of sphericity was significant in both populations, however the chi-square statistic was much higher for the Elementary School population, $X^2(171) = 5877.65$, $p<.001$, compared to the High School population $X^2(91) = 1076.53$, $p<.001$. This suggests that the Elementary School data might be better able to detect separate factors that would be more useful in terms of interpretation. Lastly, after conducting the EFA, the Elementary School model accounted for 90.78% of the variance, compared to only 80.95% in the High School model. While both analyses had adequate fit for explaining the variance in problem solving behavior within GM, the Elementary School EFA was a better model than the High School EFA overall.

Discussion

The current study explored differences between elementary and high school students in latent constructs of problem solving behaviors within *GM*. First, an exploratory factor analysis was conducted to reveal three factors in the high school data, *Engagement in Problem Solving,* *Strategic Flexibility, and Speed* (RQ1). Next, this EFA model was compared to that of the elementary school students, which revealed five factors, *Engagement in Problem Solving, Progress, Strategic Flexibility, Strategic Efficiency,* and *Speed* (RQ2). The first major difference between the elementary and high school problem solving behaviors was the difference in the number of factors. This result could mean a few different things. First, this might mean that there is less variability in problem solving behavior in high school compared to elementary school. This is likely because elementary students might be more prone to exploring with the technology. Further analysis revealed that even though the high school students had more complex problems in terms of the number of steps required to solve, elementary students took more steps on average (m=5.2) to solve than high school students (m=3.8, $p<.01$). Another possible explanation is that this difference in factors could be due to a different number of starting variables. Some of the variables recorded for the elementary school study, like go-backs and certain measures of resets, were not available for the high school data.

Despite a different number of factors, the variables in each EFA were categorized in similar ways. Three of the factors were labeled with the same name, *Engagement in problem solving, Strategic Flexibility,* and *Speed*, because they were so similar in both populations. The only two factors that appeared in the elementary data that did not appear in the high school data were *Progress* and *Strategic Efficiency.* The variables within these two factors in the elementary data were absorbed by the *Engagement in Problem Solving* factor within the high school data. Another variable that was recorded in the elementary data, but not the high school data, was extra problems completed. In the elementary study, students only had to complete 75% of the problems in each level. Extra problems were considered those attempted or completed beyond the required problems. This variable was not recorded during the high school study, which might have contributed to the *Progress* factor not appearing in the high school data. While most of the variables that loaded onto the *Strategic Efficiency* factor in the elementary data were present in the high school data, they instead loaded onto *Engagement in Problem Solving.* As seen in Table xx, a few of these variables have reverse signs. For example, User Total Step has a negative factor loading in the elementary factor, *Strategic Efficiency*, it has a positive coefficient in the high school factor, *Engagement in Problem Solving.* This makes sense as more less steps would contribute to more strategic efficiency (negative sign, negative relationship) and more steps would contribute to more engagement in problem solving (positive sign, positive relationship).

Ultimately this study takes the first step in identifying constructs of mathematical problem solving behavior across K-12 populations within the context of *GM*. While the two populations resulted in a different number of latent constructs, these constructs still resembled similar measures, such as engagement in problem solving, strategic flexibility, and speed. This shows some potential for measuring components of mathematical problem solving across different age groups and only through student interactions with *GM*. The next step in this work is addressing how these latent constructs map onto the conceptual framework proposed in Chapter 2 of this work. Also, the two studies in this section only use one method for extracting components of student problem solving behavior. EFA is a logical first choice because it uses all data to extract latent factors in the data. While this is a good method to start mapping student interactions with *GM* onto components of mathematical proficiency, other top-down methods might elicit stronger correlations to the five strands. These limitations and conclusions are addressed in the following chapter.

Chapter 5: Future Work and Conclusions

Elements and ideas presented in this chapter (Comparing Bottom-up and Top-Down Approaches) have been accepted as a research report to the 2019 National Council of Teachers of Mathematics Research Conference.

Citation:

Hulse, T., Harrison, A, Manzo, D, & Ottmar, E (2019). Developing Measures of Mathematical Proficiency in a Learning Technology. Research Report presented at the Annual Research Meeting of the National Council of Teachers of Mathematics (NCTM).

This chapter will extend the results from the last chapter by mapping the EFA constructs from the Elementary and High School populations onto the conceptual framework presented in Chapter 2. This chapter also describes how future work will address the limitations to the study in Chapter 4, including work that has already been done to compare multiple approaches to measuring mathematical proficiency.

Mapping onto Mathematical Proficiency

In addition to comparing populations, another aim of this work is to explore the theoretical framework of mathematical proficiency in the context of problem solving behavior from clickstream data. Using a bottom-up approach, such as an EFA, are the five strands of mathematical proficiency present? To explore the strands in terms of GM-based measures, let's revisit the conceptual model proposed in the introduction (Figure 2-3). In the elementary school population, five latent factors were revealed, *Engagement in Problem Solving, Progress, Strategic Flexibility, Strategic Efficiency,* and *Speed*. In the high school population, three latent factors were revealed, *Engagement in Problem Solving, Strategic Flexibility,* and *Speed*. The two factors that were only present in the elementary school population, *Progress* and *Strategic Efficiency*, seemed to all load onto the *Engagement in Problem Solving* factor in the high school population. Based on the factor loadings, data structure, and the constraints of the task, these data only clearly mapped onto procedural fluency and strategic competence. It can also be argued that one of the factors could be related to productive disposition.

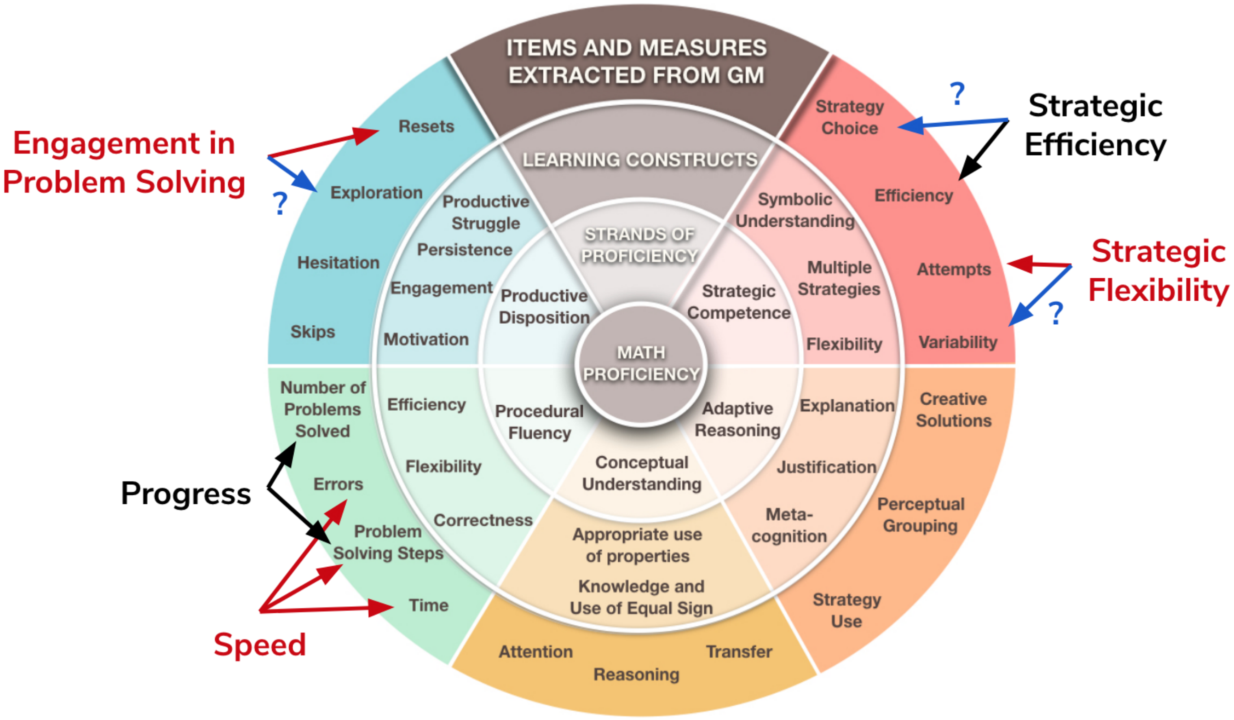


Figure 5-1. *Conceptual model of how strands of mathematical proficiency and EFA constructs map onto GM-based measures.*

Based on the factor loadings, *Progress* (Elementary only) and *Speed* (Elementary and High School) closely resemble the hypothesized measures of Procedural Fluency. *Progress* included problems completed, best step overall, and extra problems completed, which maps onto the hypothesized measures of number of problems solved and problem solving steps in the model. *Speed* may also be a factor of procedural fluency based on its factor loadings of time interaction, time per step, time per problem, and first efficiency (efficiency as measured by the first attempt of every problem).

In terms of strategic competence, which included hypothesized measures of strategy choice, efficiency, attempts, and variability, *Strategic Efficiency* (Elementary) and *Strategic Flexibility* (Elementary and High School) were closely related. *Strategic Efficiency* included efficiency-related measures such as time per step, number of steps to solve all problems on the first attempt (User First Step), number of steps to solve compared to the minimum required to solve (Stars), the total number of steps used on all problems and attempts (User Total Step), and the number of steps to solve the problem on the first attempt (First Efficiency). *Strategic Flexibility* included at least the average number of attempts and percentage of attempts completed in both populations, as well as reset-related measures in the elementary population. While *Strategic Efficiency* reflects efficiency and strategy choice as measured by a lower number of steps, *Strategic Flexibility* actually reflects variability in strategy which would be measured by a higher number of attempts. This suggests that there are two components of the strand Strategic Competence that are competing in terms of efficiency and exploration. In order to learn how to be efficient and learn how to be strategic in terms of their problem solving approach, students will need to engage in a trial and error process, which can be measured through their attempts in *GM*. This is an important finding because it reflects how these formative measures might differ from summative measures.

While an end of the unit exam might assess students' "learned" strategic competence in terms of their efficiency, formative assessment within *GM* might focus more on the trial and error process and aim to engage students in exploring strategies. In terms of measurement, this may mean that students are less efficient, however, they might gain multiple strategy usage for future problem solving.

Productive Disposition is defined as self-efficacy, motivation, and the ability to see the utility of mathematics. In terms of *GM* this was hypothesized as resets and exploration to reflect engaged behavior and hesitation and skips to reflect potential challenges and opportunities to persist. One factor within the elementary and high school populations, *Engagement in Problem Solving*, partially mapped onto the hypothesized measures of Productive Disposition. This factor was made up of go-backs, attempts, overall time interaction, total steps in the elementary population and attempts, problems completed, resets, and efficiency in the high school population. This only partially maps onto Productive Disposition for a few reasons. First, the factor is not completely consistent across both populations. This is a limitation as it may vary with each population. Second, *Engagement in Problem Solving* only reflects components of Productive Disposition such as motivation and engagement with problem solving, however ignores other components such as self-efficacy and perceived utility of mathematics. Measuring these components would only be possibly through means other than problem-solving measures within *GM*.

It was interesting that two of the factors in the high school, *Engagement in Problem Solving* and *Strategic Exploration*, data represent behaviors that relate to exploration and slow practice. Only one factor, *Speed*, relates to problem solving that is more efficient and focused on speedy practice. It was surprising that the construct of *Strategic Efficiency* was not present as its own factor in the high school data. Student who were strategically efficient could exist, but they would probably be present in the data only as having low levels of *Engagement in Problem Solving*, which could also be mistaken for students who are off-task or confused. The idea of efficiency vs. exploration has important implications on research in learning. Using this bottom-up approach, there is not a clear construct of strategic efficiency in the high school data. This could have negative implications for researching this population as it might overlook this important behavior in students. If learning technologies are to capture all five theoretical strands of mathematical proficiency, a more top-down approach might be necessary. Instead of only constructing factors based on latent constructs, grounding assessment on theory might involve designing features based on the theoretical definition of the five strands.

Future Work

One of the major limitations in the two studies presented in Part 2 of this work was that the content and method of analysis both had constraints in terms of their potential to measure all five strands of mathematical proficiency. This method of analysis was a purely bottom-up approach as EFA uses all available data to reveal underlying constructs. While this is a great first step and method to uncover trends, it is ultimately driven by the data, not the theoretical framework. Future work should also compare bottom-up and top-down approaches that start with the theoretical framework to define strands of mathematical proficiency.

Related work has taken the first step in this process by conducting a study that compares the problem solving behavior of high school and college students engaging with *GM* goal-state problems. In that method, three of the five strands of mathematical proficiency were defined using *GM*-based measures (Table 5-1). The three strands, ***procedural fluency, strategic competence,*** and ***productive disposition***, were chosen as a starting point because they do not need verbal explanation as a measure, whereas *conceptual understanding* and *adaptive reasoning* by definition require explanation of conceptual thought.

Table 5-1. *Theoretical mapping of three strands of mathematical proficiency onto measures within GM.*

Strand	**Definition (NCTM, 2001)**	**Measurement in *Graspable Math***
Procedural Fluency (RQ3)	Carrying out procedures accurately and efficiently	Number of problems solved Time to solve
Strategic Competence (RQ4)	Solving problems flexibly and with multiple strategies	Number of steps to solve Resets and Go-Backs
Productive Disposition (RQ5)	Motivation and persistence in problem solving	Time spent on challenge problem Steps taken to solve challenge problem

Procedural fluency in *GM* can be seen as a user's efficiency in solving a problem, which can be measured by the amount of time to solve a problem. Procedural fluency can also account for the number of distinct problems solved as users with higher procedural knowledge would be expected to progress further through the system than users with lower procedural knowledge. Independent samples t-tests determined that there are statistically significant differences between the mean high school (M=45.49) and college (M=75.67) distinct problems completed at the p<0.05 Level. The higher completion and clear rates at the college Level suggested that college students were also more efficient in solving problems in terms of time. To examine this further, independent samples t-tests determined significant differences between mean speed per problem and mean speed per step in high school (M=49.26, 16.79) and college populations (M=27.25, 10.57) at the p<0.05 Level. While all Levels were included for the analyses for total problems completed, only Levels 1-3 were included in the analyses for speed per problem and step, as this was the last Level where each population had at close to 50% of students.

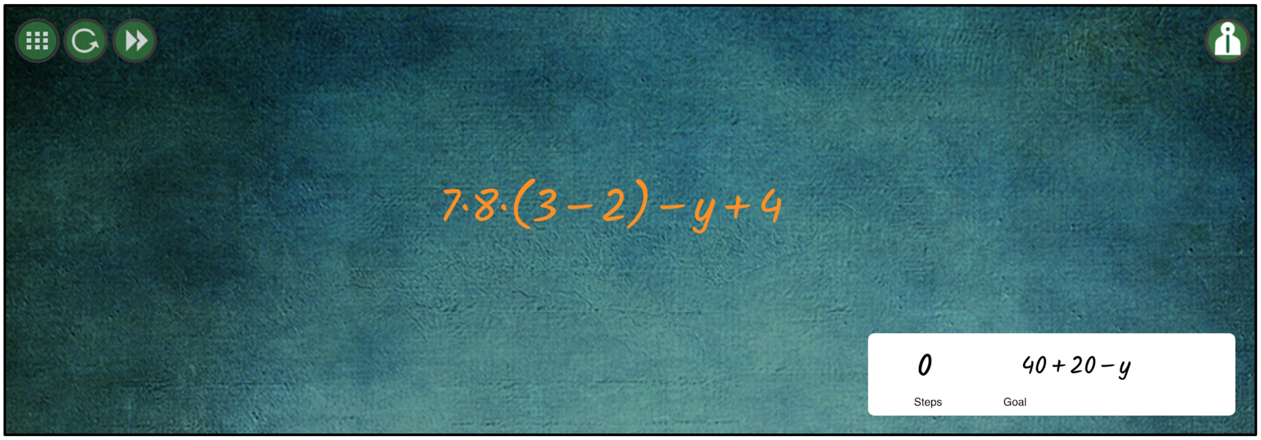
Strategic competence in *GM* was defined by behaviors related to a user's efficiency and flexibility or variability in strategy use. Independent samples t-tests determined differences in population means on efficiency (steps) and retrying behavior. In terms of efficiency, the college population (M=2.93) used significantly fewer steps to solve problems on average than the high school population (M=3.89) at the p<0.05 Level. To capture retrying behavior in *GM*, we measured the number of resets per problem and the number of times students go back to problems they have already attempted. Descriptive statistics suggest that both forms of retrying behavior were considerably low for both populations resulting in resetting in only about 10% of problems and only going back to an average of three problems for a second attempt or more. Independent samples t-tests indicate that there are no statistically significant differences in retrying behavior (resets, gobacks) between high school (M=10.19, 3.2) and college students (M=9.34, 3.14) who used *GM* in this study. While all Levels were included for the analyses for total resets and go-backs, only Levels 1-3 were included in the analyses for speed per problem and step, as this was the last Level where each population had at close to 50% of students.

This study also aimed to measure certain components of productive disposition, such as persistence. In order to measure this, students must be presented with challenging situations. In this study, all students were presented with a final challenge problem that was expected to be challenging even for adults and math "experts". Productive disposition during this challenge problem is measured in the amount of time that students persisted in working with the problem as well as the amount of steps taken in an attempt to solve the problem. Time alone would not be sufficient as a student could have the problem open on their screen without actively working on the problem. Independent samples t-tests identified significant differences in both time and steps on the final challenge problem at the $p<0.05$ Level. college students spent significantly more time and took significantly more steps (M=337.78 seconds, M=37.40 steps) than high school students (M=300.83 seconds, M=25.52 steps).

Overall, college students completed more problems, cleared more levels, and solved problems more quickly than high school students (*Procedural Fluency*). This suggests that college students display more procedural fluency than high school students in these populations. This was to be expected as the high school students were in 9th grade and most were below grade-Level in math performance while the college population were from a high-performing engineering school. What is perhaps more interesting, however, is how these two populations exhibit strategy use (*Strategic Competence*) and persistence (*Productive Disposition*). While college students were more efficient in terms of the number of steps to solve a problem, there was no difference in retrying behavior between populations, showing that very few students in both populations attempted a problem more than once. This suggests that experts and novices in both populations generally do not try multiple strategies per problem and typically stick with the results of their initial attempt. During the challenge problem, college students persisted more as measured by time and the number of steps taken. One possible explanation for this finding is that the college students' expertise may play a role in making it easier to persist in the challenge problem either due to the difficulty of the content or motivation to continue. Ultimately, this research demonstrates the potential of using *GM* as a platform for measuring multiple strands of mathematical proficiency with a top-down formative measure of assessment.

Another major limitation of this work is that all of the data comes from one type of activity. Both the elementary and high school studies are based on the goal-state activity. This is where students are given a starting expression or equation and are asked to transform it into a designated goal-state (Figure 5-2). As students solve these problems, GM can measure the steps students take, the time it takes to solve (per problem, per step), the number of times students reset problems, the number of problems students go-back to solve problems again, the overall number of problems solved, and combinations of those variables. Based on the five strands of mathematical proficiency and the constructs that were revealed using EFA methods, only certain strands could be measured using the clickstream data from a problem like this. Goal-state problems have the potential to measure procedural fluency based on the how fast the students are in their transformations and how many problems they complete, strategic competence based on the efficiency and flexibility of strategy use, and components of Productive Disposition based on student engagement in problem solving. However, other components of Productive Disposition, as well as measures of Adaptive Reasoning and Conceptual Knowledge would need to be assessed using other measures only available with alternative activity types. This issue will be addressed in the following section that describes the iterative design process of development new *GM*-based activities.

Figure 5-2. *Example of a goal-state problem in GM*



The primary goal of this research is to move beyond summative measures of correctness to formative measure of mathematical proficiency. These results give a more intricate perspective on mathematical proficiency compared to traditional summative assessments based on correctness. Instead of simply knowing the percentage of problems student got correct, these measures within *GM* provide the first steps towards teasing apart the 5 strands of mathematical proficiency. These studies serve as an example method of comparing mathematics problem solving behavior in two populations of different age groups, elementary and high school. Also, this study presents preliminary measurements of at least procedural fluency, strategic competence and components of productive disposition, but these have not yet been validated in this system. Now that this work has tested the feasibility of finding statistical differences between expert and novice populations using *GM* measures of progress, efficiency, retrying behavior, and persistence, future work can focus on validating these measures as constructs of procedural fluency, strategic competence, and productive disposition.

This research exemplifies the potential in research that can be done with a learning technology like *GM* to compare population differences at a more fine-grained level. *GM* measures clickstream behavior of users, which allows researchers and in the future, teachers, to access step by step information on mathematics problem solving behavior. This kind of rich research presents an opportunity to tease apart the components of mathematical proficiency across different populations in a method that is more efficient and at a deeper level than traditional summative assessment.

Part 3: Applications for GM in Classrooms

This section uses the theoretical framework from section 1 and the results from section 2 to describe a real-world application of this work. The primary goal of this section is to ***design GM-based tools that are grounded in theory on mathematical proficiency.*** The first chapter details the iterative design process to create GM-based activities that have the potential to measure five components of mathematical proficiency. The final chapter will highlight the potential for this work to be implemented in K-12 classrooms.

Chapter 6: Designing Activities for Promoting and Assessing Mathematical Proficiency

This work has been presented as research reports at the 2019 National Council of Teachers of Mathematics (NCTM) research conference as well as at the 2019 meeting of the International Society for Technology in Education (ISTE).

Citations:

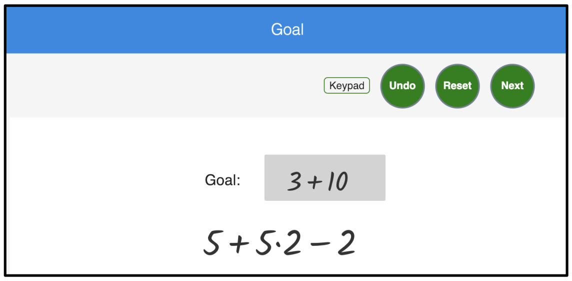
Sawrey, K., Ottmar, E., Hulse, T., Weitnauer, E., Harrison, A. (2019). Exploring Dynamic Learning Technologies for Experiencing Algebraic Notation. Discussion session presented at the National Council of Teachers of Mathematics Research Conference, April, 2019. San Diego, CA.

Weitnauer, E., Hulse, T., Sawrey, K. (2019). Graspable Math: Making Algebra Notation Accessible (and Even Fun!) to Every Student. International Society for Technology in Education. June, 2019. Philadelphia, PA.

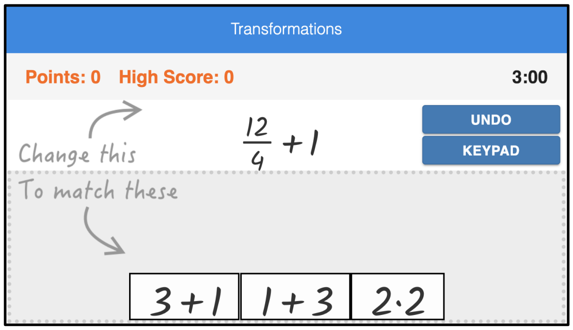
The team received a Small Business Innovation and Research (SBIR) grant to establish the technical merit, feasibility, and educational potential of Graspable Math (GM). Three major components were designed as part of the project, 1) a pre and post assessment of mathematical proficiency, 2) a set of activities to introduce GM to students and develop algebraic understandings, and 3) a platform for teachers and researchers to create their own activities, including a teacher dashboard to display student progress. The main purpose of the assessment and activities are to create a usable learning tool that measures student progress by moving beyond simple correctness and gauging student performance in all 5 strands of mathematical proficiency. The main purpose of the platform was to provide a usable teacher tool that supports classrooms in creating GM activities that meet their curricula needs, as well as presents student progress within a teacher dashboard.

This section will focus on Graspable Activities and how their designs evolved in response to input from the team members, our consultants, teachers, and students during user testing. In terms of our workflow, the core team members followed the Scrum framework, an agile development approach with short feedback cycles. This entails breaking down projects into detailed lists of prioritized tasks for each 2 week development sprint. Additionally to the sprint planning and review meetings, our team met in short daily standup meetings to synchronize our work and to quickly address roadblocks. Throughout the project, we regularly met with an educational consultant, for advice on activity structure and content.

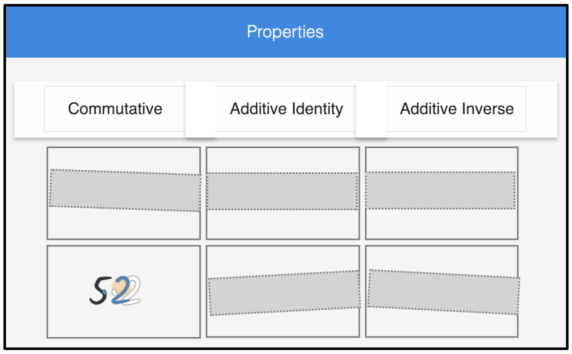
A total of five activity types and two additional template types were implemented. To see each activity played in real-time, visit tiny.cc/graspableactivities.



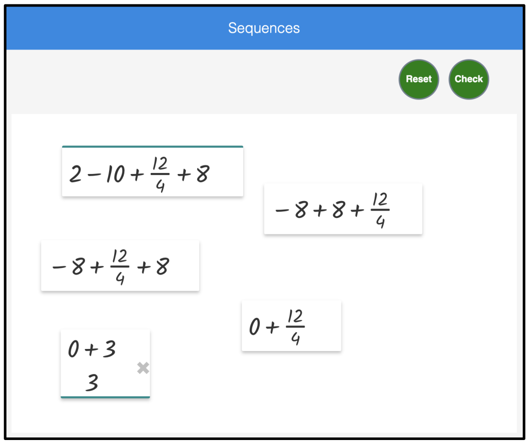
In the **Goal state activity**, users manipulate expressions to match a goal state. This breaks out of the formulaic and repetitive solve-for-x approach and let's students practice strategic flexibility and procedural fluency. The GM algebra notation let's students explore, while preventing them from committing to mathematical mistakes.



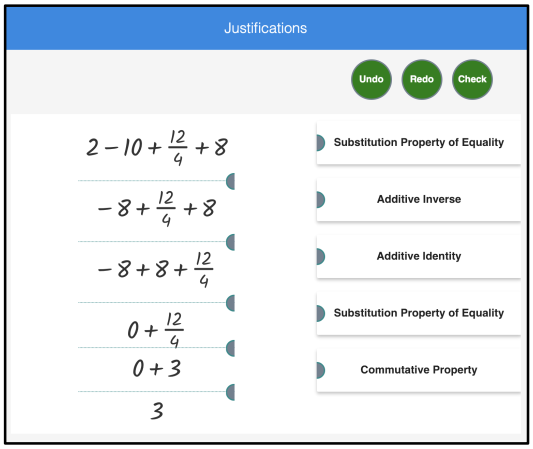
The **Transformations activity** is a quick-paced activity in which students manipulate an expression to clear blocks that appear on the screen. This activity was designed to promote procedural fluency, adaptive reasoning, and to reinforce users' understanding of equivalency.



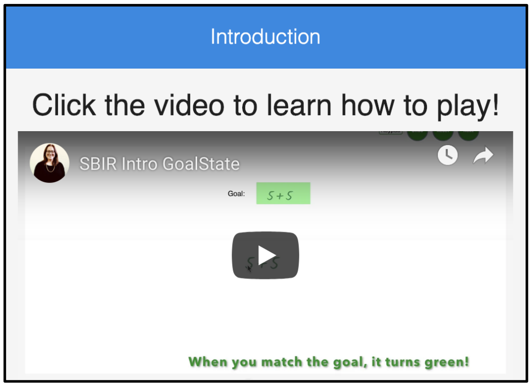
In **Connecting Properties**, students connect property terms to mathematical actions. This activity is designed to bridge the gap from procedural to conceptual knowledge and from computations to relational thinking.



In **Sequence Sort**, users sequence a set of equations or expressions to mathematically connect start and end states. This activity encourages reflecting on algebraic actions, understanding the reasoning of others, and provides scaffolding to solving equations independently.



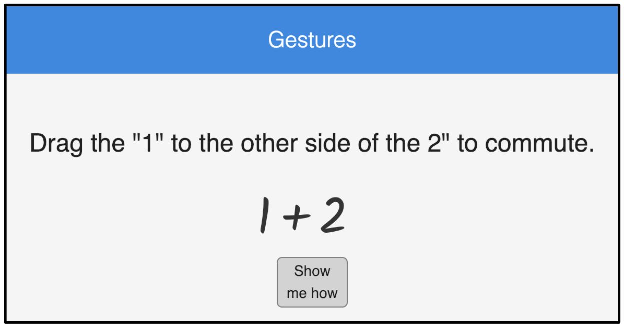
In the **Justification Match**, users are given a derivation and are asked to identify which property (as provided on cards) supports each transformation. This activity is designed to encourage students to generalize algebraic actions, moving from procedural to conceptual understandings.


Introduction

Click the video to learn how to play!

SBIR Intro GoalState

When you match the goal, it turns green!


Gestures

Drag the "1" to the other side of the 2" to commute.

$1+2$

The **Introduction** and **Gesture** activity templates allow content creators to include instructional videos, pictures, text, and GM gesture animations into an assignment. This can be used to teach the GM interface to students or to introduce them to new mathematical concepts and can be part of self-paced learning or flipped classroom instructions.

In terms of iterative development, all of the GM Activities first underwent user testing by undergraduate students. The major aims of this study were to 1) to test initial usability of the prototypes with users, and 2) to solicit feedback and information about how the prototypes could be improved. Participants included undergraduate students who were majoring in STEM fields were used for this study, rather than middle or high school students, because it was conducted over the summer and public school students were not in session. With prior knowledge on the algebraic content, undergraduate students were ideal participants to provide us with both high level (e.g., order of the activities, favorite activities) and low level (e.g., presentation of the progress bar, use of timer) feedback about their experiences. The detailed study procedure is described in the table below. Four WPI students participated in a 1.5 hour interview where they were asked to play all five activities and think-aloud as they were experiencing the system. They were asked to comment on both system ease of use, clarity of the instructions and tasks, as well as to identify any bugs that they found in the system. Data was recorded through a screen and audio recording of the session in addition to notes taken by the researcher who led the study. During these 4 user testing sessions, several initial bugs in the system were identified and many recommendations for improvement were made. These data were reviewed by the development and research teams, as well as the project consultants (mathematics educators) and informed the next iterative development cycle of the product.

Table 6-1. *User Testing Procedure*

Schedule	Content
Study Introduction (5 minutes)	Introducing user testing procedure Consent form
Goal State (15 minutes)	Commuting, Addition, Subtraction, Multiplication, Division, Decomposition, Distribution, Factoring (80 items)
Transformation (15 minutes)	Commuting, Addition, Subtraction, Multiplication, Division, Decomposition, Distribution, Factoring (6 boards, ~80 items)
Derivation Sort (15 minutes)	Commuting, Addition, Subtraction, Multiplication, Division, Decomposition, Distribution, Factoring (10 problems, ~50 steps)
Break (5 minutes)	
Connecting Properties (15 minutes)	Commuting; Adding, Subtracting, Multiplying, Dividing (Simplifying), Splitting, Distributing, Factoring (50 items)
Justification Match (15 minutes)	Simplifying, Splitting, Distributing, Factoring (~45 items)
Final Discussion (5 minutes)	Overall Discussion on usability, gaming the system, and participant background in technology and mathematics Debriefing and Compensation

As a result user testing many initial bugs, issues, and successes were identified in the GM Activities. These user issues and comments ranged from the instructional materials, to the visual cues, to activity structure, to bugs. A detailed list of user comments are described in the table below.

Table 6-2. *Comments made during user testing*

GM Component	User Issue/Comments
Instructional Materials	● Include audio in the video ● Show mouse clicks in the videos ● Goal-State practice screens with 4 problems on one page is too much ● Add undo and reset buttons for practice problems ● The intro for splitting a number is confusing ● More clearly define rules for the transformation game
Saving Progress	● Users want to be able to save their work and go back to problems
Visual cues	● Make sure the progress circles show up as green even when they click next on instructional problems ● When a problem is solved correctly, there are three flashes of green. This takes too long. One flash would suffice. ● The keyboard in the transformation game covers blocks, making it hard to remember what you're solving ● The timing of the animations in connecting properties is not consistent. Some animations are very slow while others are too fast. ● Make the "check" button more obvious in the Derivation Sort and Justification Match activities. Users wait once they connect all items for the system to automatically check.
Activity Structures	● Show a 3-2-1 step counter for when a new block will fall in the transformations ● Give the users the option to retry levels in transformations ● In connecting properties, users might want to be able to click on an incorrect card to see the original problem/animation ● Add a second level to justifications where there are distractor cards or a bottomless number of all cards for users to choose from
Activity Content	● The Goal State problems only went through some of the gestures and might need to be shortened. It took a long time to get through them and users might need less practice to understand the gestures. ● The transformation game problems ranged from too easy to too difficult. Need to find a better balance. ● Do not simplify distribution automatically. Instead of 2(x+3) → 2x + 6, it should be 2(x+3) → 2*x + 2*3 ● Can we find a better word for "balancing equations". This was unfamiliar to users
Bugs	● The keyboard isn't working correctly with pressing buttons on the screen ● In the derivation sort activity a few green lines were missing when users got the problem correct ● In connecting properties, sometimes the animations do not show 2 times

As a result of this user testing, four major changes were made to the design of *Graspable Activities*. These changes were implemented before feasibility testing in high school classrooms.

1. ***Ordering and presentation of the activities:*** In user testing, we initially presented each of the activities to students separately and played with the order in which we should present these activities to students. Throughout development, the activities started to fall into a specific order: *Goal State, Transformations, Connecting Properties, Derivation Sort,* and *Justification Match*. The decision to present the goal state activities first provided students with short gesture tutorials and practice problems that introduced them to and helped train them on the GM gestures. Once they were familiar with the gestures, procedures, and connections to the math, the transformations activity gave them additional opportunities to practice. The connecting properties activity provided the mathematical language for student, while the derivation sort and the justification match tied all of these pieces together.
2. ***Combining 2 activities into 1 connected pair:*** Though all of these activities were originally designed to be independent of each other and serve as standalone products, conversations resulting from user testing revealed that two of the activities were deeply connected. During user testing, users saw at least ten problems of *Derivation Sort* then they saw at least ten different problems of *Justification Match* afterwards. While users said this seemed acceptable and engaging to users, discussions with users, consultants, and the GM team suggested that these two activities might be stronger if they were presented in alternating problems, meaning that users would see one *Derivation Sort* and a corresponding *Justification Match* immediately after. That way, users would see the same problem twice. First, they would sort the shuffled steps to put the derivation in order. Then, they would justify those steps by matching the corresponding property names. This change was made in preparation for the classroom studies.
3. ***Modifications to the Transformation Game:*** It was clear from this study that the *Transformations* activity presented particular challenges in both the design and development in the classroom study. In study 2, the transformations game was very challenging to win or lose this game. There was too much space for blocks to fall in order to lose, but too difficult to win quickly. Anecdotally, many students became stuck in the problems without much motivation to win or lose. During an exit interview with the Study 2 teacher, we asked her opinion on each of the activities. She knew right away that this was the most frustrating. She had an idea that we should keep the blocks falling based on strategy, but also include a timer to indicate the end of the activity. That way students are always "winning" and keep playing to increase their own high score. The *GM* team thought this was a brilliant idea and have implemented it for the second classroom study. Originally we liked the idea of keeping the game as similar to Tetris as possible since it is a simple, but long lasting game that many users would be familiar with. To make our activity like Tetris, we would need to include a timer and base the game on speed. Using a timer with mathematics, however, can induce math anxiety and would likely be more frustrating than fun for many students. Instead, we decided to make blocks fall based on strategic moves. Every three moves a user makes, which could include actions like dragging to commute, tapping to add/multiply, dragging to distribute, pressing the keypad button to decompose or factor, a new block appears. That way students are judged on efficiency in terms of strategy rather than speed.
4. ***Use of Mathematical Language:*** Another concern identified in user testing was the use of mathematical language. The Goal-State, Transformation, and Sequence Sort activities focus on performing mathematical actions and steps, while the Connecting Properties and

Derivation Justification activities asked users to match mathematical terms to their corresponding actions. Much thought was given to the mathematical language used to bridge this gap between procedural and conceptual knowledge. During user testing, the GM team experimented with a variety of labels and terms in regards to varying levels of formality. Some of the terms used were direct property names, like "Distribution", while other terms were taken from common practice like "Simplifying", while other terms were described in terms of GM-based gestures like ". This sparked a long discussion with the team's consultants on the use of mathematical language in classrooms, how this language relates to formal property terms, and how the language should be implemented in Graspable Activities. The most difficult mathematical action to term is what happens when you add 2x+5x to make 7x. Most teachers and even textbooks call this "Combining like terms" or "simplifying". However, this action can be labeled as "Substitution property of equality" because 2x+5x was substituted with 7x. It can also be labeled as "Distribution" because x is distributed in x*(2+5) to get x*(7) or 7x.However, using distribution as a label for the step from 2x+5x to 7x is very distant from how combining like terms is taught in classrooms. For user testing with undergraduate students "Simplifying" was used. For the classroom study, the team took the purest approach and only used terms that were property names, so "Substitution" (for substitution property of equality) was chosen as the final term in this example.

Mapping onto Mathematical Proficiency

Graspable Activities was designed with the five strands of mathematical proficiency in mind. One of the major goals of *Graspable Activities* was to bridge the gap from procedural to conceptual knowledge. The studies in chapter 3 and 4 suggested that only procedural fluency and strategic competence (and potentially elements of productive disposition) could be measures through the goal-state activity. These new activities as part of the SBIR project provide new *GM*-based measures that could potentially measure most, if not all five strands. The new activity types introduced in *Graspable Activities* could also support the gap students when first practicing algebra thinking from arithmetic to generalizations, and from computations to relational thinking.

Figure 6-1. shows hypothetical mappings of *Graspable Activities* onto the conceptual model of *GM* measures of the five strands of mathematical proficiency. Both the goal-state activity and transformation game could measure procedural fluency, as both of these activities ask users to manipulate algebraic notation. The transformation game and derivation sorting could both provide measures of strategic competence, as they require users to engage with sequences of strategy. Connecting properties was specifically designed to connect algebraic transformations with property identification, which maps onto the strand of conceptual knowledge. Derivation justification is the first *GM*-based activity that asks users to engage with algebraic derivations and justify their actions. This activity is designed to map onto measures of adaptive reasoning, which is defined by explanation, justification, and metacognition. Productive disposition is still the most difficult to measure through computation-based practice. All five of these activities have the potential to challenge students and measure their persistence through challenge and engagement in problem solving. However, this still does not measure students' self-efficacy or attitudes towards the utility of mathematics. While these activities cannot yet get at those aspects of productive disposition, the pre and post measures created for the SBIR project included validated measures of student self-efficacy and math attitudes (described in more detail in chapter 7). These measures can then be used to predict problem solving behaviors in *GM*.

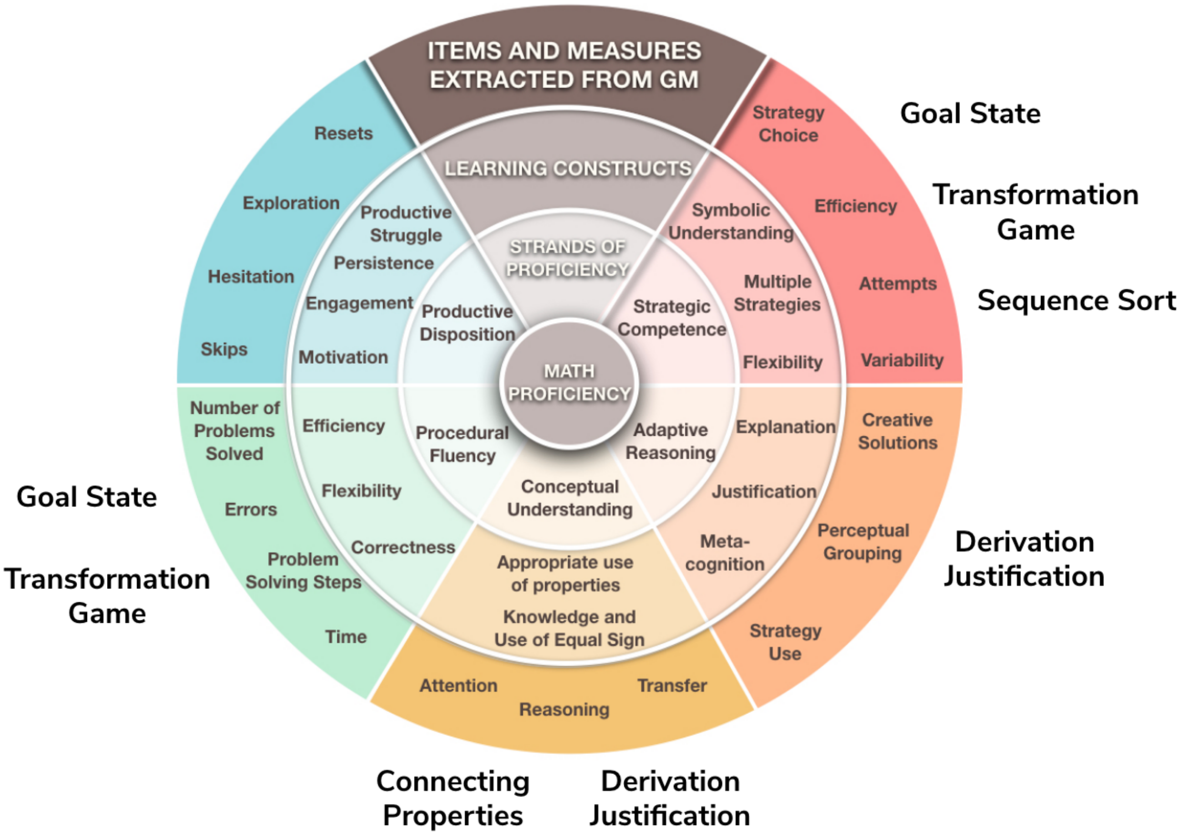


Figure 6-1. *Mapping GM activities onto the conceptual model of the five strands of mathematical proficiency.*

The work accomplished in *Graspable Activities* provides a strong foundation for the development of a final product that is usable, enjoyable, and supports students in a variety of task types that target specific learning goals within algebra. First, the final user studies confirmed that the prototype is usable by students and feasible to implement in classrooms. Students were able to adjust to the novel interface and all activity types quickly, they were highly engaged while working through the content, and the teachers we worked with were very excited about the promise of the app. Second, teacher and student feedback allowed us to refine how to structure teacher training and in-app student tutorials, how to combine tasks into coherent assignments, and how to refine the individual activity templates to address confusion or frustration with the interface. Third, teacher feedback provided us with valuable clues about how *Graspable Activities* fits their needs in the classroom. For example, all of our teachers were excited about being able to create and adjusting tasks themselves, instead of just selecting from preexisting ones. Another example is the specific language used across teachers to describe what they appreciated about the activities, such as promoting productive risk-taking of their students.

The work presented in this chapter suggests that *GM* has great potential for transforming the algebra learning experience. *GM* can be applied in a multitude of activity types that provide classrooms with options for differentiated instruction of algebraic problem solving that is often limited in traditional instruction. In addition to this, these activities were grounded in theory from the initial design and incorporate learning goals that target a variety of components of algebraic thinking and problem solving. *Graspable Activities* not only provides practice in multiple skills of mathematical proficiency in the context of algebra, but also has the potential to measure the five individual strands. In order to fully implement *Graspable Activities* in classrooms as both an instructional and measurement tool of mathematical proficiency, future work needs to be done in order to 1) validate the measures of mathematical proficiency within *GM*, and 2) work with teachers to create a teacher dashboard that displays *GM*-based measures in a way that fits their needs. Both of these areas of future research are discussed in Chapter 7.

Chapter 7: Future Work and Conclusions

After establishing the feasibility of *Graspable Courses* with students in classrooms, the next step is to create reliable measures and usable tools for teachers to utilize those measures. This chapter describes the future work of *GM* in terms of creating a suite of tools that can be easily adopted into classrooms. This includes the design of a pre and post measure of mathematical proficiency that has previously-validated measures. The pre and post measures can then be used evaluate the formative measures recorded through user interactions in *GM* described in the studies of Chapters 3 and 4. This chapter also describes the design for a platform that allows teachers to create their own problem sets through the activity types created for the SBIR project. *GM Courses* includes a course builder to build problem sets, as well as a teacher reports feature that provides teachers with *GM*-based formative measures of assessment. Lastly, this chapter will discuss major conclusions that address the overarching aims of this entire work.

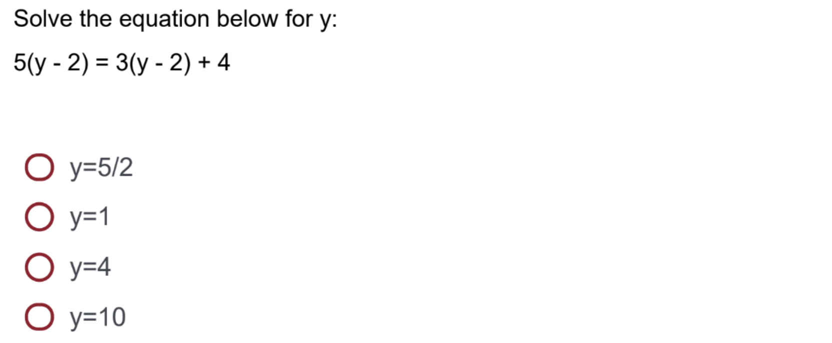
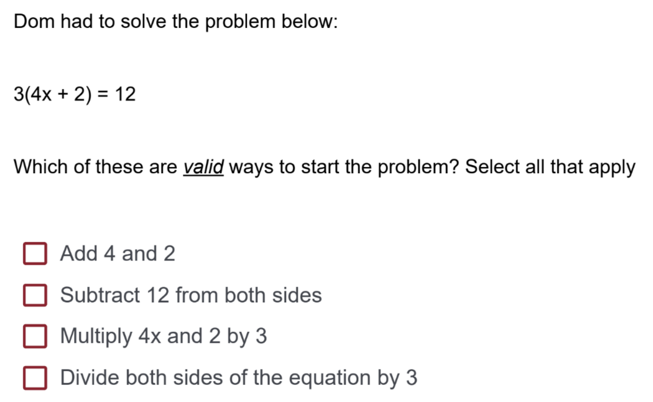
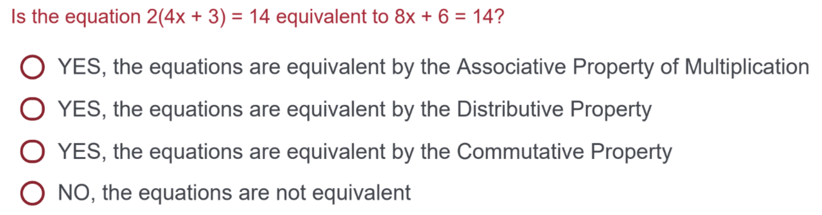
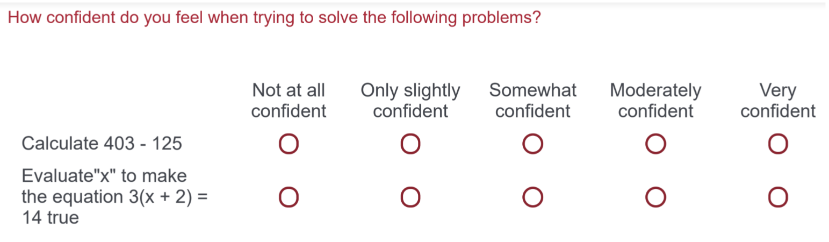
Validating Measures

Chapter 5 mentioned comparing bottom-up and top-down methods of measuring the strands of mathematical proficiency. In order to do this, an analysis should be conducted that compares the factors revealed in the EFA analysis of high school data in Chapter 4 (bottom-up) with the three strands of proficiency defined in this study (top-down). It would be critical to validate these two approaches in terms of how they correlate with or predict previously validated measures of mathematical proficiency. One approach would be to create a pre and post assessment of validated measures to compare to the *GM*-based interaction measures.

As mentioned in the previous chapter, a major component of the SBIR project was to develop pre and post assessments. These assessments were designed to measure mathematical proficiency as well as usability, math anxiety, and self-efficacy. The usability items were based on the system usability scale (SUS, Brooke, 1996)(α~.90). The SUS Likert scale questions were adapted to the SBIR project to measure student perceptions of how easy it was to use *Graspable Activities*. Items included "*I would like to use the activities for math class"* and *"I would need help from someone to work with the activities".* These measures have often been shown to be strong predictors of student math performance (CITE). They were also included because they might map onto elements of productive disposition that include student motivation and affect in the context of mathematics. These pre and post measures can be used in future analyses to correlate with or predict problem solving behavior within *GM*. This may be one method to validate the motivational and affective measures of productive disposition within *GM*.

The four math confidence rating items were designed to measure students' perception on their ability to solve math problems varied in difficulty level (e.g., calculate 403-125, and simplify 5(4+3x)). The five self-efficacy items were adapted from the Academic Efficacy subscale of the Patterns of Adaptive Learning Scales (α=.82;Midgley et al., 2000) designed to measure student perceptions of their ability in mathematics. Items on the self-efficacy scale included *"I know I can learn the skills taught in math this year"* and *"I can do almost all of the work in math if I work hard at it".* The five math anxiety items for this project was based on five items that were adapted from the Student Beliefs about Math Survey (α=.61; Kaya, 2008) (α=.61). designed to measure student perceptions of their own anxieties towards mathematics. Items on the math anxiety scale included "*I feel nervous before a math test"* and *"I can't sit still when I do math".* These measures have been used in our prior work and have been found to predict algebra learning in our system (Ottmar, Landy, & Goldstone, 2012; Ottmar & Landy, 2017).

Table 7-1. *Example Items from the pre and post assessment of mathematical proficiency*

<table>
<tr><th>Strand</th><th>Example Item</th></tr>
<tr><td>Procedural Fluency</td><td>
Solve the equation below for y:
5(y - 2) = 3(y - 2) + 4
○ y=5/2
○ y=1
○ y=4
○ y=10</td></tr>
<tr><td>Strategic Flexibility/ Adaptive Reasoning</td><td>
Dom had to solve the problem below:
3(4x + 2) = 12
Which of these are <i><u>valid</u></i> ways to start the problem? Select all that apply
□ Add 4 and 2
□ Subtract 12 from both sides
□ Multiply 4x and 2 by 3
□ Divide both sides of the equation by 3</td></tr>
<tr><td>Conceptual Understanding</td><td>
Is the equation 2(4x + 3) = 14 equivalent to 8x + 6 = 14?
○ YES, the equations are equivalent by the Associative Property of Multiplication
○ YES, the equations are equivalent by the Distributive Property
○ YES, the equations are equivalent by the Commutative Property
○ NO, the equations are not equivalent</td></tr>
<tr><td>Productive Disposition</td><td>
How confident do you feel when trying to solve the following problems?
<table>
<tr><th></th><th>Not at all confident</th><th>Only slightly confident</th><th>Somewhat confident</th><th>Moderately confident</th><th>Very confident</th></tr>
<tr><td>Calculate 403 - 125</td><td>○</td><td>○</td><td>○</td><td>○</td><td>○</td></tr>
<tr><td>Evaluate"x" to make the equation 3(x + 2) = 14 true</td><td>○</td><td>○</td><td>○</td><td>○</td><td>○</td></tr>
</table></td></tr>
</table>

Students' mathematical understanding was assessed before and after the GM sessions. The pre- and post-test was each composed of nine items adapted from a previously validated measure of procedural fluency, flexibility, and conceptual understanding in algebra (Star, Rittle-Johnson & Durkin, 2016). The post-test closely mirrored the pre-test but with different numbers in the algebraic problems. Though this measure is considered a standard in the field, it needed to be extended for the scope of this project. This scale was not designed to measure the strand of productive disposition, so this project used measures of productive disposition from a previously validated scale developed by Samuelsson (2008). This scale was validated with x-grade students in Swedish classrooms. Since this project combined two separate surveys for mathematical proficiency, some modifications needed to be made. First, the original measures from both of the the previously validated measures were much longer, so for the purposes of this project only certain items were chosen. The goal was to have 10 problems to assess mathematical proficiency. Second, it is important to note that only four of the five distinct strands of mathematical proficiency are included in the pre/post assessment of this project. Since the definitions of strategic flexibility and adaptive reasoning as proposed by NCTM (2001) are so similar that in this project they were collapsed and considered together under the Star, Rittle-Johnson, and Newton measure of flexibility. Samples of each measure are presented in Table 7-1.

Based on the initial classroom study, a few minor changes were made to the pre and post assessments for the second classroom study.

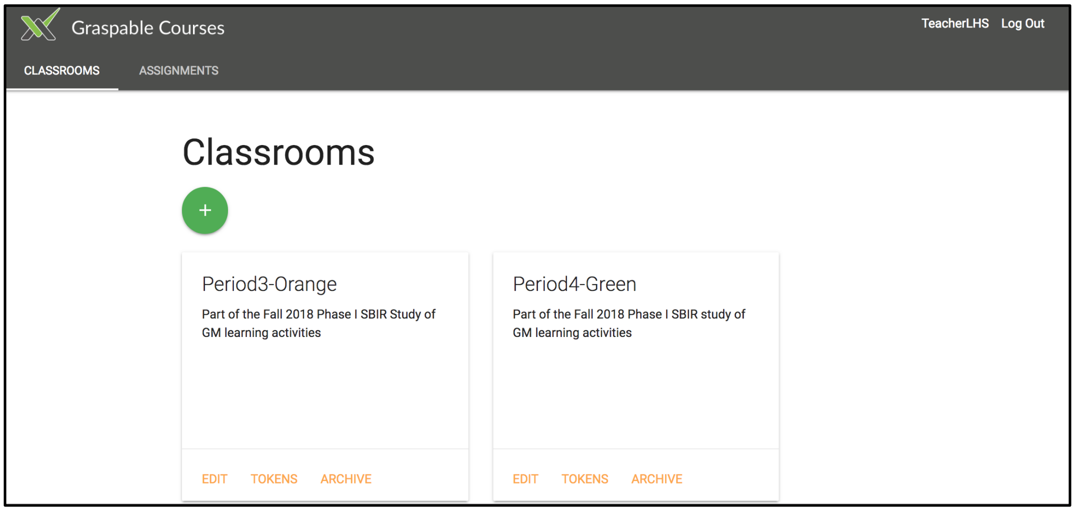
1. ***Ordering of affective measures:*** The timing of the second study was altered so that the math anxiety and self-efficacy items were included in the pre-study surveys. This change was made in order to run analyses that use these items as predictors of behaviors within *Graspable Activities* or learning gains from pre to post. This will allow the *GM* team to answer research questions like "*Do students with varying levels of math anxiety benefit more from using Graspable Activities?"*, or "*Do students with higher self-efficacy retry problems more than students with lower self-efficacy?"*.
2. ***Increasing reliability of productive disposition measures:*** Items used to measure productive disposition were slightly altered to better match the content students experienced in *Graspable Activities*. The original items asked students to rate their confidence on performing calculations like "*403-125"* and *"12-3=__+5*. In the first classroom study, the average response to all of these items was over 4 (out of 5), indicating a ceiling effect. To address this, items were modified to ask students to rate their confidence on performing calculations like "*7*(53-27)*" and "*15-3=2*(__+5)*.

In the final user study, 100% of students attempted at least 75% of both the pre and post surveys. This is partly due to the implementation of all pre and post tests online using Qualtrics. In previous high school studies, using paper and pencil pre and post tests resulted in only 35% of students attempting 75% of the problems. This kind of attrition is typical of classroom studies, but makes analysis of pre to post learning gains challenging. Learning from that previous study, the classrooms studies administered all pre and post surveys online, as well as kept the surveys short, including just 10 math problems. In addition to successful completion rates, many of the pre and post measures showed high reliability, including the *Math Self Efficacy* and *Math Anxiety* items (α = .92) and *Usability* items (α =.89). However, the reliability of the mathematics knowledge items were not ideal (α = .56). This suggests that some of the items did not reliably measure student knowledge in our population. This work highlights the reality that previously-validated measures will not always generalize to all populations. Future work should focus on adjusting these measures, which were previously-validated in other populations, so that they can be validated in the target population for this research. Once the pre and post measures are found to be reliable, then future work can continue in validating *GM* in-app problem solving behaviors as measures of individual strands of mathematical proficiency.

Implementation in Classrooms

In order for *GM* to be adopted into classrooms, it needs to serve the needs of both the teachers and the students. In the introduction of this work, a Venn Diagram was presented on the relationships between teachers, students, and devices. When focusing on teachers, learning technologies were identified as having an effective role in instruction, data collection, and classroom management. In order to make *GM* easily adoptable into classroom, its design needs to focus on supporting the teacher in those roles. The final component of the SBIR project, *Graspable Courses,* is a a tool specifically designed for teachers and is situated in those roles of instruction, data collection, and classroom management. The major goals for this platform were to be able to create and share activities to students and access the teacher dashboard that is linked to student progress reports. Currently, *Graspable Courses*, is still in the prototype phase and has tested with two teachers. Ideally, once in-app measures of mathematical proficiency are validated, they can be shared with teachers through the *Graspable Courses* dashboard. This section first describes the design features of *Graspable Courses* as a creation tool and how the dashboard could be used to display formative measures recorded by *GM* and be used to inform instruction.

In *Graspable Courses*, each teacher can create a unique account that enables them to save both classrooms and assignments (Figure 7-1). The tabs in the top left corner allow users to switch from classroom view to assignment view. Teacher can then link assignments to specific classrooms. Currently, the options for each classroom include edit, tokens, and archive. Editing a classroom allows the teacher to assign lessons. Tokens are what students need to log in. With each classroom study, we assign tokens with a random combination of numbers so that students can remain anonymous. In the future, students will have their own accounts and logins to create themselves. The archive option is there to remove any classrooms that are no longer relevant to the teacher.



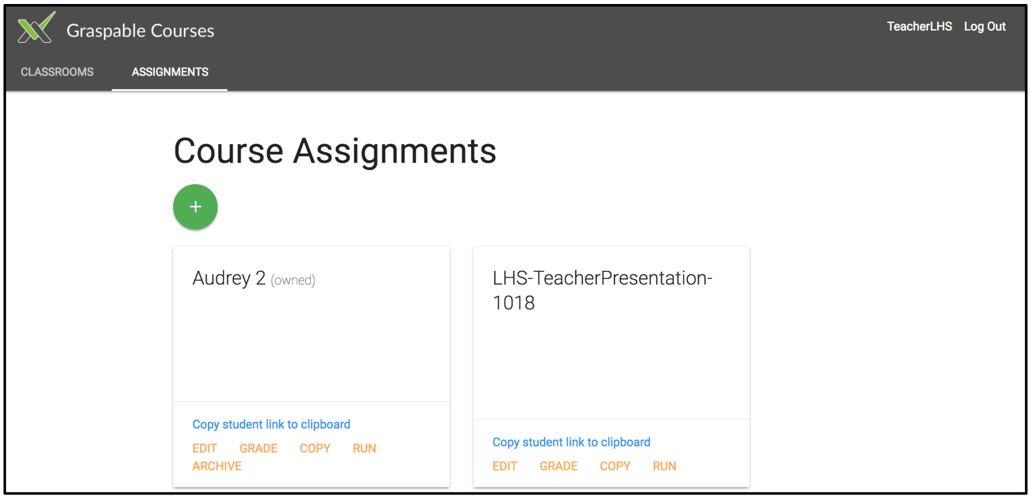
Period3-Orange

Part of the Fall 2018 Phase I SBIR Study of GM learning activities

Period4-Green

Part of the Fall 2018 Phase I SBIR study of GM learning activities

***Figure 7-1.** Graspable Courses Classrooms screen.*



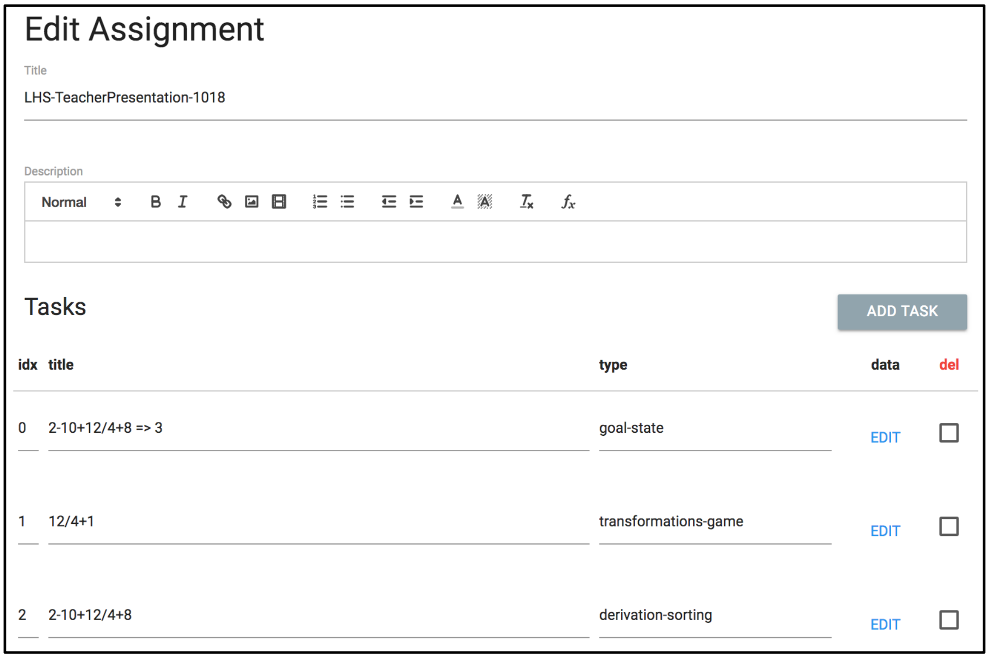
Course Assignments

Audrey 2 (owned)

LHS-TeacherPresentation-1018

Figure 7-2. *Graspable Courses Assignment screen.*

There are also options for each assignment: edit, grade, copy, run (Figure 7-2). These are also useful for teachers to create and edit assignments, as well as to access the grades dashboard. Edit allows the teacher to access the assignment editing tool (Figure 7-3). In this tool, teachers can manipulate the title and overall order of the tasks. Teachers can click add task for a new item and choose between the five *Graspable Activities*, an informational screen with text or video, or animation presentation. Within these options teacher can "copy" assignments, which makes it easier to modify their lessons. They can also "run" assignments to test out how the assignment would look in student view.



Edit Assignment

Title
LHS-TeacherPresentation-1018

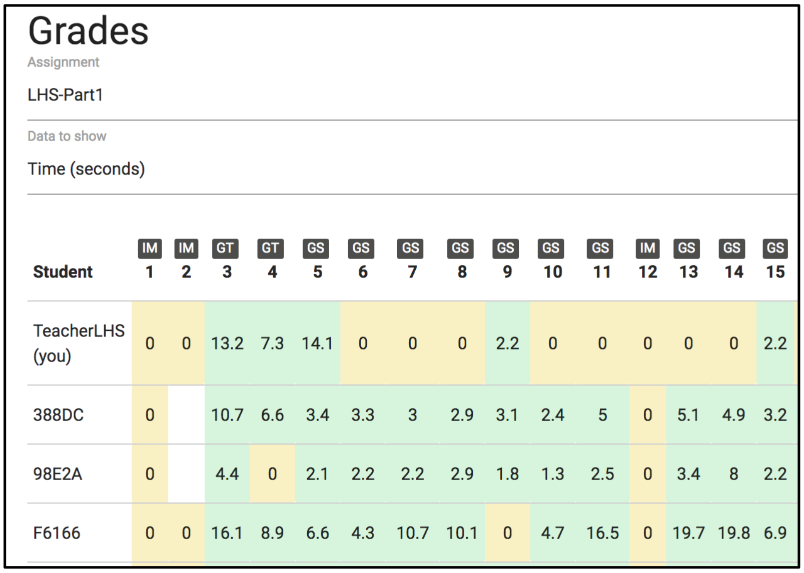
Description

Tasks

idx	title	type	data	del
0	2-10+12/4+8 => 3	goal-state	EDIT	□
1	12/4+1	transformations-game	EDIT	□
2	2-10+12/4+8	derivation-sorting	EDIT	□

Figure 7-3. *Graspable Courses assignment editing tool.*

The last option is “grade”, which brings teachers to the dashboard that shows student progress through the assignment. The classroom view is presented in Figure 7-4. In the top half of the screen, teachers can view the assignment name and the data shown. In the bottom half of the screen, the students are shown (via their randomized tokens) on the left column with each problem presented on the top row. The very top icons (IM, GT, GS) are indicators of problem type, e.g. GS means a Goal State problem. Teachers can choose which data they want to see from a dropdown list (Figure 7-5). Currently the data options are time, errors, completion, total steps taken, and resets. However, there is much more data being logged that could be made available if the teachers in our professional development session find it useful. Descriptions of the different types of data is also available by clicking an information icon on the classroom view screen.



Grades

Assignment

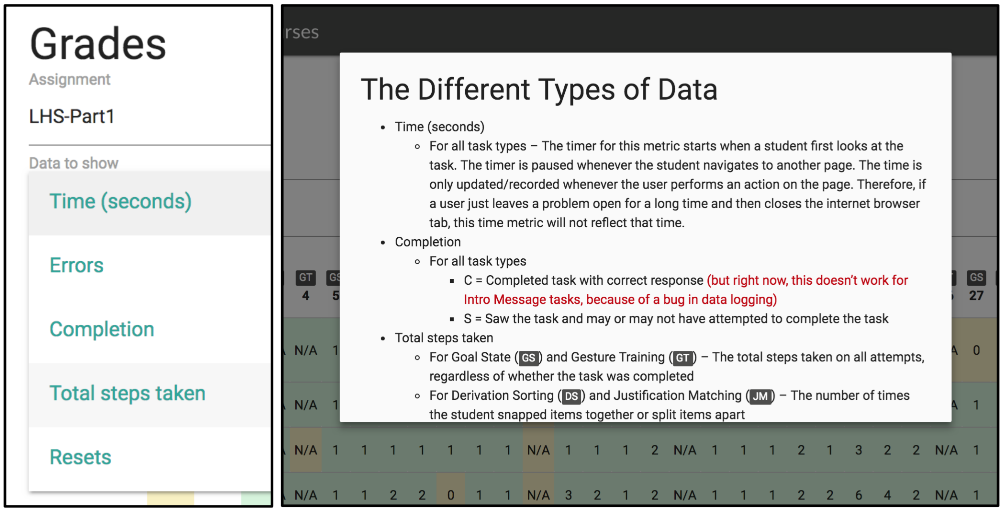
LHS-Part1

Data to show

Time (seconds)

Student	IM 1	IM 2	GT 3	GT 4	GS 5	GS 6	GS 7	GS 8	GS 9	GS 10	GS 11	IM 12	GS 13	GS 14	GS 15
TeacherLHS (you)	0	0	13.2	7.3	14.1	0	0	0	2.2	0	0	0	0	0	2.2
388DC	0		10.7	6.6	3.4	3.3	3	2.9	3.1	2.4	5	0	5.1	4.9	3.2
98E2A	0		4.4	0	2.1	2.2	2.2	2.9	1.8	1.3	2.5	0	3.4	8	2.2
F6166	0	0	16.1	8.9	6.6	4.3	10.7	10.1	0	4.7	16.5	0	19.7	19.8	6.9

Figure 7-4. *Graspable Courses teacher dashboard for student progress*



Grades

Assignment

LHS-Part1

Data to show

- Time (seconds)
- Errors
- Completion
- Total steps taken
- Resets

The Different Types of Data

- Time (seconds)
  - For all task types – The timer for this metric starts when a student first looks at the task. The timer is paused whenever the student navigates to another page. The time is only updated/recorded whenever the user performs an action on the page. Therefore, if a user just leaves a problem open for a long time and then closes the internet browser tab, this time metric will not reflect that time.
- Completion
  - For all task types
    - C = Completed task with correct response (but right now, this doesn’t work for Intro Message tasks, because of a bug in data logging)
    - S = Saw the task and may or may not have attempted to complete the task
- Total steps taken
  - For Goal State (GS) and Gesture Training (GT) – The total steps taken on all attempts, regardless of whether the task was completed
  - For Derivation Sorting (DS) and Justification Matching (JM) – The number of times the student snapped items together or split items apart

Figure 7-5. *Graspable Courses options for viewing student progress within the teacher dashboard*

The teacher dashboard is continually being developed. Teacher input on the functionality and usability of *Graspable Courses* is invaluable to the project and will continue to drive the design of the dashboard. Once the formative measures of mathematical proficiency are validated with the pre and post tests, they will be implemented as an option for teachers to display in their reports. This option could give teachers insight on how each individual student, as well as the class as a whole, is developing each strand of mathematical proficiency. For example, the dashboard could notify teachers that the class is generally high on measures of procedural fluency and strategic competence, but they lack in their ability to justify why their strategies are effective (adaptive reasoning). The teacher could then use this information to drive their instruction for the next day. The teacher may decide to add a ten minute activity to the beginning of class to practice adaptive reasoning, such as the Derivation Sort and Match activities within *Graspable Courses*.

A major goal of this work is to make tools and activities that are useful for the teachers and students who use them. Part of this is making activities that are engaging and helpful to student learning. The other part is making data available to teachers that helps them better address their students' needs. To accomplish this, the research team was comprised not only with learning sciences researchers, but also developmental psychologists, mathematicians, computer scientists, and educators who want to use these tools in their classrooms. These diverse perspectives are key in creating tools that are effective for users and easily implemented into authentic contexts. These perspectives, especially from teachers, will continue to drive the design changes for the next iterations of *Graspable Courses*, including the activities, builder, and teacher reports.

Conclusions

Overall, the major objectives of this work were to *1) develop a theoretical framework to assess mathematical proficiency within GM, 2) explore GM-based measures of mathematical proficiency across K-12 populations, and 3) design GM-based tools that are grounded in theory on mathematical proficiency.* Together, the three sections that addressed these objectives provide a rich perspective on the evolution of research on mathematical proficiency, how this research is applied in practice, and an in-depth example of how one technology-based learning environment has been developed to measure mathematical proficiency.

By grounding measurements in educational theory, there is greater potential to evaluate current research in the field. First, this work suggests there is an added benefit of including formative measures within predictive models. Above and beyond background characteristics and summative measures of knowledge, formative measures of the learning process revealed subtle interactions based on student behaviors and prior knowledge. Second, this work showed the feasibility of using a learning technology, like *Graspable Math*, as a formative assessment tool. The second study compared the EFA factors of student problem solving behavior between an elementary and high school population. While the number of factors differed, the variables in each EFA resembled similar measures, such as engagement in problem solving, strategic flexibility, and speed. This shows some potential for measuring components of mathematical problem solving across different age groups and only through student interactions with *GM*. Lastly, this work described the development process of new *GM*-based activities designed to support and measure student progress within each strand of mathematical proficiency. This work suggests that *GM* has great potential for transforming the algebra learning experience. *GM* can be applied in a multitude of activity types that provide classrooms with options for differentiated instruction of algebraic problem solving that is often limited in traditional instruction.

There are many potential applications for the *GM*-based measures of mathematical proficiency proposed in this work. First, these measures could give students a better representation of their learning process by assessing their progress in more incremental measures. This could give a broader population of students more opportunities to express their learning process. Some students may think differently from the way in which tests are written, disadvantaging them from the start because they conceptualize information differently than the way it is presented. This includes students with special needs or students from different cultures who are learning English as a second language. Developing better measures of math proficiency has the potential to better represent the process of learning as well as create more measures that express both learning and effort. This is one way that learning technologies could make learning accessible and equitable for all types of learners.

Second, these measures could be used by teachers to create data-informed practices that meet the needs of their unique students. The *GM* development team is already in the process of creating teacher dashboards that allow teachers to see the most relevant information about their students' progress within the system. Many learning technologies can measure the number of attempts that students take or the time they take to solve, but *GM* has the ability to show teachers every step that a student took to solve a given problem. GM also has a visualizer feature that can show a classroom-wide visual of the most popular steps to take and how other students deviated from that method. With these detailed measures, teachers can better assess the learning process rather than the learning outcome and inform their instructional practices in the classroom to meet students where they are in the process.

Third, these measures could also be made into a student dashboard where students can assess their own learning at any point. Learning is a complex process that involves many self regulatory behaviors such as planning, metacognitive monitoring, and reflection (Azevedo, 2007). Learning can be enhanced with instructional activities that focus on metacognitive practices, which can be efficiently and automatically facilitated with learning technologies (Aleven & Koedinger, 2002). *GM* has the potential to extend typical metacognitive practices to incorporate all five strands of mathematical proficiency. Together, these applications have major practical implications to increase data-driven practices that provide more individualized classroom experiences for students, as well as more metacognitive engagement in student reflection of their own learning. All of which stem from, and could only exist from, a perceptual-motor-based learning technology like *GM*.

The work presented in this thesis has shown that the potential for *Graspable Math* is immense, not only as a learning tool, but as a teaching and measurement tool as well. *GM* is not only applicable to algebra I classrooms, but any classroom that uses algebra, including physics, chemistry, and engineering courses. *GM* could also be a great tool for elementary students who are seeing expressions for the first time. The immediate feedback and fluid gestures allows users to explore the laws of mathematics in a risk-free environment. In terms of measurement, *GM* is also a great platform for exploring the components of mathematical proficiency. With *GM's* great logging capabilities, researchers can tease apart the five strands and continue to work with teachers to determine how these formative measures could inform teaching practices and benefit students algebraic reasoning. Ultimately the goal of this work is to serve as an example method for other researchers, educators, and designers to move beyond summative measures of assessment and enhance the formative assessment capabilities of learning technologies by grounding measures in theories of learning.

This research is supported by the U.S. Department of Education, through Grant R305A110060 to University of Richmond and Indiana University, Grant R305A180401 to Worcester Polytechnic Institute, Grant 91990018C0032 to Graspable Math for a Small Business and Innovation Research (SBIR) award, and through Grant P200A150306 and P200A180088 for the Graduate Assistance in Area of National Need (GAANN) fellowship. Any opinions, findings, and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S Department of Education.

References

Aleven, V., and Koedinger, K. R. 2000. Limitations of student control: Do students know when they need help?. In Intelligent tutoring systems, 1839, 292-303.

Allsopp, D., Lovin, L. H., & van Ingen, S. (2017). Supporting Mathematical Proficiency: Strategies for New Special Education Teachers. *TEACHING Exceptional Children*, *49*(4), 273-283.

Angeli, C., & Valanides, N. (2009). Epistemological and methodological issues for the conceptualization, development, and assessment of ICT–TPCK: Advances in technological pedagogical content knowledge (TPCK). *Computers & education*, *52*(1), 154-168.

Arroyo, I., Beal, C., Murray, T., Walles, R., & Woolf, B. P. (2004). Web-based intelligent multimedia tutoring for high stakes achievement tests. In *International Conference on Intelligent Tutoring Systems* (pp. 468-477). Springer, Berlin, Heidelberg.

Arroyo, I., Muldner, K., Schultz, S. E., Burleson, W., Wixon, N., & Woolf, B. P. (2016). Addressing Affective States with Empathy and Growth Mindset. In *UMAP (Extended Proceedings)*.

Arroyo, I., & Woolf, B. P. (2005). Inferring learning and attitudes from a Bayesian Network of log file data. In *AIED*(pp. 33-40).

Arroyo, I., Woolf, B. P., Cooper, D. G., Burleson, W., & Muldner, K. (2011). The impact of animated pedagogical agents on girls' and boys' emotions, attitudes, behaviors and learning. In *Advanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on* (pp. 506-510). IEEE.

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, *24*(4), 387-426.

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

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students game the system. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 383-390). ACM.

Baker, R. S., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment-by-moment. *International Journal of Artificial Intelligence in Education*, *21*(1-2), 5-25.

Bergan, J. R., Sladeczek, I. E., Schwarz, R. D., & Smith, A. N. (1991). Effects of a measurement and planning system on kindergartners' cognitive development and educational programming. American Educational Research Journal, 28, 683–714.

Bieg, M., Goetz, T., & Lipnevich, A. A. (2014). What students think they feel differs from what they really feel–academic self-concept moderates the discrepancy between students' trait and state emotional self-reports. *PloS one*, *9*(3), e92563.

Booth, J. L., Barbieri, C., Eyer, F., & Paré-Blagoev, E. J. (2014). Persistent and pernicious errors in algebraic problem solving. *The Journal of Problem Solving*, *7*(1), 3.

Braith, L., Ottmar, E., & Skorinko, J. (2017). *Even Elementary Students Can Explore Algebra!* (Undergraduate Major Qualifying Project No. E-project-042717-103336). Retrieved from Worcester Polytechnic Institute Electronic Projects Collection: https://web.wpi.edu/Pubs/E-project/Available/E-project-042717-103336/

Braith, L, Daigle, M, Manzo, D, & Ottmar, E. (2017). *Even Elementary Students Can Explore Algebra!: Testing the Feasibility of from Here to There!, a Game-Based Perceptual Learning Intervention.* Poster Presented at the American Psychological Society Conference, Boston, MA.

Botelho, A. F., Adjei, S. A., and Heffernan, N. T. 2016. Modeling interactions across skills: A method to construct and compare models predicting the existence of skill relationships. In Proceedings of the 9th International Conference on Educational Data Mining, 292-297.

Buabeng-Andoh, C. (2012). Factors Influencing Teachers' Adoption and Integration of Information and Communication Technology into Teaching: A Review of the Literature. *International Journal of Education and Development using Information and Communication Technology*, *8*(1), 136-155.

Carraher, D. W., Schliemann, A. D., Brizuela, B. M., & Earnest, D. (2006). Arithmetic and algebra in early mathematics education. *Journal for Research in Mathematics education*, 87-115.

Cayton-Hodges, G. A., Feng, G., & Pan, X. (2015). Tablet- based math assessment: what can we learn from math apps?. Educational Technology & Society, 18(2), 3-20.

Conati, C., Gertner A. and VanLehn K. 2002. Using Bayesian networks to manage uncertainty in student modeling. Journal of User Modeling and User-Adapted Interaction, 12, 4, 371-417.

Darling-Hammond, L., Zielezinski, M. B., & Goldman, S. (2014). Using technology to support at-risk students' learning. *Stanford Center for Opportunity Policy in Education. Online https://edpolicy. stanford. edu/publications/pubs/1241*.

D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., ... & Graesser, A. (2008). AutoTutor detects and responds to learners affective and cognitive states. In *Workshop on emotional and cognitive issues at the international conference on intelligent tutoring systems* (pp. 306-308).

Drasgow, F. (Ed.). (2015). Technology and testing: improving educational and psychological measurement. Routledge.

Dweck, C. S. (2002). Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways). In *Improving academic achievement* (pp. 37-60). Academic Press.

Feinberg, S., & Murphy, M. (2000, September). Applying cognitive load theory to the design of web-based instruction. In *Proceedings of IEEE professional communication society international professional communication conference and Proceedings of the 18th annual ACM international conference on Computer documentation: technology & teamwork* (pp. 353-360). IEEE Educational Activities Department.

Ferguson, K., Arroyo, I., Mahadevan, S., Woolf, B., and Barto, A. 2006. Improving intelligent tutoring systems: Using expectation maximization to learn student skill levels. In Intelligent Tutoring Systems, 4053, 453-462.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22*(4), 521-563.

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. Topics in Cognitive Science, 2(2), 265-284.

Goleman, D. (1996). Emotional Intelligence. Why It Can Matter More than IQ. *Learning*, *24*(6), 49-50.

Gotwals, A. W., Philhower, J., Cisterna, D., & Bennett, S. (2015). Using video to examine formative assessment practices as measures of expertise for mathematics and science teachers. *International Journal of Science and Mathematics Education*, *13*(2), 405-423.

Gülbahar, Y. (2007). Technology planning: A roadmap to successful technology integration in schools. *Computers & Education*, *49*(4), 943-956.

Holstein, K., Hong, G., Tegene, M., McLaren, B. M., & Aleven, V.: The classroom as a dashboard: Co-designing wearable cognitive augmentation for K-12 teachers. In: LAK. pp. 79-88. ACM (2018).

Holstein, K., McLaren, B. M., & Aleven, V. (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In *International Conference on Artificial Intelligence in Education* (pp. 154-168). Springer, Cham.

Jiang, M. J., Cooper, J. L., & Alibali, M. W. (2014). Spatial factors influence arithmetic performance: The case of the minus sign. *The Quarterly Journal of Experimental Psychology*, *67*(8), 1626-1642.

Kaput, J. (1998, May). Transforming algebra from an engine of inequity to an engine of mathematical power by "algebrafying" the K-12 curriculum. In *The nature and role of algebra in the K-14 curriculum: Proceedings of a national symposium* (pp. 25-26). Washington, DC: National Research Council, National Academy Press.

Kellman, P. J., Massey, C. M., & Son, J. Y. (2010). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. Topics in Cognitive Science, 2(2).

Kiili, K., Devlin, K., Perttula, T., Tuomi, P., & Lindstedt, A. (2015). Using video games to combine learning and assessment in mathematics education. International Journal of Serious Games, 2(4), 37-55.

Kirkwood, A., & Price, L. (2014). Technology-enhanced learning and teaching in higher education: what is 'enhanced' and how do we know? A critical literature review. *Learning, media and technology*, *39*(1), 6-36.

Kirshner, D. (1989). The visual syntax of algebra. Journal for Research in Mathematics Education, 274-287.

Koedinger, K. R., & MacLaren, B. A. (2002). Developing a pedagogical domain theory of early algebra problem solving. In *Carnegie Mellon University*.

Koedinger, K. R., and Mathan, S. 2004. Distinguishing qualitatively different kinds of learning using log files and learning curves. In ITS 2004 Log Analysis Workshop, 39-46.

Koole, M. L. (2009). A model for framing mobile learning. *Mobile learning: Transforming the delivery of education and training*, *1*(2), 25-47.

Kulik, J. A., & Fletcher, J. D.: Effectiveness of intelligent tutoring systems: a meta- analytic review. RER, 86, 1, pp. 42-78. (2016).

Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 720.

Landy, D., & Goldstone, R. L. (2010). Proximity and precedence in arithmetic. *The Quarterly Journal of Experimental Psychology, 63*(10), 1953-1968.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, *38*(1), 43-52.

McFarland, J., Hussar, B., de Brey, C., Snyder, T., Wang, X., Wilkinson-Flicker, S., ... & Bullock Mann, F. (2017). The Condition of Education 2017. NCES 2017-144. *National Center for Education Statistics*.

Mills, S. C., & Tincher, R. C. (2003). Be the technology: A developmental model for evaluating technology integration. *Journal of Research on Technology in Education*, *35*(3), 382-401.

National Research Council (NRC). (2001a). Adding it up: Helping children learn mathematics: National Academies Press

Nord, C., Roey, S., Perkins, R., Lyons, M., Lemanski, N., Brown, J., & Schuknecht, J. (2011). The Nation's Report Card [TM]: America's High School Graduates. Results of the 2009 NAEP High School Transcript Study. NCES 2011-462. *National Center for Education Statistics*.

Nutt, J. (2010). Professional educators and the evolving role of ICT in schools. *CfBT Education Trust*.

Ostrow, K. S., and Heffernan, N. T. 2015, June. The role of student choice within adaptive tutoring. In International Conference on Artificial Intelligence in Education, 752-755. Springer International Publishing.

Ottmar, E., & Landy, D. (2017). Concreteness fading of algebraic instruction: effects on learning. *Journal of the Learning Sciences*, *26*(1), 51-78.

Ottmar, E., Landy, D., Goldstone, R. L., & Weitnauer, E. (2015). Getting From Here to There!: Testing the Effectiveness of an Interactive Mathematics Intervention Embedding Perceptual Learning. In CogSci.

Ottmar, E., Landy, D., Weitnauer, E., & Goldstone, R. (2015). Graspable mathematics: Using perceptual learning technology to discover algebraic notation. In *Integrating touch-enabled and mobile devices into contemporary mathematics education*(pp. 24-48). IGI Global.

Pardos, Z. A., & Heffernan, N. T. (2011). KT-IDEM: introducing item difficulty to the knowledge tracing model. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 243-254). Springer, Berlin, Heidelberg.

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, *102*(3), 531.

Pekrun, R., Vogl, E., Muis, K. R., & Sinatra, G. M. (2017). Measuring emotions during epistemic activities: the Epistemically-Related Emotion Scales. *Cognition and Emotion*, *31*(6), 1268-1276.

Popham, J. 2009. Assessing Student Affect, Educational Leadership, 66, 8, 85-86.

Roll, I., Baker, R. S. J. d., Aleven, V., and Koedinger, K. R. 2014. On the benefits of seeking (and avoiding) help in online problem-solving environments. The Journal of the Learning Sciences, 23, 537-560.

Romero, C., and Ventura, S. 2010. Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetircs—*Part C: Applications and Reviews*, 40, 6, 601-61.

Schofield, J. W.: Computers and classroom culture. Cambridge University Press (1995).

Sfard, A., & Linchevski, L. (1994). The gains and the pitfalls of reification—the case of algebra. In *Learning mathematics* (pp. 87-124). Springer, Dordrecht.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. Computer games and instruction, 55(2), 503-524.

Speece, D. L., Molloy, D. E., & Case, L. P. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. Learning Disabilities: Research and Practice, 18(3), 147–156.

Stein, M. K., Kaufman, J. H., Sherman, M., & Hillen, A. F. (2011). Algebra: A challenge at the crossroads of policy and practice. *Review of Educational Research*, *81*(4), 453-492.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, *4*(4), 295-312.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197-221.

Weitnauer, E., Landy, D., & Ottmar, E. (2016). Graspable math: Towards dynamic algebra notations that support learners better than paper. In *Future Technologies Conference.*

Wixon, M., Arroyo, I., Muldner, K., Burleson, W., Rai, D., and Woolf, B. 2014. The opportunities and limitations of scaling up sensor-free affect detection. In Educational Data Mining 2014.

Yacef, K.: Intelligent teaching assistant systems. In: ICCE. pp. 136-140. IEEE (2002)