THE IMPACT OF A TECHNOLOGY-ENHANCED MATH PERFORMANCE TASK ON STUDENT COGNITIVE ENGAGEMENT IN MATHEMATICS

by

MEG GUERREIRO

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2017

DISSERTATION APPROVAL PAGE

Student: Meg Guerreiro

Title: The Impact of a Technology-Enhanced Math Performance Task on Student Cognitive Engagement in Mathematics

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Leadership by:

Dr. Kathleen Scalise	Chairperson
Dr. Keith Hollenbeck	Core Member
Dr. Joanna Smith	Core Member
Dr. Joanna Goode	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2017

DISSERTATION ABSTRACT

Meg Guerreiro

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

June 2017

Title: The Impact of a Technology-Enhanced Math Performance Task on Student Cognitive Engagement in Mathematics

Technology may play a critical role in impacting student engagement, specifically within an assessment context. Using a mixed methods approach, I examined the relationship between varying degrees of technology-enhancements applied in a mathematics performance task on the outcome of student cognitive engagement. Using a counterbalanced quasi-experimental design, I evaluated the impact of three performance task platforms on student self-reported cognitive engagement in from a sample of students in grades 6-8 in Oregon, Washington, and North Carolina ($N = 450$). The three performance task platforms (a) included technology-enhanced (technology-based including animation and interactivity), (b) technology-enabled (computer-based without including animation and interactivity), (c) and paper-and-pencil. The measure used for cognitive engagement (CE-S-DSP & SOS) was a hybrid of previously used self-reporting tools and showed preferable reliability for the overall score of cognitive engagement. The data were not able to be explored using a 5-factor confirmatory factory analysis, due to model fit limitations.

Results from the between subjects analysis of variance and did not suggest a relationship between performance task platform (modality type) and student cognitive

engagement. Qualitative interview data indicated that students preferred using technology to take tests and overall showed favorability for the technology-enhanced performance task, specifically the interactivity and animations to help visualize and work through the problem. Yet, despite the positive links to technology-enhancements, there were features of paper-and-pencil tasks that students appreciated such as the ability to navigate between the items and the ability to take notes. Results indicated that just putting tests on computers may not be enough and technological affordance should be purposefully implemented. Findings from this study can help inform future use of platform type, technological enhancements employed, and strategies for technology use within an assessment context.

CURRICULUM VITAE

NAME OF AUTHOR: Meg Guerreiro

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Wilkes University, Wilkes-Barre, PA
Temple University, Philadelphia, PA

DEGREES AWARDED:

Doctor of Philosophy, Educational Leadership, 2017, University of Oregon
Master of Science, Instructional Media, 2010, Wilkes University
Bachelor of Science, Elementary Education, 2008, Temple University
Bachelor of Science, Special Education, 2008, Temple University

AREAS OF SPECIAL INTEREST:

Educational Technology
Measurement and Assessment
Self-Efficacy and Engagement

PROFESSIONAL EXPERIENCE:

Research Scientist, NWEA, 2017-Present

Senior Research Associate, NWEA, 2015-2017

Student Support Services Coordinator, The International School, 2013-2015

Educational Technology Coordinator, The International School, 2012-2013

Elementary Teacher, New Hope-Solebury School District, 2010-2012

Elementary Teacher, The School District of Philadelphia, 2008-2010

GRANTS, AWARDS, AND HONORS:

Roe L. Johns Travel Grant, Assessing Oregon Schools' Preparedness to Implement the Smarter Balanced Assessment, Association for Education Finance and Policy, 2015

Department of Educational Methodology, Policy, and Leadership Travel Grant, Analyzing Opportunity to Learn Common Core and Preparedness of Smarter Balanced Implementation in Oregon, University of Oregon, 2015

College of Education Travel Grant, The Impact of an iPad-Delivered Mathematics Intervention on Number Identification Skills, University of Oregon, 2014

Department of Educational Methodology, Policy, and Leadership Travel Grant, Exploring the Role of Teacher Familiarity with Technology in the Implementation of an iPad Delivered Kindergarten Mathematics Intervention, University of Oregon, 2014

PUBLICATIONS:

Adkins, D. & Guerreiro, M. (2017). (in press). Learning Styles: Considerations for Technology Enhanced Item Design. British Journal of Educational Technology.

Anderson, R., Guerreiro, M., & Smith, J. (2016). Are all biases bad? Collaborative grounded theory in developmental evaluation of education policy. Journal of MultiDisciplinary Evaluation, 12 (27), 44-57.

Shanley, L., Strand Cary, M., Clarke, B., Guerreiro, M., & Thier, M. (2016). Instructors' Technology Experience and iPad Delivered Intervention Implementation: A Mixed Methods Replication Study. Educational Technology Research & Development, 64 (4).

ACKNOWLEDGMENTS

This research was supported in part by NWEA; the findings and conclusions expressed do not necessarily represent the view or opinions of NWEA. NWEA funded the development and implementation of the assessments used in this research. Special thank you to Mike Nesterak, VP of Advanced Research and Development, for his support in the completion of this project.

I would like to thank my mentor and advisor, Kathleen Scalise, for her continued support, expertise, and for challenging me in all the right ways; I have learned so much these past two years and I am forever grateful for her knowledge, insight, and reassurance. Thank you to Jo Smith, Joanna Goode, and Keith Hollenbeck, for their support and encouragement in preparation of this manuscript.

On a more personal note, I would like to thank my partner and best friend, Steph, for her constant and never ending support, patience, many hugs, and for holding the flashlight in the dark tunnel ahead. She has had a profound impact on this journey and, without her support, I would not be where I am today. Thank you to my son, Gray, whose conception and first year of life spanned the writing of this dissertation and whose curiosity and love of life encouraged me to never give up. Thank you to so many friends for continued encouragement and support and for understanding when plans were canceled last minute or when my computer showed up at dinner. Special thanks to my family (near and far) for believing in me and providing encouragement at all the right times, especially my grandparents whose brave move decades ago made it possible for me to pursue my dreams. Thank you to my sister and dad for being my first teachers in

life and to my parents for instilling in me the value of education and the drive to always chase my dreams.

Thank you to all of my former students who strengthened my love of teaching and provided inspiration to keep asking questions and seeking answers. Finally, thank you to my former teachers, colleagues, and mentors who helped me begin this journey; their passion, service, and support provided inspiration to set and reach unthinkable goals.

"And once the storm is over, you won't remember how you made it through, how you managed to survive. You won't even be sure, whether the storm is really over. But one thing is certain. When you come out of the storm, you won't be the same person who walked in. That's what this storm's all about."

Haruki Murakami

For my partner, Steph, and son, Gray,

and to all children in hope that, one day, each will have the equal opportunity to pursue their dreams.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Table Page

Table Page

CHAPTER I

INTRODUCTION

A strong focus over recent years in the U.S. on institutional accountability has sparked a national focus on student assessment in education, including as a result of mandates from the Every Student Succeeds Act (ESSA). Signed by President Obama in 2015, this measure reauthorizes the 50-year-old Elementary and Secondary Education Act (ESEA), the nation's national education law.

While the exact implementation policies of ESEA are somewhat in flux under the new administration, the nationwide conversation in the U.S. regarding educational assessment remains largely centered on two formats: large scale more summative assessments and classroom-based more formative assessments. Both types of assessments are often used to make decisions about students, teachers, schools, or districts (Garrison & Ehringhaus, n.d.; Hidden Curriculum, 2014; Northern Illinois University, n.d.).

Large-scale assessments are often used for accountability purposes, so the preparation and outcomes of these assessments can be a key focus of school districts and teachers. A focus on assessment for learning, or the use of assessment data to drive instruction, has also prompted a need for evaluation of student performance as well as program effectiveness through various types of assessment models. These components may in some cases when used appropriately help to satisfy the necessity for evidence of student learning (Smiley & Anderson, 2011).

As an example of both of these focal points for assessment, the Common Core State Standards (CCSS), a national set of educational academic standards which have been adopted or adapted by many states, call for all students to be college and career ready upon high school graduation. As a result of the evolving educational landscape, a number of the high interest

assessment needs for these programs involve *hard-to-measure* constructs (Scalise, 2012), for which successful paper-and-pencil assessments have sometimes proven inadequate, unwieldy, or too expensive for a sufficient data collection process. In response, and due to numerous technology affordances now available, assessments both for large-scale use and in the classroom have started to include technology-enhanced items and tasks. These may be used to meet CCSS college and career readiness benchmarks such as problem solving and 21st century skills. Both CCSS standardized assessments such as originally developed by the Smarter Balanced Assessments [SBA] and the Partnership for Assessment of Readiness for College and Careers [PARCC] include technology-enhanced tasks, for large-scale assessments and for use in the classroom.

Statement of Problem

The use of assessment for accountability purposes as well as student achievement and instructional needs calls for a more in-depth understanding of student effort expended during assessments, or test events. When students have expended low motivation during assessment, the validity of program evaluation results that utilize student scores can be erroneous (Wise & DeMars, 2005b). Such approaches to program evaluation, therefore, require further information on student engagement as well as ways to attempt to encourage student effort to provide an accurate demonstration of knowledge and understanding. Thus, it is important to understand what students actually know versus the level at which they are willing to perform during a test event (Thelk et al., 2009; Wise & DeMars, 2005b), especially if student data are to be employed in the evaluation and adoption of school programs or for educational decision making, such as graduation.

In the research literature, student effort is often measured by evaluating cognitive engagement and motivation during test events (Miller, Greene, Montalvo, Ravindran, & Nichols, 1996; Smiley & Anderson, 2011; Sundre, 1997; Wise, 2006). This has been measured by either student self-report in the form of surveys or measured by time spent per item, also referred to as response time effort (RTE). Yet, in order to accurately understand and report on student assessment outcomes, students must show motivation and effort during assessment practices (Wise & DeMars, 2005b). This is particularly true when assessment outcomes during a high stakes test are used to measure student academic achievement and program effectiveness (Smiley & Anderson, 2011).

To some degree, student outcomes have been studied through test modality (Buchanan, 2002; Gallagher, Bridgeman, & Cahalan, 2002; Hargreaves, Shorrocks-Taylor, Swinnerton, Tait, & Threlfall, 2004; Lankford, Bell, & Elias, 1994). Yet, there is minimal research on the degree of interactivity and agency in assessment item types. Additionally, modes of measurement to evaluate student cognitive engagement have differed substantially. Without accurate measures of student motivation and engagement during the test event, outcome measures can be unreliable, therefore, making it uncertain whether scores are reflective of students' true ability or whether scores are reflective of an indication of student performance when students are not trying their best (Thelk et al., 2009; Wise & DeMars, 2005b).

In addition to consistent and accurate measurement of student motivation to evaluate program effectiveness, student motivation is a strong influence for the involvement and pursuit of science, technology, engineering, and mathematics (STEM) among youth. Despite the increased nationwide focus on STEM among policy makers, overall student interest has been

reported as only marginally increasing in recent years, such as by one percent between 2010-2014, as reported by survey research (ACT, 2014b).

Better understanding student effort, as measured by cognitive engagement and motivation, may help educators improve student experiences in STEM through curricular programs, extracurricular activities, instructional practices, and/or interventions and assessments. The implementation of well-designed interventions and assessments, particularly utilizing technology-enhancements, can yield high quality data about student effort and interest as well as student performance ability within STEM settings. This additional information may help to broaden understanding of student effort and interest in STEM activities as well as provide an ideal opportunity to contribute to STEM interest investigations and help recognize how to encourage students in other STEM efforts.

Understanding student effort is also imperative to understanding student outcomes of a test event. It is important to distinguish between student "actual proficiency (i.e., what a student knows and can do) [and] demonstrated proficiency (i.e., how well a student performs on a test)" (Wise & DeMars, 2005b, p. 14) as a result of assessment performance. This distinction can help to avoid a threat to test score validity (Thelk et al., 2009; Thelk, Sundre, Horst, & Finney, n.d.; Wise & DeMars, 2005b). Often, assessment scores are used to make significant judgments about academic programs (Kane, 2001; Messick, 1994; Thelk et al., n.d.) without taking into account student engagement and motivation during the data collection process. Without the consideration of student engagement and motivation, interpreting student assessment outcomes can be difficult and erroneous. The lack of consequential outcomes for students (e.g., graduation, grades) as a result of an assessment, as well as the lack of intrinsic motivation for performance during a test event, can have a direct impact on student outcome measures and,

consequently, institutional accountability (Smiley & Anderson, 2011; Thelk et al., 2009; Wise & DeMars, 2005b). Additionally, the absence of sufficient engagement expended by students during an assessment may lead to instructional decision-making that is not supported by high quality data.

Additionally, advancements in technology-enhanced assessments, yet lack of literature on these advancements, call for the need to examine modality differences between paper-and-pencil assessments and computer-based assessments (CBA), including different types of CBA. Modality differences and varying degrees of interactivity, agency, animation, and other components may be key contributing factors to how students experience technology-enhancements, specifically during assessments. Previous research resulted in mixed outcomes regarding the impact of mode on student achievement and engagement, with few research results in mathematics (Bodmann & Robinson, 2004; Buchanan, 2002; Clariana & Wallace, 2002a; Gallagher et al., 2002; Hargreaves et al., 2004; Neuman & Baydoun, 1998). Great variation in the purpose and design of assessments, as well as the perceptions of what represents students' skills and abilities, have also contributed to lack of clarity and breadth (Miller et al., 1996).

Purpose of the Study

The purpose of this study is to help address the need for technology-enhanced assessment research in order to evaluate modality differences as well as student cognitive engagement. A better understanding of how technology-enhanced tasks and assessment modality impact student effort during assessments may help to promote better outcome measures as well as improve the interpretation of outcomes. Thus this study examines an adapted measure of cognitive engagement (CE-S-DSP & SOS) used within assessments, with the context here employing a mathematics assessment developed in three different types of delivery.

The cognitive engagement measure used here expands on the extant Student Opinion Scale (SOS; Sundre, 1999) and the Cognitive Engagement Survey (CES; adapted from Miller, Greene, Montalvo, Ravindran, & Nichols, 1996). It creates a Cognitive Engagement Scale – Short – Deep, Shallow, Persistence combined with the SOS, or "CE-S-DSP & SOS" for the combined instruction. The SOS portion of the combined instrument measures affective aspects of student motivation, including self-reported beliefs of importance (perceived value) and effort (theoretical value). The CES portion measures motivation as evidenced by degree of self-reported cognitive engagement with two scales of processing (deep and shallow) and persistence. The CE-S-DSP & SOS is used here to estimate the construct of cognitive engagement within middle school students (grades 6-8) on three mathematics performance task modes.

Research questions are:

Research Question One (RQ1): Investigating the performance of the CE-S-DSP & SOS within the context of the mathematics performance instrument:

1a. Does the CE-S-DSP & SOS (see Appendices A-B) show variance across the components of cognitive engagement in the context of the mathematics performance instruments (see Appendices C-F), for the sample dataset?

1b. Does the CE-S-DSP & SOS show internal consistency in the context of the mathematics performance instruments for the sample dataset?

Research Question Two (RQ2): Investigating quantitative and qualitative relationships between affective measure outcomes and use of modality types:

2a. What relationship is found between student cognitive engagement and assessment modality type, following the use of the mathematics performance instrument provided in three

different modality conditions? The three different conditions are (1) paper-and-pencil, (2) technology-enabled, which was converted with fidelity to paper-and-pencil but ported on the computer device with the inclusion of scaffolding, and (3) technology-enhanced, with technology enhancements employing more technology affordances and innovations to support interactivity and agency than the technology-enabled fidelity performance instrument or the paper-and-pencil instrument. Disaggregation by sex and race/ethnicity was also considered in this research question.

2b. What relationship is found between home technology use patterns and cognitive engagement?

2c. Using a more in-depth qualitative interview protocol (see Appendix G), can descriptions of student attitudes towards technology-enhanced assessments be developed and associated with various aspects of student performance?

Research Literature

The following research literature section provides a foundation to contextualize the current study. The sections aim to present previous literature, discuss gaps in the research, and provide the connection to the current study. The literature section begins by reviewing the nation's increased focus on STEM, notably motivation, extraneous factors, inclusion of marginalized groups, and shift in aspirations during middle school. The research then intends to define the theory of cognitive engagement and motivation and review each within an assessment context including how they relate to each other and impact program evaluation as well as the theoretical framework for the cognitive engagement instrumentation used in the current study. Previously used survey tools and statistical procedures to measure cognitive engagement are discussed. Finally, the need for authentic assessment, technology-enhanced items, modernized

delivery modes are explored, as evidenced by the CCSS and College and Career Readiness. The impact of computer use on academic and nonacademic outcomes are also discussed as well as research needs within the subject of performance tasks. Additionally, the section reviews the theoretical framework for the development of the performance task. The section concludes by outlining the research gap as well as the literature and gap connection to the current study.

STEM focus. The increased focus on STEM in western developed nations has been notable (U.S. Department of Labor, 2007; Wang, 2013); yet, many students have low STEM aspirations (DeWitt et al., 2013; Elster, 2014; Lyons & Quinn, 2010) and, as a result, choose not to pursue STEM focused courses or careers (Lyons & Quinn, 2010). Additionally, the overall student interest in STEM between 2010 and 2014 has only increased by one percent (ACT, 2014b). Of the 2014 high school graduates, 53% expressed an interest in mathematics and 46% expressed an interest in science (ACT, 2014a). The lack of aspirations towards STEM is particularly prominent in marginalized populations such as females (Archer et al., 2010; Elster, 2014), racial and ethnic minorities (DeWitt et al., 2013; Wang, 2013), and students from low socio-economic backgrounds (Archer et al., 2012; Aschbacher, Li, & Roth, 2010).

Of high school graduates in 2014, interest in mathematics was expressed by 18% of African American students, 25% of American Indian students, 36% of Hispanic/Latino students, and 39% of Pacific Islander students in comparison to 75% of Asian students and 58% of White students (ACT, 2014b). Similarly, interest in science was expressed by 13% of African American students, 21% of American Indian students, 26% of Hispanic/Latino students, and 29% of Pacific Islander students in comparison to 59% of Asian students and 52% of White students (ACT, 2014b). Additionally, Wang (2013) expressed the importance of addressing the

gender bias in STEM by increasing female students' self-efficacy in order to encourage female interest in STEM fields.

The lack of students pursuing STEM fields is problematic and detrimental to societal employment demands (Langdon, Beede, & Doms, 2011) particularly as STEM fields are increasing. The STEM education and workforce challenge is difficult; many students do not pursue STEM because of inadequate preparation in the K-12 system (U.S. Department of Labor, 2007). Of 1,845,787 high school graduates in 2014, 43% were considered college ready in mathematics and 37% were considered college ready in science (ACT, 2014a). These statistics are a startling indication of the lack of student proficiency in STEM fields. Further, there is a greater need for more diverse populations of STEM professionals in today's global market which would require not only academic STEM proficiency but student aspirations and interest in STEM fields. Yet, despite these needs, the influences in trends of career interest are impacted by several factors including student self-efficacy, various extraneous factors, and inequitable marginalization of groups, aspirations, and gender.

Motivation for STEM. Motivation is a strong factor in STEM interest, with a significant and positive link between mathematics attitudes, mathematics self-efficacy, aspirations to pursue STEM, and higher education aspirations (Wang, 2013). Additionally, self-efficacy and confidence present continuing challenges in STEM (DeWitt et al., 2013; Elster, 2014; Lyons & Quinn, 2010; Wang, 2013), specifically student aspirations (Bandura, Barbaranelli, Caprara, & Pastorelli, 2001). Substantial research has linked student aspirations in STEM to student engagement (Elster, 2014; Lyons & Quinn, 2010) and self-efficacy (DeWitt et al., 2013; Elster, 2014; Lyons & Quinn, 2010; Wang, 2013); yet, considerable literature indicates that STEM is failing to engage students (Aschbacher et al., 2010; Lyons & Quinn, 2010).

There are also substantial gender differences in STEM (DeWitt et al., 2013; Elster, 2014; OECD, 2006b) that can shape student STEM interest by age 14. This difference in gender could also be reflective of a disconnect with school experiences (Archer et al., 2010; Aschbacher et al., 2010; Elster, 2014) or a gender dichotomy (Archer et al., 2010) with girls identifying science as too dangerous and boys identifying science as too tame. Some studies reported no difference in gender aspirations towards science (Elster, 2014), rather, focused on gender self-concept (DeWitt et al., 2013; Elster, 2014), gender expectations (Elster, 2014), or gendered family career goals (Aschbacher et al., 2010) as most concerning. Additionally, teachers' science-as-male stereotypes can also greatly influence students' gender differences in motivational beliefs in science (Thomas, 2017).

Although female interest in STEM has increased in recent years, males are more likely than female counterparts to be interested in STEM. This difference is portrayed by statistics showing 55% of males interested in mathematics in comparison to 45% of females as well as 48% of males interest in science in comparison to 38% of females (ACT, 2014b). Also notable, female interest in STEM has a different focus than males counterparts with female interest areas in medical/health and biology and male interest focused on engineering and mathematics (ACT, 2014b). This difference could be indicative of gender expectation (Elster, 2014) or gendered family career goals (Aschbacher et al., 2010) which continues to contribute to the gender gap, particularly in STEM.

Extraneous factors. Low minority self-efficacy and aspirations to participate in STEM-focused programs is problematic due to notably lower percentages of minority students meeting expected benchmarks in mathematics and science as well as overall lower student STEM interest among minority students, in comparison to White and Asian peers (ACT, 2014b). Self-efficacy

is influenced by early mathematics achievement among marginalized populations (DeWitt et al., 2013; Wang, 2013); when early achievement factors are negative, students are less likely to pursue a STEM focus or career. Although exposure to mathematics and science has a direct link to students' STEM interests, the impact accrues most to White students and least to minority students (Wang, 2013). High school racial background largely impacts a students' STEM aspirations with underrepresented populations experiencing the least gain in aspirations to pursue a STEM field (Wang, 2013). With the proper research, training, and curriculum, strategic interventions in education can play an important role in the development of STEM aspirations among minority youth (DeWitt et al., 2013) along with support for families to promote and encourage extracurricular activities (DeWitt et al., 2013). With these supports, STEM participation policies can be created to help engage working-class families in STEM career trajectories (Archer et al., 2012).

Marginalized groups in STEM. Traditionally, science is seen as a subject that often attracts more white, middle class students, and often males (Archer et al., 2010; Lyons & Quinn, 2010). As a result, many young people have a difficult time imagining themselves pursuing science-related degrees despite high aspirations in younger grades (Archer et al., 2010). This is particularly evident for girls (Archer et al., 2010; DeWitt et al., 2013; Elster, 2014; OECD, 2006) and students who are non-White (DeWitt et al., 2013; Wang, 2013). Similarly, fewer than 50% of American Indian and Native-Alaskan students in high school have access to the full range of mathematics and science courses (Shilling, 2015). Additionally, 78% and 83% of the schools serving the lowest percentages of Black and Latino students offer Chemistry and Algebra II courses; yet, only 66% and 74% serving the highest percentages of Black and Latino students offer Chemistry and Algebra II courses (Shilling, 2015).

Archer et al. (2012) links high aspirations to socio-economic factors; while, other research indicates contradictory findings (Aschbacher et al., 2010). Archer et al. (2012) identified a relationship between family habitus with science aspirations; middle class families engender a natural identification of science within their children while working class families perceive science as less familiar. Results from The Programme for International Student Assessment (PISA) suggests that families with higher socio-economic status are more likely to have an interest in science (OECD, 2006b). Additionally, some students, particularly African American girls, pursue science for an altruistic purpose (Aschbacher et al., 2010) while South Asian students are more inclined to view themselves as pursuing a science-related path in comparison with White students (Archer et al., 2012; Aschbacher et al., 2010; DeWitt et al., 2013; Wang, 2013).

Additional statistics show that of the 2014 high school graduates, college readiness mathematics benchmarks were met by 14% of African American students, 20% American Indian students, 29% Hispanic/Latino students, and 30% Pacific Islander students in comparison to 69% of Asian students and 52% of White students. Similarly, in science, college readiness benchmarks were met by 10% of African American students, 17% American Indian students, 21% Hispanic/Latino students, and 22% Pacific Islander students in comparison to 53% Asian students and 46% White students (ACT, 2014a). Although none of the race or ethnicity benchmarks are noteworthy, there is clearly a gap where Asian and White students are outperforming other racial and ethnic students such as students who are Hispanic/Latino, Pacific Islander, or African American.

STEM aspirations between ages 10 to 14. Students studied at the age of 10 often enjoy doing science and have aspirations to pursue a science-related career regardless of gender, race,

or ethnicity (Archer et al., 2012; Archer et al., 2010); however, these aspirations dissipate as students leave elementary school (Archer et al., 2010; DeWitt et al., 2013) and aspirations have been completely formed by age 14 (Archer et al., 2010; Tai, Qi Liu, Maltese, & Fan, 2006). This interest trajectory may also extend beyond science to all STEM related fields. Some research (Archer et al., 2012; Tai et al., 2006) already recognizes many working class children as disadvantaged in science by the age of ten. Research also identifies and connects middle class aspirations as well as working class resistance towards STEM along with demographic factors such as gender (DeWitt et al., 2013; Elster, 2014; OECD, 2006b) and race/ethnicity (DeWitt et al., 2013).

Despite the state of students' science aspirations at age ten, substantial literature indicates that student interest (or lack of interest) in science has been completely formed by age 14 (Archer et al., 2010; Tai et al., 2006) and can extend beyond science to include mathematics, engineering, and technology. This interest has been shaped by life experiences (Aschbacher et al., 2010) such as teachers and curriculum (Aschbacher et al., 2010; Elster, 2014; Lamb, Akmal, & Petrie, 2015; Wang, 2013), parental attitudes (Archer et al., 2012; Archer et al., 2010; Aschbacher et al., 2010; DeWitt et al., 2013; Lamb et al., 2015), peer interests and social influence (Aschbacher et al., 2010; DeWitt et al., 2013; Lamb et al., 2015), and STEM-related extracurricular activities (Archer et al., 2010; Aschbacher et al., 2010; Lamb et al., 2015), expectations or interests (Wang, 2013), or experiences (Aschbacher et al., 2010; DeWitt et al., 2013; Lyons & Quinn, 2010; Wang, 2013) all of which impact students' interest to pursue STEM courses (Elster, 2014). This divergence of initial interest can also be formed by subject difficulty (Aschbacher et al., 2010), mathematics achievement (Wang, 2013), or the impact of personal and professional ideas swaying a student from taking more difficult STEM courses (Aschbacher et

al., 2010). Further research by Wang (2013) also attributed general mathematics and science exposure as a strong predictor of students' future STEM entrance.

Additionally, student demographic factors such as gender (DeWitt et al., 2013; Elster, 2014; OECD, 2006b), race, and ethnicity (DeWitt et al., 2013) have also shown to substantially shape student STEM interest by age 14. These differences in aspirations could also be reflective of a disconnect with school experiences (Archer et al., 2010; Aschbacher et al., 2010; Elster, 2014) or a gender dichotomy (Archer et al., 2010).

Cognitive engagement and motivational theory.

Cognitive engagement . Newmann, Wehlage, & Lamborn (1992) define cognitive engagement in academic work as a "student's psychological investment in and effort directed toward learning, understanding, or mastering the knowledge, skills, or crafts that academic work is intended to promote" (p. 12). Marks (2000) supplements this definition by including attention and interest and the application of "both affective and behavioral participation in the learning experience" (p. 155). Although there are countless definitions of engagement, the literature includes an overview of measurement tools that have clearly, theoretically, and operationally, defined constructs of cognitive engagement. The construct of engagement includes numerous components that aim to measure overall student engagement. The tools used to measure cognitive engagement include both self-report measures (e.g. surveys) as well as time spent per item (e.g. RTE). The self-report tools are primarily the measures requiring a specific definition of cognitive engagement.

Although there are many components to consider when measuring a students' cognitive engagement and effort towards learning and assessment, the literature helps define and make connections between the numerous factors as well as to the overall construct of cognitive

engagement. Maehr and Meyer (1997) discuss components measured in engagement such as direction, intensity, persistence, quality, and outcome; yet, many of these components are not sufficient when explored deeper than face value. For example, intensity (number of tasks completed) may be associated with physiological factors (e.g., illness, fatigue, substance abuse) and may not provide a clear estimate; yet, direction (on or off task behaviors), persistence (time engaged or items attempted), and quality (type of investment) have been noted as primary facts of motivation (Maehr & Meyer, 1997). Furthermore, research has shifted standard measures of schooling outcomes to include critical and creative thinking and other outcomes of lifelong learners (Maehr & Meyer, 1997) which have historically not been included in measures of cognitive engagement.

Other literature (Miller et al., 1996) includes constructs such as learning goals, future consequences, performance goals, pleasing the teacher/family, and perceived ability. Additional adapted forms of a second cognitive engagement measure by Greene and Miller (1996) explore the constructs of self-regulation, deep strategy, shallow processing, and persistence. Work by Sundre (1997) explores importance (perceived value) and effort (theoretical value). These various constructs share minimal overlap; yet, all have significantly predicted cognitive engagement when included in self-report measures.

Once adequately measured, cognitive engagement has been cited by research as having a strong, positive relationship with student achievement (Finn, 1989; Miller et al., 1996; Saeed & Zyngier, 2012; Smiley & Anderson, 2011; Sundre, 1999; Thelk et al., n.d.; Walker, Greene, & Mansell, 2006). In an academic setting, students who are cognitively engaged may demonstrate many favorable actions such as positive behavior (Saeed & Zyngier, 2012), consistent attendance, assignment completion (Saeed & Zyngier, 2012), and academic proficiency (Saeed

& Zyngier, 2012). Students who are engaged are more likely to learn, find the academic learning experience rewarding, graduate, and pursue higher education (Marks, 2000). In contrast, not as favorable behaviors such as rote memorization may be indicative of shallow engagement (Smiley & Anderson, 2011). Students who are disengaged may display behaviors of a highly engaged student such as consistent attendance, assignment completion, display of positive behaviors (Saeed & Zyngier, 2012; Smiley & Anderson, 2011), yet still lack intrinsic value, utility, attainment, and perceived costs (Wise & DeMars, 2005b). Disengagement in an academic setting can have adverse effects on achievement, behavior, attendance, and/or graduation rate (Finn, 1989; Newmann et al., 1992).

Cognitive engagement in assessment. The passing of the ESSA and national adoption of the CCSS has prompted an increase in institutional accountability through assessment. This shift in accountability reform produced the need to evaluate student cognitive engagement within the scope of educational assessment. Although school engagement encompasses a myriad of educational areas, the display of disengagement within educational assessment appears vastly different in the literature. There is minimal consensus about the definition and measurement of cognitive engagement which elucidates themes minimally including academic or cognitive components (Appleton, Christenson, & Furlong, 2008). Within educational assessment, cognitive engagement often falls under the theme of student engagement within academic work (Smiley & Anderson, 2011). In this context, students who demonstrate cognitive engagement may arrive to school prepared, read carefully, and formulate thoughtful answers to master learning with the highest academic results (Newmann et al., 1992; Saeed & Zyngier, 2012); while, disengaged students may arrive the school unprepared and provide vague or unrelated responses in academic discussions (Smiley & Anderson, 2011).

Many researchers have pursued measurement of cognitive engagement in assessment measures (Smiley & Anderson, 2011; Sundre, 1997; Wise & DeMars, 2005b). In educational assessment, Wise & DeMars (2005b) discuss engagement and motivation as "giving one's best effect to the test, with the goal being to accurately represent what one knows and can do in the content area covered by the test" (p. 2). Furthermore, Wise & DeMars (2005b) outline the definition of test taking effort as student "engagement and expenditure of energy toward the goal of attaining the highest possible score on the test" (p.2).

Motivation as evidenced through cognitive engagement. Motivation and cognitive engagement are often mistakenly considered synonymous terms. In fact, both intend to measure effort; yet, motivation measures tend to include effort as a subscale (Smiley & Anderson, 2011). According to Smiley and Anderson (2011), "engagement implies more than motivation, although motivation is necessary for cognitive engagement" (p. 19). Motivation often reflects direction, intensity, and quality (Maehr & Meyer, 1997); while engagement reflects involvement in a task or activity (Reeve, Jang, Carrell, Jeon, & Barch, 2004). Motivation is necessary but not sufficient in order for someone to be engaged (Appleton et al., 2008; Smiley & Anderson, 2011). As a result, engagement is a construct worth an independent investigation (Appleton et al., 2008).

Cognitive engagement can shift and change across different contexts (Marks, 2000; Smiley & Anderson, 2011). This shift could be indicative of a number of factors such as subject ability, self-efficacy, or curriculum. Walker, Greene, and Mansell (2006) highlight the impact of academic value on the prediction of cognitive engagement above and beyond connection with self-efficacy and motivation; therefore, understanding cognitive engagement provides more meaning beyond what self-efficacy and motivation can explain.

The multidimensionality of student engagement has led to an examination of a combination of indicators. Most of the literature focuses on an observable indicator (e.g., student behavior, academic proficiency); while, the less overt indicators (e.g., cognitive and psychological engagement) helps to provide additional information on a students' level of engagement (Appleton et al., 2008). Further, the less overt indications include numerous items that attempt to measure factors related to the construct of student cognitive engagement. Despite countless factors used in measurement, few components that are used to measure cognitive engagement make regular appearances in the literature.

Of the factors discussed, shallow and meaningful processing are one of the most common subsets of items (Miller et al., 1996; Smiley & Anderson, 2011; Walker et al., 2006). The type of processing refers to type of understanding demonstrated by the student within an academic context. When a student demonstrates meaningful processing (also referred to as deep processing or deep strategy) it suggests a student can make strong connections between new and prior knowledge (Kardash & Amlund, 1991; Smiley & Anderson, 2011; Walker et al., 2006) including self-regulatory skills (Miller et al., 1996). On the contrary, shallow processing (also referred to as shallow strategy) involves the demonstration of rote memorization and superficial engagement with material (Smiley & Anderson, 2011; Walker et al., 2006). When measuring cognitive engagement, the inclusion of meaningful processing can lead to enhanced performance on achievement measures (Kardash & Amlund, 1991; Miller et al., 1996) and, therefore, subsequent increases in cognitive engagement.

Additional variables include self-reported effort and perceived importance of the test. Research argues that future consequences are some of the best predictors of achievement (Sundre, 1999), with future consequences being a significant predictor of self-regulation and

deep processing (Miller et al., 1996). Furthermore, Appleton et al. (2008) investigated the construct of engagement through the cyclical interaction with contextual variables (e.g., structure, support, involvement) towards the outcome (e.g., academic, social, emotional) between level of engagement and quantity of support. Miller et al. (1996) also evaluated social support (pleasing the teacher, pleasing the family), academic predictors (learning goal), as well as self-efficacy variables such as perceived ability.

Motivation in STEM. Motivation is a strong factor in student STEM interest, as discussed earlier, with a significant and positive link between mathematics attitudes, mathematics self-efficacy, aspirations to pursue STEM, and higher education aspirations (Wang, 2013). Additionally, self-efficacy and confidence present continuing challenges in STEM (DeWitt et al., 2013; Elster, 2014; Lyons & Quinn, 2010; Wang, 2013), specifically student aspirations (Bandura et al., 2001). Substantial research has linked student aspirations in STEM to student engagement (Elster, 2014; Lyons & Quinn, 2010) and self-efficacy (DeWitt et al., 2013; Elster, 2014; Lyons & Quinn, 2010; Wang, 2013). Yet, considerable literature indicates that STEM is failing to engage students (Aschbacher et al., 2010; Lyons & Quinn, 2010).

Few studies include mathematics performance outcomes. Notably, Miller et al. (1996) refers to mathematics as a complex area of focus and Wang (2013) refers to mathematics as a subject worthy of additional investigation. Historically speaking, mathematics has been a prominent factor in the achievement gap between girls and boys, with more boys pursuing and succeeding in mathematics courses and fields in comparison to female peers (Miller et al., 1996). Additionally, girls, in comparison to boys, tend to have higher mathematics aspirations during earlier schooling years; yet, those aspirations dissipate over time (Archer et al., 2010; DeWitt et al., 2013; Elster, 2014). Moreover, a mathematics achievement gap also exists in the literature

between minority groups, with Asian and White students outperforming other racial and ethnic groups (ACT, 2014a).

Expectancy-value model. The expectancy-value model of achievement motivation, originally developed by Pintrich (1989) but later expanded (Eccles et al., 1983), includes a foundation related to value through the measurement of student effort and reported value (or importance) during a test event or instructional session. Effort, as defined by Wolf, Smith, and Birnbaum (1995) includes the amount of mental taxation the student is willing to exert when responding to items. Additionally, the measurement of effort also corresponds to the theoretical feature of value (Thelk et al., n.d.). In the expectancy-value model, individual perceptions and task-specifics directly influence expectancies, values, and achievement choices, effort, performance, and persistence (Wigfield & Eccles, 2000); student effort is based on student perceived success and student perceived value of scoring well on the test (Eccles et al., 1983; Thelk et al., n.d.; Wigfield & Eccles, 2000; Wise & DeMars, 2005b). The scale is intended to assess students' perceived value of the task. When the scale measures that the perceived value is high, the student is likely to be more engaged (Thelk et al., n.d.). Additionally, belonging and value are known to help predict cognitive engagement, both meaningful and shallow (Walker et al., 2006), above and beyond cognitive engagement and self-efficacy (Walker et al., 2006). Many motivation and engagement measurement tools (i.e. SOS) have reflected changes of scores across various testing consequences (Thelk et al., n.d.).

Cognitive engagement survey tools. There are many self-reporting tools that aim to measure student engagement and motivation. The commonly used tools intend to measure the cognitive engagement construct within different environments: some measure engagement within instructional tasks while others aim to evaluate engagement within test events. One of the

more commonly known tools that measures engagement within an assessment context is the Student Opinion Scale (SOS) developed by Donna Sundre (1997). The SOS was an extension of the 5-point Likert scale used by Wolf, Smith, and Birnbaum (1995) but also incorporated the expectancy-value model (Pintrich, 1989) exploring the factors of importance and effort to measure the construct of cognitive engagement. The SOS is a student self-report 10-item Likert tool mainly used in low stakes testing situations as an efficient means for estimating motivation and has been supported for over a decade of use (Thelk et al., n.d.). The SOS is comprised of two subscales, importance and effort; the scale of importance intends to assess perceived value of the tasks while effort corresponds to the theoretical value (Thelk et al., n.d.). The SOS allows for user reporting on separate subscales as a measure of examinee motivation (Rios, Liu, & Bridgeman, 2014).

Another commonly used instructional engagement tool is the Cognitive Engagement Survey (CES) developed by Greene and Miller (1996) as a scale of cognitive engagement. The scale includes 54 Likert-type items to "examine the links from learning goals, perceived ability, and performance goals to cognitive engagement and then from cognitive engagement to achievement [in order to] help synthesize the interpretation of relationships" (Greene & Miller, 1996). The measure spans across five components aimed to measure different constructs of cognitive engagement including: (1) learning goal orientation, (2) performance goal orientation, (3) perceived ability, (4) meaningful cognitive engagement, (5) shallow cognitive engagement (Greene & Miller, 1996).

Greene and Miller's (1996) scale was later adapted by Smiley and Anderson (2011) and included an abridged scale to measure cognitive engagement within the context of educational assessment. This shortened tool became known as the Cognitive Engagement – Short Form (CE-

S; Smiley & Anderson, 2011). The CE-S was adapted to incorporate language specific to a large-scale assessment context (Smiley & Anderson, 2011) and included deep and shallow processing as measures of cognitive engagement across five items (in total). The 4-point Likert scale that was used in the CE-S was anchored with strongly disagree and strongly agree.

Self-report measures of cognitive engagement such as SOS (Sundre, 1997), the CES (Greene & Miller, 1996), and CE-S (Smiley & Anderson, 2011) are useful tools because they require minimal resources for proper administration; however, as with any tool, the use includes limitations (Rios et al., 2014). For one, examinees may exaggerate estimations when minimal effort was expended (Rios et al., 2014) or examinees may report reduced effort when great effort was expended (Wise & DeMars, 2005b); therefore, making accurate estimates difficult. The exaggeration of estimations as well as the underrepresentation of estimations is a common concern among self-report measures. Additionally, motivation within a test event may fluctuate across items (Wise & Kong, 2005) which would be difficult to measure with an summative self-reporting measure. This could be especially true with interactive items, adaptive items, or items that vary in subject matter or type.

Response time effort measurement. Due to the fact that, self-report measures of engagement are not always useful tools, researchers have attempted more objective measurements for examining engagement within an assessment context. Assessments assume solution behavior by examinees, that is, that an examinee reads and considers an item before responding. Therefore, rapid responses on assessment items can indicate a lack of motivation, specifically in a low stakes assessment context (Wise, Ma, Kingsbury, & Houser, 2010). Deborah Schnipke (1997) suggests that item response time could be useful in spotting engagement at the end of speeded test event. Response time effort (RTE; Wise & Kong, 2005) is

another way to measure student motivation similar to the measurement of effort scores in the SOS (Sundre, 1997; Thelk et al., n.d.). RTE is based on a development by Wise and Kong (2005) which attempts to measure examinee effort during assessments based on behavior rather than self-reporting. Evaluating RTE on individual items, also known as rapid-guessing behavior (Wise et al., 2010), can help to parse guessing (through rapid response) and solution behavior. Wise and colleagues (2010) define rapid-guessing behavior as arriving at a response before being able to read and/or consider an item; while, solution behavior would be considered everything else. The presence of many instances of rapid-guessing behavior within an assessment could affect a students' RTE for the measurement.

The process of RTE identification involves the setting of a threshold (zero to one) to indicate guessing behavior, or low effort, that is being expended (Rios et al., 2014) in comparison to solution behavior. This threshold is set in an attempt to determine non effortful responses (rapid-guessing) and effortful responses (solution behavior) (Wise et al., 2010). Many thresholds have been explored (Rios et al., 2014) such as utilization of a common criterion (Wise, Kingsbury, Thomason, & Kong, 2004), the number of characters in an item (Wise & Kong, 2005), response time frequency distributions (Wise, 2006), statistical estimation using mixture modeling response time and response accuracy (Lee & Jia, 2014), normative threshold percentages (Wise & Ma, 2012), test characteristics (Silm, Must, & Täht, 2013), as well as modeling approaches such as the effort-moderated IRT model (Wise & DeMars, 2005a). Setting threshold values allows for further examination of response time, given the predetermined threshold value (Rios et al., 2014). Literature indicates an acceptable threshold value of 0.90 (Swerdzewski, Harmes, & Finney, 2011; Wise & Kong, 2005) suggesting 90% or more of items should display motivation (Rios et al., 2014). Therefore, it is often the case that a RTE threshold

is set at a 10% response time of the average RTE (e.g. if the average student completes an item in 40 seconds, the threshold for rapid-guessing behavior might be set at 4 seconds).

RTE is particularly useful when measuring effort during low stakes assessments. Typically, during low stakes assessment, student consequences are low; during which, students may choose to not respond to items (Wise & DeMars, 2005a; Wise & Kong, 2005). Additionally, the measurement of RTE could also indicate if students are participating in rapid-guessing behavior; defined as behavior that exhibits rapid responses to items or solutions (Silm et al., 2013; Wise & Kong, 2005). Silm et al. (2013) indicates that there was a lower average time spent on incorrect responses in comparison to correct responses.

RTE has also been evaluated in different testing situations. First, students in older grades exhibit greater rapid-guessing behavior during low stakes test events than students in lower grades (Ma, Wise, Thum, & Kingsbury, 2011; Wise et al., 2010). This could be due to many factors such as item difficulty or motivation. Further, low stakes mathematic items solicit less rapid-guessing in comparison to reading items (Wise et al., 2010).

The RTE model is reliable and theoretically (as well as empirically) related to the SOS (Sundre, 1997) effort scores (Thelk et al., n.d.). Additionally, effort has a positive correlation with achievement, as demonstrated through RTE (Thelk et al., n.d.). The measurement of RTE within a test event can help identify students whose scores may not be indicative of actual proficiency, rather, suggesting a lack of engagement or effort. In doing this, RTE measurement can help to identify spurious student scores within a sample; therefore, identifying when scores fail to demonstrate proficiency (Silm et al., 2013; Wise & Kong, 2005). Omitting students with low RTE scores may provide a better estimate of proficiency within a sample (Silm et al., 2013;

Wise & Kong, 2005) as well as provide a more concrete understanding of student motivation and effort (Wise & Kong, 2005) within a test event.

Comparison of cognitive engagement tools and RTE. Self-report measures of cognitive engagement and RTE have both successfully identified low examinee effort within a low stakes test event; however, minimal literature have compared the two approaches to classifying student engagement within both large samples (Swerdzewski et al., 2011; Wise & Kong, 2005) and small samples (Rios et al., 2014). Swerdzewski et al. (2011) reported strong consistency between the measures of self-reporting and RTE, claiming minimal differences (if any) between the scores. Swerdzewski et al. (2011) suggests researchers may find utility using either method, with self-report removing more suspect data than RTE due to the conservative thresholds of RTE and potential bias in self-reporting. Furthermore, Wise and Kong (2005) found differing results when comparing self-report measures to RTE, indicating the measurement of different constructs or influence by examinee self-reporting. Results also proposed that self-reporting may provide a larger proportion of unmotivated examinees, in comparison to RTE (Wise & Kong, 2005); therefore, supporting the aforementioned claim that self-reporting could lead to a more liberal identification of suspect data (Swerdzewski et al., 2011).

Rios et al. (2014) compared the results of self-reported measures of cognitive engagement and RTE within a small sample size and found a slightly stronger relationship between test performance and RTE than test performance and self-reported measures. This difference led to a slight increase in students being filtered because of low examinee effort when using RTE as a measure (Rios et al., 2014). Researchers suggest that, although both measures filter low-effort examinees, self-report measures and RTE may be "measuring different aspects of examinee effort profiles" (Rios et al., 2014, p. 73).

Rios et al. (2014) suggest that self-reported measures may threaten validity due to the exaggeration or underrepresentation of self-reported effort. The exaggeration or reduction of reported error could be due to a lack of ability (Antin & Shaw, 2012); yet, may introduce threats to validity within a self-report measure. On the contrary, RTE removes this validity threat by using a theoretical threshold value across examinees. However, regardless of the slight differences, Rios et al. (2014) suggest both measures could be an effective method for filtering invalid data.

Cognitive engagement theoretical framework for instrumentation. The subsequent sections briefly describe the reasoning for the development of the cognitive engagement measure and scale. The subscales were developed and used as theoretical subscales but may be treated to yield a single scale or subscales depending on empirical evidence following data collection.

There are have been various methods developed to measure student cognitive engagement in a K-12 educational assessment context (Smiley & Anderson, 2011; Sundre, 1997; Wise & DeMars, 2005b). These methods include observational instruments, self-report measures (Greene & Miller, 1996; Miller, Behrens, Greene, & Newman, 1993; Smiley & Anderson, 2011; Sundre, 1997, 1999, Thelk et al., 2009, n.d.), as well as measures of response time, also known as response time effort (RTE) (Wise, 2006; Wise & DeMars, 2005a, 2005b, Wise et al., 2004, 2010; Wise & Kong, 2005; Wise & Ma, 2012).

Despite the development of numerous measures of cognitive engagement, definitions of how to measure cognitive engagement have minimal consensus among researchers (Appleton et al., 2008). As a result, there are various components that are included as factors within these measures that are believed to contribute to a students' overall measure of cognitive engagement. Maehr and Meyer (1997) discuss components such as direction, intensity, persistence, quality,

and outcome; yet, Maehr and Meyer (1997) suggest that many of these components are not sufficient when explored deeper than face value. For example, intensity (number of tasks completed) may be associated with physiological factors (e.g., illness, fatigue, substance abuse) and may not provide a clear estimate; yet, direction (on or off task behaviors), persistence (time engaged or items attempted), and quality (type of investment) have been noted as primary facts of motivation (Maehr & Meyer, 1997).

Additionally, research has shifted standard measures of schooling outcomes to include critical and creative thinking and other outcomes of lifelong learners (Maehr & Meyer, 1997) which have historically been excluded in measures of cognitive engagement. Other literature (Miller et al., 1996) includes constructs of cognitive engagement such as learning goals, future consequences, performance goals, pleasing the teacher/family, and perceived ability; yet, some of these factors may not apply directly to measurement of cognitive engagement within a K-12 assessment context. Additional adapted forms of a second cognitive engagement measure by Greene and Miller (1996) explore the constructs of self-regulation, deep strategy, shallow processing, and persistence. Work by Sundre (1997) explores importance (perceived value) and effort (theoretical value). All of the various aforementioned constructs share minimal overlap; yet, each has significantly predicted cognitive engagement when included in self-report measures.

For this study, numerous factors of measurement of cognitive engagement were taken into consideration, specifically, for measurement within the context of a low-stakes assessment in grades 6-8. As a result, previously developed instruments were adapted to include specific components of cognitive engagement including: (a) deep strategy, (b) shallow processing, (c) persistence, (d) importance, and (e) effort. The current study used an adapted form of the CES

(Miller et al., 1996) to measure cognitive engagement by including deep strategy, shallow processing, and persistence while also including an un-adapted version of the SOS (Sundre, 1997, 1999) measuring importance and effort, see Figure 1.1. These five components were selected in order to create a theoretically holistic measurement of cognitive engagement within a low stakes assessment context in grades 6-8.

Additional constructs omitted for the adaptation of instruments in this study included future consequences, learning goals, pleasing the teacher, pleasing the family, and perceived ability (Miller et al., 1996). Constructs such as pleasing the teacher, pleasing the family, and learning goals are primarily measured in instructional contexts while future consequences, learning goals, and perceived ability may be similar constructs to others already included; future consequences is a predictor of deep regulation (Miller et al., 1996) and perceived ability may be more related with a self-efficacy construct.

In addition to self-report measures, the inclusion of RTE (measurement of engagement by time spent per item) as a measure may also provide additional information on cognitive engagement; however, there were many factors that led to developing an adapted self-report measure of cognitive engagement rather than including RTE. The RTE model is reliable and theoretically (as well as empirically) related to the SOS (Sundre, 1997) effort scores (Thelk et al., n.d.). Additionally, literature shows minimal differences between RTE and self-reporting (Swerdzewski et al., 2011), suggesting self-reporting can be more rigorous (Swerdzewski et al., 2011; Wise & Kong, 2005) and may be more conservative at flagging students who are disengaged. Lastly, measurement of RTE is a strong indicator of rapid guessing behavior; defined as behavior that exhibits rapid responses to items or solutions (Silm et al., 2013; Wise & Kong, 2005). The subject of mathematics has less rapid guessing behavior in comparison to

reading items (Wise et al., 2010), which aligns closely with the mathematics performance tasks being utilized in this study. Despite the omission of RTE for this study, further exploration of RTE in similar contexts may contribute to the overall measurement of cognitive engagement due to older grades exhibiting great rapid-guessing behavior during low stakes assessments than students in lower grades (Ma et al., 2011; Wise et al., 2010), particularly when collected and analyzed in combination with self-report measures.

Deep strategy measures, also known as self-regulatory skills, measure mastery of academic work (Smiley & Anderson, 2011) while shallow processing measures rote memorization and basic understanding. Both of these constructs were included as measures of cognitive engagement for this study; previous research indicates that when instructional tasks are approached with the goal of increasing understanding or skills, greater self-regulatory and deep cognitive strategies are utilized (Miller et al., 1993) which are related to academic achievement (Miller et al., 1996). Deep and shallow processing are measured across seven Likert items (four for deep strategy and three for shallow processing). Persistence, originally included in Miller et al.'s (1996) cognitive engagement measure but omitted when Smiley and Anderson (2011) adapted the measure, was included in the cognitive engagement measure for this study as an additional factor used towards a holistic measure of cognitive engagement. Persistence is measured across four Likert items. In sum, the CE-S-DSP (adapted from Miller et al., 1996 and Smiley & Anderson, 2011) survey measured three factors across a total of eleven items.

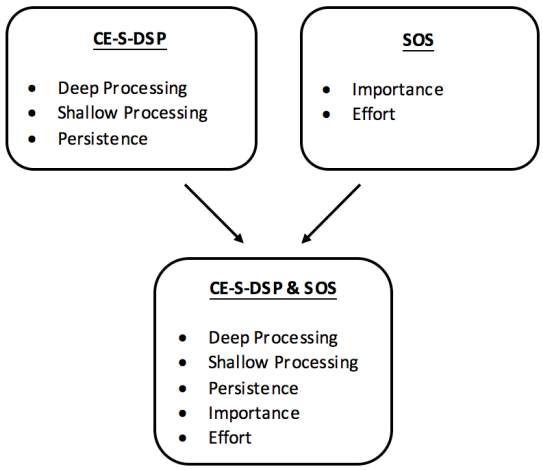


Figure 1.1. The combination of the CE-S-DSP & SOS.

Cognitive engagement survey, CE-S-DSP. The previously developed survey by Miller et al. (1996) and Smiley & Anderson (2011) showed acceptable to good internal reliability coefficients between both the original subscales (Miller et al., 1996) as well as the adapted subscales (Smiley & Anderson, 2011).

Measurement characteristics. Miller et al. (1996) conducted reliability analyses within two samples including: (a) $N = 297$ high school student volunteers ($N = 144$ males, $N = 144$ females, $N = 9$ no report of gender) from a large, middle class suburban school in the mid-south and (b) $N = 269$ students from the same high school as study 1 ($N = 117$ males, $N = 150$ females, $N = 2$ no report of gender). Missing data were treated with listwise deletion (Miller et al., 1996).

Construct validation was obtained through a factor analysis of the constructs (Miller et al., 1996) using varimax rotation; however, due to intercorrelation, a factor analysis using oblique rotation was examined (Miller et al., 1996). Seven factors emerged that had Eigen values greater than one; however, the seventh value was dropped from subsequent analyses due to lack of conceptual value (Miller et al., 1996). A re-examination was further tested by Smiley

and Anderson (2011) by exploration of factor structure using a confirmatory factor analysis (CFA) to determine if the factor structure of the original scale applied to the adapted measure. Smiley and Anderson (2011) evaluated deep engagement and shallow engagement as a new measure of cognitive engagement, specifically within a 45-minute assessment session instead of a full course of study (Miller et al., 1996). The CFA helped to determine the dimensionality of the scale as a two-factor model and concluded that shortening the scale did not affect the factor structure (Smiley & Anderson, 2011).

Reliability coefficients for the original CES (Miller et al., 1996) showed acceptable internal reliability across two studies for each subscale (ranging from .63 to .92). Specifically, the target subscales used within the study had acceptable reliability coefficients for deep strategy ($\alpha = .63/.69$), shallow processing strategy ($\alpha = .65/.73$), and persistence ($\alpha = .75/.81$) (Miller et al., 1996). Additionally, analyses reported $\alpha = .90$ for the longer version of the deep processing subscale, and $\alpha = .81$ for the longer version of the shallow engagement subscale (Greene & Miller, 1996; Smiley & Anderson, 2011). Furthermore, the target subscales (deep processing, shallow processing, persistence) showed no significant differences between male and female students (Miller et al., 1996). Moreover, correlations among the subscales with achievement showed moderate correlation with persistence (r = 0.36) and modest correlation with deep strategy (r = 0.26) (Miller et al., 1996). Overall correlations between the subscales and achievement ranged from 0.22 to 0.40 (Miller et al., 1996).

Smiley and Anderson (2011) conducted reliability analyses for the CE-S, which was a shortened and adapted form of the original CES (Miller et al., 1996); five items were adapted and reworded specifically for a large-scale assessment context. The sample included students from a mid-Atlantic university (N = 243) who participated in a university-wide assessment activity

(Smiley & Anderson, 2011). The factor analysis for the CE-S indicated a two-factor model with appropriate values for the fit indices as well as the standardized polychoric residuals (Smiley & Anderson, 2011). The internal reliability measures were not as high as anticipated due to the small number of items within each factor: deep processing subscale $\alpha = .56$ (three-items) and shallow processing subscale $\alpha = .71$ (two-items) (Smiley & Anderson, 2011). Parameter estimates were measured to determine variance in items accounted for by the latent factor; standardized coefficients ranged from .56 to .91 and were all significant ($p < .05$) (Smiley & Anderson, 2011). Further, R^2 values ranged from .31 to .83, indicating the percentage of variance explained by each item, 31% and 85% accordingly (Smiley & Anderson, 2011).

Student Opinion Survey (SOS). The SOS (Sundre, 1999) was used to measure importance (perceived value) and effort (theoretical value), see Appendix B. The SOS scale (Sundre, 1999) is a revised version of the Wolf and Smith (1993) instrument and was created in order to increase the items from eight to ten and improve the measurement of the two prominent factors: importance and effort (Sundre, 1999).

Measurement characteristics. Thelk et al., (2009) references reliability analyses within a variety of samples including: (a) General Education, Mid-Atlantic, 4-Year, Public Liberal Arts University with N = 3,111 first-year students for study one and N = 3,343 first-year students for study two, N = 1,965 sophomores for study one and N = 2,210 sophomores for study two; (b) General Education; Mid-Western; 4-Year; Public Liberal Arts University with N = 1,002 seniors; and (c) Exit Exams; Mid-Atlantic, 2-Year; Public Community College with N = 332 graduating students (Thelk et al., 2009).

Construct validation was obtained through an iterative process involving the collection of evidence to support or refine the theory or measure under study (Thelk et al., 2009). To begin,

the SOS (Sundre, 1999) was designed with carefully defined constructs (e.g., a theoretical standpoint driven by a theory of motivation) followed by an operational definition of the construct covering domains of all possible items (Thelk et al., 2009). Then, the structure of item responses were examined for covariance, as predicted by theoretical basis for the scale (Thelk et al., 2009). Lastly, external validation was obtained through hypothesizing that scores from the SOS (Sundre, 1999) would relate to specific constructs (e.g., test performance) which would differ between specific groups of students (e.g., students among different testing contexts) (Thelk et al., 2009).

Reliability coefficients for the SOS (Sundre, 1999) showed high internal reliability for each subscale: importance ($\alpha = .80\ to\ .89$) and effort ($\alpha = .83\ to\ .87$). Further internal reliability analyses indicated slightly lower results: effort ($\alpha = .74$) and importance ($\alpha = .77$) (Smiley & Anderson, 2011). Additionally, a factor analysis yielded a two-factor model as a better representation of the data indicating that effort and importance are distinct; however, similar wording between items may provide a poor representation of the data (Thelk et al., 2009).

Furthermore, the subscales (importance and effort) showed no significant differences between male and female students; therefore, providing construct validity of the SOS (Sundre, 1999), higher interpretation confidence across gender, and aggregation of gender responses (Thelk et al., 2009). Moreover, correlations among the subscales with RTE yielded a positive correlation ($r = 0.54$) and correlations among the subscales with achievement showed moderate correlation with effort ($r = 0.30$) (Thelk et al., 2009).

The impact of student effort and motivation on program evaluation. One of the common areas of research centers on student effort and motivation during test events. The proliferation of high stakes assessments has prompted a need for increased evaluation of program

effectiveness. This evaluation often comes in the form of low stakes assessment or other evidence of student learning (Smiley & Anderson, 2011). Similarly, student effort has strong implications on the validity of score inferences (Swerdzewski et al., 2011; Thelk et al., n.d.), providing test scores that may not be reflective of true ability levels (Swerdzewski et al., 2011). With the increased scrutiny of program effectiveness, mainly measured by assessment, it is important to understand variance that may impact student assessment performance (Wise & Kong, 2005).

Wise & DeMars (2005b) suggest that expectancies and values are influenced by attainment (e.g., test importance), utility (e.g., future plans), intrinsic value (e.g., enjoyment), and perceived costs (e.g., what has been given up to complete the task). Yet, many low stakes assessments have minimal to no consequences to students for performance on the assessment, which may directly impact expectancies and values. Specifically, in low stakes environments, students do not perceive personal benefit; therefore, many students hold weak values, leading to low effort on the assessment (Sundre, 1999; Wise & DeMars, 2005b) which is demonstrated by rapid-guessing behaviors (Wise & Kong, 2005). In contrast, during high stakes situations, students may be aware of associated consequences of outcome performance (i.e., placement, college acceptance, licensure) (Wise & DeMars, 2005b); consequently, increasing student perceived importance and, therefore also increasing student expended effort. As a result, value could be related to test or item design (engagement) and/or assessment outcome purpose (consequence). Subsequently, as stakes of testing increase (high stakes assessment), scores on the importance scale of motivation increase (Sundre, 1999; Thelk et al., n.d.); when students understand the testing environment is consequential, they place more importance on their performance and outcome; therefore, demonstrating increase engagement.

The utilization of measurement procedures, such as RTE or self-report instruments measuring cognitive engagement and motivation, can help to provide an estimate of student effort expended during test events. Employing motivational theory procedures helps to find ways to encourage student effort; yet, additional factors such as cognitive engagement, may offer additional insight (Smiley & Anderson, 2011). Wise (2006) proposes that rapid-guessing behavior, as reported by RTE, does not need to be very high in order to impact reliability. Through these measurements, students with low RTE or self-report measures of cognitive engagement could be filtered from the sample providing a more accurate estimate of student proficiency from the remaining scores (Rios et al., 2014; Swerdzewski et al., 2011; Wise, 2006; Wise & DeMars, 2005b; Wise & Kong, 2005). The addition of filtering techniques could help to ensure that remaining scores are from motivated examinees; therefore, providing confidence in inferences made about the construct of interest being measured (Swerdzewski et al., 2011). By choosing not to filter scores from unmotivated examinees, the administrator or practitioner should treat the outcome of interest with caution. Without proper filtering or consideration, the decisions that may be made from the outcome measures may be jeopardized (Swerdzewski et al., 2011).

Common Core State Standards and college and career readiness. The current state of the United States' assessment system is continually changing, particularly as a result of the ESSA and the with the national implementation of the CCSS and focus on College and Career Readiness. The CCSS implementation is part of a national education reform, with states supporting content standards that reflect student readiness for college and career success, "…aligning states behind a select set of essential content standards that reflect the academic knowledge and skills that research suggests are more crucial for college and career success"

(Oregon Department of Education, 2014; Quay, 2010). This shift is changing the way we implement and practice instruction and assessment in schools. Additionally, the CCSS initiative "responds to the increasing concern among the public, business community, and policy makers that American students are ill-equipped to meet postsecondary and career demands and are falling behind their international peers" (Quay, 2010, p.1); the implementation of the CCSS and College and Career Readiness aimed to provide a more rigorous approach to standards reform and holding educational constituents (students, teachers, administrators, state leaders) to higher academic (teaching and learning) standards.

The implementation of CCSS has prompted a stronger focus on the national assessment policy, which includes state and district-led decisions on assessment vendors and policies, as well as additional supports and connections to the community to include College and Career Readiness, real world skills, and 21st century experiences. The CCSS and current standardized assessments (specifically the SBA) aim to foster uniformity, higher student achievement, and subsequently impact community involvement; yet, the results of this work are still being scrutinized.

Few of the CCSS standardized assessments have implemented performance-based tasks in an attempt to measure CCSS and College and Career Readiness constructs including 21st century skills and higher order thinking. These assessments, most notably, SBA and the Partnership for Assessment of Readiness for College and Careers (PARCC), are considered high-stakes assessments and help to advise future state, district, and school goals as well as impact student placement, high school graduation, or even help to inform college acceptances. Additional low-stakes formative and summative performance-based assessments, however, can also impact similar consequential outcomes and decisions. Regardless of outcomes, schools that

implement performance tasks have the ability to help support college and career readiness (Gagnon, 2010) which, in turn, may also generate a more authentic measure of student achievement. However, despite these optimistic perspectives, performance task assessments are not as widely utilized or included in school curricula or large-scale research initiatives.

Even further, the utilization of performance-based tasks may positively contribute to student engagement and motivation; therefore, adding to the dependability of outcome measures. Limitations for authentic performance tasks include high cost as well as reliability and validity concerns (McGaw, 2006; Messick, 1994). If an assessment task is authentic and includes high content and construct validity, it often includes a high price tag and could pose concerns on reliability and consistency of conditions (Garmire, 2005; Newhouse, 2011). Further, if unengaging, performance-based assessments, could limit the usability of outcomes. Additionally, the scoring of performance-based assessments can be cumbersome and lead to automatically scored items, where there depth of knowledge (DOK; Webb, 2002) is lacking, or human-scored items, which can lead to subjectivity concerns and can be time consuming.

Need for authentic assessments. The conversation around competency-based education models is gaining popularity, particularly with opportunities for students to move at an individual pace as well as provide a clearer picture of student knowledge (McClarty & Gaertner, 2015). This model of competency-based education is an advantageous approach to learning and assessment and has the ability to replace traditional practices (Newhouse, 2011). Many educational researchers and practitioners would argue that our assessments have an authenticity deficit and do not adequately measure higher-order thinking or practical skills (Lin & Dwyer, 2006; Newhouse, 2011) have called for the need to improve the validity of student assessment in order to better reflect these more difficult areas of measurement as well as continual

improvement of teaching, learning, and preparation of students for college and career readiness (Newhouse, 2011). Specifically, the need for classroom experiences, including student assessment, to reflect complexities that exist in 21st century work (Rosenbaum, Klopfer, & Perry, 2007). If implemented, these new practices, both teaching and assessment, should provide students with the ability to "work with incomplete information, adapt to changing conditions, manage complexity, and fluidly create and share knowledge" (Rosenbaum et al., 2007, p. 32).

Additionally, the paper-and-pencil assessment model has not yet demonstrated authenticity and often lacks alignment (Clarke-Midura & Dede, 2010). Areas that are harder to assess are often overlooked in typical standardized assessments (McGaw, 2006). Whereas, typical paper-and-pencil assessments attempt to demonstrate student understanding in the shortest amount of time possible (Clarke-Midura & Dede, 2010), technology-based assessments provide the opportunity to meet students at their unique academic level and provide a holistic measure of proficiency along with better alignment to curriculum and pedagogical practices.

The evolution of technology to provide authentic contexts has the ability to foster situated learning and collaborative problem solving (Rosenbaum et al., 2007). These advancements allow for exploration and experience of content within authentic contexts (Bressler & Bodzin, 2013). In order for an assessment to be authentic it must (1) contain realistic, real-world situations, (2) require judgment, (3) allow the learner to "do" or carry out tasks, (4) simulate contexts, (5) allow for integration of knowledge, and (6) provide appropriate opportunities to practice and refine performance and product (Wiggins, 1998). Many performance-based assessments have the ability to provide these components, particularly when utilizing technological enhancements and interactivity.

The use of technology to enhance learning environments has the ability to provide learners with many affordances and greater agency in their learning process (Alfieri, Brooks, Aldrich, & Tenenbaum, 2007; Slavin, Lake, Hanley, & Thurston, 2014), particularly in STEM subjects (Slavin et al., 2014). Additional improvements can take the shape of inquiry-based learning using technology, which many researchers have identified to be the best mode (Furtak, Seidel, Iverson, & Briggs, 2012; Gerjets, Scheiter, & Schuh, 2007; Jong, 2006). Additionally, the use of technology applications also have the ability to provide strong illustration of content (Slavin et al., 2014), active inquiry (Jong, 2006; Slavin et al., 2014), collaboration (Bressler & Bodzin, 2013; Slavin et al., 2014), and increased student motivation and subject relevance (Bressler & Bodzin, 2013; Rosenbaum et al., 2007; Slavin et al., 2014).

Technology-enhanced items and assessments. The use of technology in educational assessments has been examined for decades and is becoming more widely used due to innovations in technology, advanced statistical methods, and the need for the evaluation of more complex skills. Particularly, the shift to the CCSS and focus on 21st century skills has prompted a need to include the use of technology within both curriculum and assessments; existing models of assessments may not adequately measure the skills, knowledge, attitudes, and characteristics that are needed within the shifting educational landscape (Ripley, 2009).

Ripley (2009) discusses the need for technology-based assessments and cites efficiency and transformation as key factors, specifically with authentic tasks in a simulated environment. Additionally, some studies have found a significant improvement in student ability after interacting with dynamic models, particularly the ability to consider more advanced and dynamic concepts (Levy, 2013). Dynamic models provide the opportunity to follow a reasoning process based on interactions (Levy, 2013); yet, these types of authentic, technology-based assessments

are not widely implemented in current assessment practices. One particular area of deficient is the use of technology-enhanced items particularly exploring the effect of technology enhancements on student academic performance, cognition, and engagement.

Use of educational technology in high-stakes standardized assessments has been a relatively new implementation due to previous technology limitations and concerns around efficacy. The shift to technology-based assessments can bring about administrative gains and service improvements; yet, the use of technology-based performance tasks is relatively rare. This gap calls for the need to examine additional research and understanding, particularly as the use of technology-based assessments continue to gain popularity and evolve into more advanced tools with greater abilities and enhancements.

Enhanced interactivity of items allows for the incorporation of multimedia objects and evaluation of skills are not easily measured in traditional assessments (Csapo, Molnar, & Toth, 2009; Halldorsson, McKelvia, & Bjornsson, 2009; Kikis-Papadakis & Kollias, 2009; Kyllonen, 2009; Lee, 2009; Martin, 2009; Ripley, 2009; Sorensen & Andersen, 2009; Zacharia et al., 2015). Thus, allowing for the modeling of real-world complex systems allows for manipulation or participation and observing conditions (Rosenbaum et al., 2007). Particularly, technology-enhanced items in STEM can provide more realistic context through video and animations (Martin, 2009) by providing stimulating content that is not easily observed in real time (Halldorsson et al., 2009). Linn and Eylon (2011) assert that by taking advantage of technology-enhanced visualizations, "advances in technology can enable learners to explore phenomena that are too small (molecules), fast (electrons), abstract (forces), or massive (the solar system) to observe directly" (p. 186). Visualizations can also be considered a low cost alternative to real

experiments (Feurzeig & Roberts, 1999) while providing students with the opportunity to interact with traditionally inaccessible content (Levy, 2013).

An additional benefit of technology-enhanced environments results in offering learners more agency in their learning process (Zacharia et al., 2015). This agency is demonstrated by students' self-regulated learning which allows students to be responsible for their learning endeavors and managing any challenges that arise (Zacharia et al., 2015). Furthermore, the additional benefit of performance-based assessments paired with technology enhancements may provide these unique opportunities to engage in (1) realistic, real-world situations, (2) judgment, (3) 'doing' or carrying out tasks, (4) simulation contexts, (5) integration of knowledge, and (6) the practice and refine performance and product (Wiggins, 1998); yet, within the context of a computer-based assessment program. Technology-based assessments have been shown to provide more motivation and enjoyment (Halldorsson et al., 2009; Lee, 2009).

It is important to note, however, that technology-enhanced assessments may lead to academic and cognitive differences between groups of students, particularly gender differences when incorporated in STEM fields (Halldorsson, McKelvia, & Bjornsson, 2009; Lee, 2009; Ripley, 2009; Sorensen & Andersen, 2009). Gender differences can be seen on both high and low interactivity items, despite non-significant differences in paper-and-pencil assessments (Halldorsson et al., 2009). One reason for this difference may be familiarity with technology by gender (e.g., boys reporting more frequent use than girls) (Halldorsson et al., 2009; Lee, 2009; Martin, 2009) thus, a positive correlation between familiarity and achievement outcomes (Martin, 2009). Lack of technology confidence among girls may also impact performance on computer-based assessments (Sorensen & Andersen, 2009).

Assessment delivery mode. The use of technology in education has increased exponentially due to advancements in educational technology-based curricular programs, mobile applications (apps), and assessments. Through these advancements, research calls for the need to examine mode differences between paper-and-pencil assessments and non-adaptive computer-based assessments, particularly in technology-enhanced assessments. Research has mixed outcomes regarding mode impact on student achievement and engagement. The literature indicates that the psychometric qualities between the modes are the same (Ford, Vitelli, & Stuckless, 1996) but may have significantly different results between the measures (Buchanan, 2002; Gallagher, Bridgeman, & Cahalan, 2002; Hargreaves, Shorrocks-Taylor, Swinnerton, Tait, & Threlfall, 2004; Lankford, Bell, & Elias, 1994; Lee, 2004; Noyes & Garland, 2008; Smither, Walker, & Yap, 2004) particularly with personality measures (Lankford et al., 1994). Research that indicated significant differences between platforms included web-based versus paper-and-pencil comparisons (Buchanan, 2002), computer-based versus paper-and-pencil comparisons (Gallagher et al., 2002) general differences between mode types (Hargreaves et al., 2004; Noyes, Garland, & Robbins, 2004; Smither et al., 2004), as well as differences in writing performance (Lee, 2004).

There were numerous studies that found marginally significant effects (Clariana & Wallace, 2002b; Neuman & Baydoun, 1998; Vispoel, Boo, & Bleiler, 2001), particularly when timing was examined (Vispoel et al., 2001). While, other research resulted in no significance differences between modes (Bodmann & Robinson, 2004; Donovan, Drasgow, & Probst, 2000; Finegan & Allen, 1994; Horton & Lovitt, 1994; King & Miles, 1995; Mason, Patry, & Bernstein, 2001; Özalp-Yaman & Çağiltay, 2010), specifically non-significant differences between reading groups (Horton & Lovitt, 1994), measurement equivalence (King & Miles, 1995), speed and

performance (Bodmann & Robinson, 2004), and mode based questionnaires (Finegan & Allen, 1994). Additional research shows that when students are motivated, there are no mode differences (Mason et al., 2001); yet, motivation or engagement were not examined throughout most of the previous research.

The impact of computer use on academic and nonacademic outcomes. Effect on students' use of technology has been a strong focus in the literature across numerous fields. In education, researchers have studied the effect of technology use on both academic and non-academic constructs, particularly with the proliferation of technology for academic and home use. As a result, children are spending an increased amount of time consuming media each day. A 2010 study found that children consume an average of 7.5 hours of media each day (Kaiser Family Foundation, 2010). Consumption of media at this level is important to explore in order to understand time investments youth are making regarding time spent on technology.

Numerous studies have linked computer use at home with positive impacts on academic proficiency outcomes (Attewell & Battle, 1999; Bennett, Persky, Weiss, Jenkins, & Russell, 2010; Casey, Layte, Lyons, & Silles, 2012; Fiorini, 2009; Halldorsson et al., 2009; Naevdal, 2007; OECD, 2006a; Papanastasiou, Zembylas, & Vrasidas, 2003; Schmitt & Wadsworth, 2006; Tsikalas, Lee, & Newkirk, 2007; Wenglinsky, 2006) while other studies have linked heavy use of home computer to negative impacts on academic outcomes (Fuchs & Woessmann, 2004; Malamud & Pop-Eleches, 2011; Vigdor, Ladd, & Martinez, 2014), specifically in mathematics (Wittwer & Senkbeil, 2008). Other literature found no effects between home computer use and academic proficiency (Fairlie & Robinson, 2013).

Despite significant effects of home computer use on academic outcomes, it is important to note that marginalized groups, particularly females, differ in their use of technology.

Research indicates a positive impact of computer time on test scores, specifically for girls (Fiorini, 2009; Naevdal, 2007) or children from low to mid SES families (Fairlie, 2012; Fairlie & London, 2011; Fiorini, 2009; Naevdal, 2007). Other research (Fairlie, 2015) suggests differences in computer use across marginalized groups do not lead to different grades, test scores, or other educational outcomes. Further exploration on the impact of home computer use on academic proficiency outcomes is needed.

This need for research also translates to understanding how technology impacts non-cognitive skills. The literature also reports mixed outcomes. Fiorini (2009) did not find a link between computer use and the Restless and Emotional Index; however, there is a positive effect of computer use on the relationship index, specifically for girls, but that effect vanishes after two years. Literature has also explored the effects of type of computer use. Programs such as surfing the internet (Bennett et al., 2010; Casey et al., 2012), doing projects for school (Casey et al., 2012), problem solving activities (Wittwer & Senkbeil, 2008) and e-mailing or using a word processer (Bennett et al., 2010; Casey et al., 2012) are associated with higher mathematics and reading scores; while, instant messaging and downloading music (Casey et al., 2012), creating artwork (Bennett et al., 2010) or watching movies (Casey et al., 2012) are negatively associated with mathematics and reading scores.

Performance task research needs. Interest and research on performance tasks spiked in the 90s and early 2000s with paper-and-pencil performance assessments. Performance tasks are understood as concrete, goal-oriented work performed on a specific occasion and evaluated by a rater (Shavelson, Baxter, & Gao, 1993). Many considered this reform in assessment practice as a shift from traditional multiple choice testing to authentic assessment. This shift was often described as a new practice that focused on 3 Ps: performance, portfolios, and products (Madaus

& O'Dwyer, 1999) to provide measurement of authentic contexts (Wiggins, 1998). The method of presentation for the task may be paper-and-pencil (open ended problems), computer simulations, or real-time observation (Shavelson et al., 1993) and was often evaluated with a rubric or scoring guideline (Wiggins, 1998). These advancements occurred at the cusp of technology innovations; therefore, simulations may have included minimal technology enhancements in comparison to current resources. The literature discusses the need for a more direct link between instruction and assessment and the use of performance-based tasks to provide "authentic assessment formats, models, and rich descriptions of performance expectations, along with feedback to the learner … [which] work in tandem to connect instruction and learning to assessment" (Adair-Hauck, Gilsan, Koda, Swender, & Sandrock, 2006, p. 362).

The interest in authentic, performance-based assessments hit a decline post 2000, with more schools focusing on multiple choice assessments which were standards aligned and helped to identify readiness for the high-stakes, end of year, summative assessment. Newhouse (2011) indicates that this shift was due to the increasing policy on educational accountability which required accurate and reliable measures. Further, many districts utilize commercially available tests which still feature easily scored, discrete-point items (Adair-Hauck et al., 2006).

The shift to the CCSS and College and Career Readiness initiated a pursuit of alternate assessments to measure student academic and nonacademic skills, specifically higher-order tasks. These new assessments under the CCSS consider the utilization of Bloom's six levels of cognitive learning (Bloom, 1956) to include components often unmeasurable in traditional standardized assessments (Mitri, 2003) while also implementing greater DOK (Webb, 2002). Evaluations of these higher order tasks, often referred to as tacit assessment, consider more faculty to implement intuition, judgment, and feeling (Mitri, 2003) and re framed within the lens

of performance based assessments.

Despite the interest in updated performance-based assessments, minimal research has investigated the impact of new performance measures. Current CCSS standardized assessments (SBA and PARCC) have started to utilize technology-enhanced tasks; yet, results of these assessments are still under investigation. The advancements in technology-enhanced assessments and shift to performance-based assessments call for the need to examine student outcomes and potential impacts on academic proficiency and motivation. Further, the increase in game-based learning offers an entirely new lens to performance-based assessment and is only beginning to be evaluated.

Performance tasks in STEM would be an advantageous avenue to explore, particularly considering the increased focus on STEM, yet, minimal overall student interest. Of the 2014 high school graduates, 53% expressed an interest in mathematics and 46% expressed an interest in science (ACT, 2014a). The lack of aspirations towards STEM is particularly prominent in marginalized populations such as gender (Archer et al., 2010; Elster, 2014), race/ethnicity (DeWitt et al., 2013; Wang, 2013), and socio-economic status (Archer et al., 2012; Aschbacher et al., 2010). Upon high school completion, graduates report struggling in mathematics more than other subject areas; therefore, indicating a need to focus on mathematics, particularly within the scope of teaching, learning, and assessment in mathematics (Gagnon, 2010). Gagnon (2010) calls for a priority focus on graduates who are less likely to pursue STEM fields; whereas, performance assessment structures could be factors to help change and explain this trend.

Performance task framework for instrumentation. The subsequent sections briefly describe the reasoning for the development of the numerical and algebraic expressions and equations performance task measure and use of the scale. The subscales were developed and

used as theoretical subscales but may be treated to yield a single scale or subscales depending on empirical evidence following data collection.

Performance task internal consistency. The performance instrument was developed based on the CCSS 7.EE.3 standard: *Solve real-life and mathematical problems using numerical and algebraic expressions and equations* (National Governors Association, 2010) which includes the integration of positive and negative numbers, specifically within equations and tying in other mathematical content strands such as number systems and mathematical practice (Schwols & Dempsey, 2013). The performance measure construct is a mathematical proficiency construct measured by the CCSS of solving real-life and mathematical problems using numerical and algebraic expressions and equations. The mathematical construct (Solving real-life and mathematical problems using numerical and algebraic expressions and equations) is measured within a formative assessment performance instrument for examinees to display understanding of solving algebraic expressions.

Solving real-life and mathematical problems using numerical and algebraic expressions and equations is a CCSS that includes the integration of positive and negative numbers, specifically within equations and tying in other mathematical content strands such as number systems and mathematical practice (Schwols & Dempsey, 2013). The measurement of this skill includes solving multi-step, real world problems using positive and negative numbers (in any form), applying properties of operations, converting between forms, assessing the reasonableness of answers, and using mental computation and estimation strategies, specifically within equations that compare algebraic solutions (Schwols & Dempsey, 2013).

The measurement of one's mathematical ability within the construct of solving real-life and mathematical problems using numerical and algebraic expressions and equations is

measured across five main items, each containing several sub items and, therefore, utilizing polytomous scoring (see Appendix F). Appendix F displays the specific grading rubric including the items, number of points awarded, and DOK for items. The range for the total number of points possible for the performance task is zero (minimum score) to 29 (maximum score) across five items: (a) item one worth eight; (b) item two worth one point; (c) item three worth seven points; and (d) item four worth six points, and (e) item five worth seven points. Maximum points are awarded for complete answers while partial credit is awarded for limited answers. Moreover, the study aims to focus on solving real-life and mathematical problems using numerical and algebraic expressions and equations for students specifically within a low stakes formative assessment context.

The mathematical construct of solving real-life and mathematical problems using numerical and algebraic expressions and equations is evaluated by the performance instrument item and total score. The performance instrument includes four tasks (i.e., items) with each task including multiple items. The five main items within the performance task are measured for difficulty using Webb's Depth of Knowledge and curricular elements (i.e., assessment components) are categorized based on cognitive demands reflecting a depth of knowledge required to correctly solve the item (Mississippi Department of Education, 2009). Webb's models include four main depths of knowledge: (a) level one – recall and reproduction; (b) level two –skills and concepts; (c) level three – short-term strategic thinking; and (d) level four – strategic thinking (Mississippi Department of Education, 2009). The four levels reflect the amount of work required to solve the problem (Mississippi Department of Education, 2009).

The five main items within the performance task are measured for difficulty using Webb's Depth of Knowledge and curricular elements (i.e., assessment components) and

categorized based on cognitive demands reflecting a depth of knowledge required to correctly solve the item (Mississippi Department of Education, 2009). The items used range in difficulty and DOK, which is reflected in the rubric scoring. Four of the five items (i.e., items one, three, four, five) within the mathematics performance task include a DOK score of three; therefore, items are ranked in difficulty by specific requirements to solve each item, as classified by NWEA® senior mathematics content specialists. As mentioned earlier, items are polytomously scored with item one worth the highest score with eight points, item three and five worth seven points, item four worth six points, and item two worth one point (Appendix F; Table 1.1), as indicated by NWEA® senior mathematics content specialists. Item one is the most difficult of the items because it requires a full solution. Then, item five is ranked as the next difficult item because of the number of different variables required to solve, including like terms followed by items three and four because they break down components of item one with item four including more variables than previous items. Item two is ranked as least difficult because it requires basic strategy identification. Table 1.1 outlines the items by DOK, difficulty, and points.

The current study uses the range of total number of points possible from zero (minimum score) to 29 (maximum score) across five items: (a) item one worth eight; (b) item two worth one point; (c) item three worth seven points; and (d) item four worth six points, and (e) item five worth seven points. Maximum points are awarded for complete answers while partial credit is awarded for limited answers.

Table 1.1

Mathematics Performance Instrument Ranking of Items by DOK, Points, and Difficulty

Item	DOK	Total points	Difficulty rank
Item 1	3	8	1
Item 2	1	1	5
Item 3	3	7	3
Item 4	3	6	4
Item 5	3	7	2

Note. Difficulty is ranked from 1 (hardest) to 5 (easiest).

Interim assessment framework for instrumentation. Measures of Academic Progress (***MAP®) scores.*** The NWEA MAP® assessment is used in the study as a measure of student achievement. MAP® assessment is a low-stakes interim assessment that is administered seasonally (fall, winter, spring). MAP® is a multiple choice, computerized adaptive test (CAT) constructed based on a students' unique performance in response to items constrained in content by standards (Thum & Hauser, 2015). MAP® utilizes an algorithm to estimate student achievement level after each item response; subsequent items are selected based on a matched difficulty level to the student achievement and a 50% probability of a correct response (Northwest Evaluation Association, 2011). Items are selected from the NWEA™ item pool and are based on the test taker ability estimate (Thum & Hauser, 2015). Subject specific items (e.g., mathematics) are calibrated to the same vertical scale known as the RIT[1] scale which is based on a one-parameter (1PL) logistic Item Response Theory (IRT) model, also known as the Rasch model (Northwest Evaluation Association, 2011), see Footnote 1.

The reliability of the MAP® assessment examines the consistency of the assessment and

[1] RIT scale corresponds to Rasch-Unit and is expressed as $P_{ij} = \frac{e^{(\theta_j - \delta_i)}}{1+e^{(\theta_j - \delta_i)}}$.

cannot be measured using traditional methods; test-retest or parallel forms are not possible due to the adaptive nature of the assessment (Northwest Evaluation Association, 2011). Therefore, a "stratified, randomly-parallel form reliability" (Green, Bock, Humphreys, Linn, & Reckase, 1984, p. 353) is used as a measurement of reliability for a CAT, such as MAP®. This measurement of reliability can be framed by "correlations between two tests administered from two different but related item pools and those administered twice but from different item pools" (Northwest Evaluation Association, 2011, p. 55). According to the Northwest Evaluation Association (2011), reliability across 40 states for sixth grade ranges between 0.792 and 0.906, reliability for seventh grade ranges between 0.730 and 0.910, and reliability for eighth grade ranges between 0.716 and 0.905. These estimates were derived from minimum and maximum correlations for MAP® mathematics tests with different item pool structures between both spring 2008-fall 2008 as well as spring 2008-spring 2009 for 40 states[2] (Northwest Evaluation Association, 2011).

Additional reliability evidence is derived from correlations of MAP® scores between terms (e.g. Spring 2012) with the same students tested the following spring (e.g., Spring 2013) (Northwest Evaluation Association, 2011). According the Northwest Evaluation Association (2011), reliability across states for sixth grade ranges between 0.778 and 0.925, reliability for seventh grade ranges between 0.827 and 0.917, and reliability for eighth grade ranges between 0.820 and 0.927. These estimates were derived from minimum and maximum test-retest correlations for MAP® mathematics tests with common item pool structures between spring 2008-fall 2008[3], fall 2008-spring 2009[4], and spring 2008-fall 2009[5] (Northwest Evaluation

[2] 40 states include AK, AR, AZ, CA, CO, DE, GA, IA, ID, IL, IN, KS, KY, MA, ME, MI, MN, MO, MT, NC, ND, NE, NH, NJ, NM, NV, NY, OH, OK, OR, PA, RI, SC, TX, UT, VA, VT, WA, WI, and WY.
[3] Spring 2008-Fall 2008 states include all states from Footnote 2 except for ME, MO, NM, RI, UT as well as the addition of HI.

Association, 2011).

Lastly, reliability estimates for MAP® state content-alignment in mathematics for spring and fall of 2008 and spring of 2009[6] show reliability across all states for sixth grade ranges between 0.952 to 0.970, reliability for seventh grade ranges between 0.958 and 0.973, and eighth grade ranges between 0.961 and 0.975 (Northwest Evaluation Association, 2011).

The validity of the MAP® assessment examines whether the test measures what it intends to measure and if the results can be used in decision making (Kane, 2001; Northwest Evaluation Association, 2011). The evidence of validity provides "adequacy and coverage of a test's content, to its ability to yield scores that are predictive of a status in some area, to its ability to draw accurate inferences about a test taker's status with respect to a construct, to its ability to allow generalizations from test performance within a domain to like performance in the same domain" (Northwest Evaluation Association, 2011, p. 182). This evidence can be measured through content validity, concurrent validity, predictive validity, and criterion-related validity.

Content validity for MAP® is ensured through test and item development based on procedural evidence; content specialists group state standards into test design by goals and sub-goals (Barker, 2015). Additionally, goals and sub-goals are grouped by state standards and content standards and are mapped through advanced software and content expert verification (Barker, 2015; Northwest Evaluation Association, 2011). This process occurs through numerous iterations as well as bias and sensitivity validation by internal NWEA™ content specialists and item validation by internal NWEA™ researchers. Classification accuracy and decision

[4] Fall 2008-spring 2009 states include all states from Footnote 2 except for MT and NV as well as the addition of CT, HI, LA, and MS.
[5] Spring 2008-Spring 2009 states include all states from Footnote 2 except MT and NV as well as the addition of CT, LA, and MS.
[6] Spring 2008 to Fall 2009 states include all states from Footnote 2 with the addition of FL, MT, NV, and TN.

consistency was evaluated across twenty-six states[7] based on state content aligned MAP® mathematics tests administered in the same term as the state accountability test in the 2008-2009 year (Northwest Evaluation Association, 2011).

Concurrent validity for MAP® is expressed as a Pearson correlation coefficient between total domain area RIT score and the total scale score of another assessment within the same domain area (Northwest Evaluation Association, 2011). Both tests must be administered within two to three weeks of each other; strong correlations would be indicated by Pearson correlation coefficients in the mid .80s (Northwest Evaluation Association, 2011). However, tests that do not use multiple choice items tend to have lower correlations than tests that have multiple choice items (Northwest Evaluation Association, 2011). Concurrent validity for mathematics is based on performance on state accountability tests by state content aligned MAP® tests show a Pearson correlation coefficient ranging from .746 to .876 for sixth grade, .698 to .871 for seventh grade, and .704 to .878 for eighth grade[8] (Northwest Evaluation Association, 2011).

Predictive validity for MAP® measures the relationship between MAP® performance to performance on another test in the same domain that is taken at a future date (Northwest Evaluation Association, 2011). Tests must be administered several weeks apart and strong predictive validity can be inferred for correlations in the low .80s (Northwest Evaluation Association, 2011). Predictive validity of predicted performance on state accountability tests in mathematics and by state content aligned MAP® tests show a correlation coefficient ranging from .745 to .859 for sixth grade, .637 to .869 for seventh grade, and .583 to .868 for eighth

[7] States include AZ, CA, CO, DE, GA, IL, IN, KS, MA, ME, MI, MN, MT, ND, NH, NJ, NM, NW, OH, OR, RI, SC, TX, WA, WI, WY.
[8] Twelve states were included in the measurement of concurrent validity including AR, CA, CO, FL, GA, KY, ND, NC, PA, SC, WI, and WY.

grade[9] (Northwest Evaluation Association, 2011).

Criterion-related validity for MAP® measures the extent test scores relate to external performance (e.g. graduate-not graduate) (Northwest Evaluation Association, 2011). For MAP®, proficiency level on the state assessment is often used as a measure of external criterion (e.g. proficient-not proficient) (Northwest Evaluation Association, 2011). Correlation coefficients for criterion-related validity estimates will also be smaller than correlations from test performances expressed as scale scores (Northwest Evaluation Association, 2011). Validity of criterion-related performance on state accountability tests for mathematics by state content aligned MAP® tests show a correlation coefficient ranging from .624 to .715 for sixth grade, .589 to .722 for seventh grade, and .570 to .724 for eighth grade[10] (Northwest Evaluation Association, 2011).

Summary and Study Context

The purpose of this study is to help to address the need for technology-enhanced assessment research, particularly in mathematics, and to evaluate mode differences as well as student cognitive engagement. Motivation for the study includes the need to understand student effort and interest in STEM fields, better understanding of student motivation and effort during assessment, measurement of academic program effectiveness, and the need for current technology-enhanced assessment research, specifically within performance-tasks. Previous studies have explored the use of technology in assessment, mode differences, student effort, and the need for authentic assessments to measure the CCSS and College and Career Readiness; however, minimal research has been completed to measure the effect of a technology-enhanced

[9] Nine states were included in the measurement of predictive validity including AR, CA, CO, FL, GA, NC, ND, PA, and SC.

[10] Seven states were included for the measurement of criterion-related validity for sixth and seventh grade (AR, CA, CO, FL, GA, KY, SC), six states were included for the measurement of criterion-related validity for eighth grade (CA, CO, FL, GA, KY, SC).

performance task on student cognitive engagement in mathematics.

The study builds on the expectancy-value model using the Student Opinion Scale (SOS; Sundre, 1997) and the CES and CE-S (Greene & Miller, 1996; Smiley & Anderson, 2011) to develop a self-report measure (CE-S-DSP & SOS) to examine cognitive engagement within low-stakes assessment contexts, namely within a mathematics performance task instrument. Cognitive engagement was measured across three performance instrument modes: (a) paper-and-pencil; (b) technology-enabled, which was converted with fidelity to paper-and-pencil but ported on the computer device; and (c) technology-enhanced, with technology enhancements employing more technology affordances and innovations to support interactivity and agency than the technology-enabled fidelity performance instrument or paper-and-pencil instrument. The self-report measure (CE-S-DSP & SOS) explored the engagement subscales of importance, effort, processing (deep and shallow), and persistence of middle school students (grades 6-8). The measure expands on the Student Opinion Scale (SOS; Sundre, 1999) and the Cognitive Engagement Survey (CES; adapted from Miller, Green, Montalvo, Ravindran, & Nichols, 1996) to create the Cognitive Engagement Scale – Short – Deep, Shallow, Persistence (CE-S-DSP) combined with the SOS, see Figure 1.1. The SOS measures affective aspects of student motivation, including self-reported beliefs of importance (perceived value) and effort (theoretical value). The CES measures motivation as evidence by degree of self-reported cognitive engagement with two scales of processing (deep and shallow) and persistence. The CE-S-DSP & SOS was used to estimate the construct of cognitive engagement with middle school students (grades 6-8) on three mathematics performance task modes. A student interview survey was administered on student attitudes towards technology-enhanced assessments.

The study aims to help better understand student cognitive engagement within low-stakes

performance-based instruments, specifically in STEM. Results of the study could help explore factors influencing student interest in STEM programs and careers, measurement of program effectiveness, and differences in cognitive engagement across assessment modes, particularly technology-enhanced modes. Gender differences and frequency and type of computer use at home was also explored.

Research Questions and Contributions

The literature pool of research supports the need for further investigation of student cognitive engagement within assessment, specifically technology-enhanced performance-based assessments. Furthermore, additional research is needed on the effect of assessment modality differences, sex, and race/ethnicity on student academic proficiency and cognitive engagement. The literature findings suggest that race/ethnicity, SES, and sex are significant factors in student pursuit and self-efficacy in STEM and should be considered for further investigations (Louise Archer et al., 2010; DeWitt et al., 2013; Elster, 2014; Wang, 2013). The body of literature proposes that further examination of cognitive engagement within an assessment context is needed and may help provide better academic outcomes for students, more reliable measures of program effectiveness, and a better understanding of the impact of assessment modes on student engagement.

The findings of this literature review create an impetus for further research on (a) factors impacting student interest in STEM, (b) appropriate use of student outcomes for the measurement of program effectiveness, (c) differences in student cognitive engagement across assessment modes, particularly technology-enhanced modes, (d) the impact of sex on cognitive engagement across assessment modes, and (e) the impact of type of computer use at home on student cognitive engagement. Additionally, more mixed methods research, is needed to fully

understand the factors that impact student engagement within technology-based assessment modes.

Since this initial research in these areas ware conducted, there have been numerous advancements in the educational landscape. First, the national implementation of the CCSS and emphasis on College and Career Readiness has shifted the focus of educational curricula, standards, and assessment practices. Additionally, the increase in technological advancements for teaching, learning and assessment has prompted the creation of new standardized assessments (often high stakes) with minimal understanding of implications and impact on student cognitive engagement and performance. Finally, research rigor including experimental or quasi-experimental, pre/posttest design, and treatment/control groups could be implemented to ensure best practices in research methods. There is also a need for research on the impact of technology-enhanced assessments on special education and/or English language learners, as none of the studies included evaluate either population.

The findings from this literature review may help inform district-wide decisions about new assessment programs, primarily technology-enhanced assessments, as well as the impact on student cognitive engagement, achievement, and marginalized populations. Schools may choose to implement technology-enhanced assessments, alter curricular components, or encourage more computer use among students. Additionally, the findings could help inform policy makers and assessment companies, particularly as the national interest in technology-enhanced assessment grows, specifically within STEM. Through careful planning and implementation, leaders can see improved results in students' cognitive engagement, STEM interest, and achievement, as well as a better understand of how to interpret outcomes for program evaluation.

The findings from this literature review suggest that there is a need to evaluate student

cognitive engagement within a technology-enhanced performance assessment, specifically across modality type. The proposed research study represents an important step in filling the gap in researching the measurement of cognitive engagement within a technology-enhanced performance instrument in mathematics.

CHAPTER II

METHODS

This chapter introduces the study methodology, beginning with defining the study design, conditions, and division of subjects within the design. The chapter then discusses the components of the performance task across the three modalities (i.e., platforms) in the study, along with the mathematics features and items within each modality. This chapter also discusses the sampling design and participants. Next, the instrumentation section summarizes the use of cognitive engagement measures discussed in the literature review in the last chapter and concludes with the data analysis plan used to address the hypotheses and research questions of the study.

The study implements a single group counterbalanced design with three conditions: (1) paper-and-pencil, (2) technology-enabled, and (3) technology-enhanced. Here, counter-balanced means the systematic division of subjects across groups. The counterbalanced design in the current study allows for the student sample (N =450) to be divided into six groups to systematically organize the order of treatment (Shuttleworth, 2009), see Figure 2.1. The independent variables of the study are the three mathematics performance instrument assessment conditions (categorical variable): (a) paper-and-pencil, (b) technology-enabled, and (c) technology-enhanced. For the quantitative analyses, the dependent variable is the total score on the measure of cognitive engagement. For the qualitative analyses, the dependent variable is total score on the mathematics performance task.

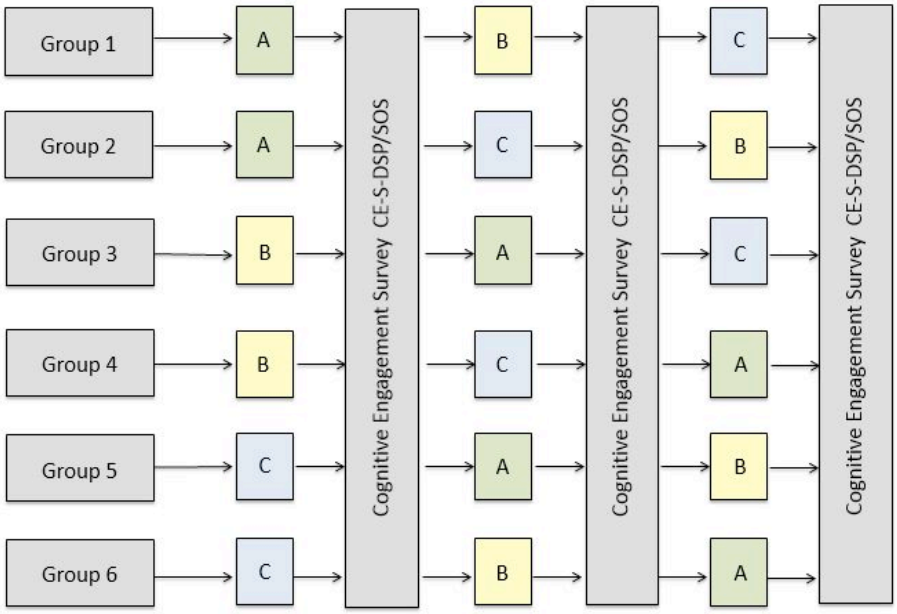


Figure 2.1. Single group with counterbalancing. Adapted from Shuttleworth, M. (2009). Counterbalanced Measures Design. Retrieved September 7, 2015, from https://explorable.com/counterbalanced-measures-design. *n* =450 students are divided between the six groups within the design.

Sample

Sampling design. Participants for this study include a convenience sample ($N = 450$) of students in grades 6, 7, and 8 from one K-12 public school in Oregon ($n = 122$), one K-8 private school in Washington ($n = 74$), and one K-8 private school in North Carolina ($n = 254$). All students within grades 6-8 at the participating schools who did not choose to opt-out assisted in the study. A subset of participants ($n \approx 7$) from the Washington school were selected to participate in qualitative follow-up interviews, see Appendix G, to gain additional information about the features of the technology-enhanced mathematics performance instrument.

The data were collected and made available through a Portland-based NWEA™ research project, employing the cognitive engagement and mathematics instruments as described in the *Instruments* section of this chapter. The schools used in the study are part of the NWEA™ *Partners in Innovation Program* which provides monetary incentives to schools for voluntary

participation in annual research studies (*Partners in Innovation* brochure provided in Appendix H and *Partners in Innovation* legal documentation provided in Appendix I). However, student participants do not receive direct compensation for participation.

Grades 6, 7, and 8 were selected as the focus group for this study in large part due to the research that discusses the shift in STEM aspirations between the ages of 10 and 14 (Louise Archer et al., 2010, 2012; DeWitt et al., 2013; Tai et al., 2006), see Chapter I section on "Aspirations between Ages 10 to 14." Further, the mathematics performance instrument focuses on seventh grade CCSS mathematics curriculum (CCSS 7.EE.3 *Solve real-life and mathematical problems using numerical and algebraic expressions and equations;* National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). The inclusion of sixth and eighth grade provided a larger sample size within the same grade band for this study on how aspects of engagement may be related to modality.

Number of participants. The number of students selected to participate in the study was largely due to availability (size of participating schools); however, in order to obtain meaningful outcomes, the sample size needed for statistical tests within the study was also a strong factor in obtaining an appropriate sample size. I conducted a power analysis (Faul, Erdfelder, Lang, & Buchner, 2007, 2009) in order to determine the number of participants needed to detect effects from the independent variables based on (a) size of effect in the population variables, (b) statistical tests used, and (c) level of significance (Rudestam & Newton, 2007). Measures of an effect size can provide standardized and objective magnitude measures of an observed effect, helping to allow comparison of effects across numerous studies (Field, 2013).

The level of power describes the probability that a statistical test will find an effect assuming there is an effect in the population, or in other words, the probability of correctly

rejecting the null hypothesis when it is false. Of course, bigger effects are easier to detect, and smaller effects require a larger sample size to detect. Statistical power is inversely related to beta or the probability of making a Type II error (power = 1 – β). Power of a statistical test involves four main parameters (effect size, sample size, alpha significance level, and beta), all of which are mathematically related so that identifying any three then specifies the value of the fourth. For this study, the alpha significance level will be set at .05 and beta will be set at .2 (thus power = .8), which are typical values for social science research (Field, 2013).

A one-way independent analysis of variance (ANOVA) evaluated the independent variance of modality type with three levels: (a) technology-enhanced, (b) technology-enabled, and (c) paper-and-pencil on the dependent variable of total score on the cognitive engagement measure (CE-S-DSP & SOS). A power analysis for the one-way ANOVA was conducted to evaluate sample sizes necessary to achieve a power level of .80 and a significance level of .05, with three treatment groups (levels)[11]. Table 2.1 shows the power analysis for a one-way independent ANOVA indicating the sample size needed for a small effect (f = .10), the sample size needed for a medium effect (f = .25), and the sample size needed for a large effect (f = .40) (Reid, 2013). Based on Table 1.1, in order to detect a medium effect size (f = .25) in the study using these parameters, a sample size of n = 159 is needed for one-way ANOVA.

[11] Note that the design in Figure 2.1 does indicate a potential violation to some degree of independence for the ANOVA tests, since the same students are used in each treatment condition through the rotated design. The rotated design is typically used in software development to balance treatment conditions. Limitations of the design are discussed at the end of chapter.

Table 2.1

Power Analysis for a One-Way ANOVA

Effect size	Total *N*
Large	66
Medium	159
Small	969

Note. The above power analysis is based on a power level of .80, a significance (alpha) level of .05, *df* = 3, and three groups. Power analysis was conducted using G*Power 3.1 software (Faul et al., 2007, 2009). Note that other programs may slightly reduce the number of subjects by five or less as compared to results in Table 2.1, due to how rounding error is handled within the algorithms. The larger number is used here for caution in estimating subject recruiting needs.

A three-way independent ANOVA evaluated the effect of three independent variables on the dependent variable of total score on the cognitive engagement measure (CE-S-DSP & SOS). The independent variables modality type with three levels: (a) technology-enhanced, (b) technology-enabled, and (c) paper-and-pencil on, sex with three levels: (a) male, (b) female, and (c) other, and race/ethnicity with seven levels: (a) White, (b) Black, (c) Hispanic, (d) Asian Pacific Islander, (e) Native American, (f) two or more races, and (g) other. A power analysis for the three-way ANOVA was conducted to evaluate sample sizes necessary to achieve a power level of .80 and a significance level of .05. Table 2.2 shows the power analysis for the three-way ANOVA indicating the sample size needed for a small effect (*f* = .10), the sample size needed for a medium effect (*f* = .25), and the sample size needed for a large effect (*f* = .40) (Reid, 2013). Therefore, to achieve a medium effect size (*f* = .25) in order to achieve a power level of .80 and a significance level of .05, a sample size of *n* = 314 would be needed, see Table 2.2.

Table 2.2

Power Analysis for a Three-Way ANOVA

Effect Size	Total *N*
Large	130
Medium	314
Small	1894

Note. The above power analysis is based on a power level of .80, a significance (alpha) level of .05, *df* = 3, and three groups. Power analysis was conducted using G*Power 3.1 software (Faul et al., 2007, 2009). Note that other programs may slightly reduce the number of subjects by five or less as compared to results in Table 2.2, due to how rounding error is handled within the algorithms. The larger number is used here for caution in estimating subject recruiting needs.

I compared sample size due to the availability of student participants total across all schools and grades to the above power analyses results to help guide the sample selection. The overall combined sample size ($N = 450$) included enough participants to have a probability at the defined levels above of detecting a medium-sized effect for both the one-way ANOVA and three-way ANOVA analyses used here in the results section across the combined sample if, in fact, there is an effect to detect. The additional one-way ANOVA evaluating the effect of time spent on technology at home (in hours per day) has a smaller sample size ($n = 309$) that also meets the criteria of an appropriate sample size to detect a medium-sized effect. The breakdown of sample sizes between schools and grades are shown in Table 2.3.

Table 2.3

Sample Sizes Between Schools and Grades: (N = 450)

School and grade	*n*	*%*
Washington		
Grade six	20	0.04
Grade seven	21	0.05
Grade eight	33	0.07
Oregon		
Grade six	37	0.08
Grade seven	46	0.10
Grade eight	39	0.09
North Carolina		
Grade six	72	0.16
Grade seven	86	0.19
Grade eight	96	0.21

Note. Total grade six sample $n = 129$; total grade seven sample $n = 153$; total grade eight sample $n = 168$. Total sample size (grades 6-8) $N = 450$.

Students in each school were randomly assigned to one of six treatment groups, as per computer generated identification numbers (IDs). The counterbalanced design allows for division of subjects (N =450) into six groups which systematically organizes the order of treatment (Shuttleworth, 2009), see Figure 2.1. The division of subjects between groups were computer generated and random, with each group receiving approximately equal participants. All participants (n =450) completed all three assessment conditions (Appendices C-D), three subsequent student engagement surveys (Appendices A-B), demographic data entry, and a final user survey on technology use at home (Appendix J). A purposive subset of participants (n = 7) was selected for qualitative interviews in order to gather a ride range of cases for variation on

dimensions of interest (Patton, 2001). As a result, participants from the Washington school were selected based on sex and time of completion of the performance tasks (early-finisher, mid-finisher, and late-finisher) to participate in an interview to gain additional information about the feature of the technology-enhanced mathematics performance instrument (see Appendix G) and overall satisfaction of the technology used. Selection criteria (sex and completion time) was used to ensure diversity of interview participants. For example, a student who was an early-finisher may have been selected for an interview but, due to the duration of the interview and time constraints of the classes, a second student may not have been interviewed during the same class period. As a result, the second class may have targeted a mid- or late-finisher for an interview. In one rare instance, two students were able to interviewed during the class period resulting in three students who participated in the interview in grade six. As a result, seven students were selected to participate in the interview.

Instruments

The next section describes the instruments utilized in the study. The instruments include the measures of cognitive engagement (CE-S-DSP & SOS), the performance task, and student MAP® scores from the NWEA™ interim assessment. The first instrument includes the cognitive engagement measures include a self-report Likert scale survey that measures engagement across five factors. The second instrument is the mathematics performance task includes five tasks (i.e., items) with each task including multiple items in an attempt to measure student mathematics proficiency. Finally, the third instrument is the NWEA™ MAP® assessment which is a low-stakes interim assessment that measures the construct of mathematics using a multiple choice, computerized adaptive test (CAT) format.

Cognitive engagement measures. This study uses the CE-S-DSP as a self-report measure to operationalize the factors of deep strategy, shallow processing, and persistence, as described in Chapter I. It is intended to measure subscales of (a) deep processing, (b) shallow processing, and (c) persistence, see Appendix A. The study also incorporates the SOS (developed by Sundre, 1999) to include the constructs of importance (perceived value) and effort (theoretical value) as additional factors to measure cognitive engagement, see Appendix B. The complete measure of cognitive engagement is shown in Appendices A-B (CE-S-DSP & SOS).

Administration and scoring. The present study utilizes the cognitive engagement survey (CE-S-DSP) to measure the subscales of deep processing, shallow processing, and persistence across a total of ten questions (adapted from Miller et al., 1996 and Smiley & Anderson, 2011). Study participants completed the survey after each mathematics performance instrument for a total completion of three measures of cognitive engagement (see Figure 2.1). The CE-S-DSP uses a four-point Likert scale anchored with strongly disagree and strongly agree. Four points were chosen to omit a neutral option and use forced choice. Deep strategy has four questions yielding a total score of 16, shallow strategy has three questions yielding a total score of 12 (28 total for the processing subscale), and persistence has four questions yielding a total score of 16. The outcome measure for the CE-S-DSP yields a total score of 44.

Additional factors of cognitive engagement include perception of importance (perceived value) and amount of effort exerted on the test (theoretical value) (Thelk et al., 2009). The SOS (Sundre, 1999), which contains the original items for this study, is based on the expectancy-value model re-developed by Pintrich (1989), originally by Eccles et al. (1983) and includes ten Likert items across the subscales of importance (five items) and effort (five items). The original SOS (Sundre, 1999) uses a five-point Likert scale anchored with strongly disagree and strongly agree

including a neutral choice. The current study revised the SOS (Sundre, 1999) scale to include a four-point Likert scale by omitting the neutral choice. The scale in the current study was anchored with strongly disagree and strongly agree and included four points to omit the neutral option and used forced choice. Response data collected is categorical, sum data is continuous. Importance has five questions yielding a total score of 20 and effort has five questions yielding a total score of 20. The outcome measure for the SOS fields a total score of 40. The SOS (Sundre, 1999) was included as an additional measure within the current study because it may include additional components beyond Miller et al.'s (1996) CES scale. It may help provide a more robust measurement of cognitive engagement, specifically within a low stakes assessment context.

Achievement measures. The study utilizes two measures of academic achievement in mathematics. Both measures aim to determine student achievement outcome as a result of a test event. One of the achievement measures (performance instrument) is a newly created assessment and the second achievement measure (MAP®) is an existing measure.

Performance instrument model and rubric. The study utilizes a mathematics performance instrument as a student achievement measure for mathematics. The current study measured the mathematical construct through the use of three mathematics performance instrument assessment modes: (1) paper-and-pencil, (2) technology-enabled, and (3) technology-enhanced. The paper-and-pencil assessment includes the performance instrument in a traditional paper test format and does not include use of technology. The technology-enabled assessment provides the performance instrument on a computer modality which is the assessment converted with fidelity from paper-and-pencil and ported on the computer. The technology-enhanced assessment employs more technology affordances and innovations than the technology-enabled

assessment including use of an avatar, animation, and interactivity of items. The performance task specifications across the three modes are outlined in Appendix C. Examples of the performance task are shown in Appendices D-E.

In the current study, the measurement of one's mathematical ability of solving real-life and mathematical problems using numerical and algebraic expressions and equations is measured across five main items, each containing several sub items and, therefore, utilizing polytomous scoring (see Appendix F). Appendix F displays the specific grading rubric including the items, number of points awarded, and DOK for items. The range for the total number of points possible is zero (minimum score) to 29 (maximum score) across five items: (a) item one worth eight; (b) item two worth one point; (c) item three worth seven points; and (d) item four worth six points, and (e) item five worth seven points. Maximum points are awarded for complete answers while partial credit is awarded for limited answers. Moreover, the study aims to focus on solving real-life and mathematical problems using numerical and algebraic expressions and equations for students specifically within a low stakes formative assessment context.

Due to scoring limitations that occurred mid-development of the performance tasks, scoring for the performance tasks was not automated, as originally anticipated. As a result, for the purposes of the current study, performance tasks were scored for the purposive subset of students only and were analyzed qualitatively. Scoring for the remainder of the student tasks was completed on an as needed basis as part of the larger NWEA™ research study project but were not included in this study. Scoring guidelines for both the current study as well as the larger scope of work at NWEA™ were the same.

Administration and scoring. The study implemented all three assessments per student with each round of assessments comprised of one of the three 20-minute (60-minute total) mathematics performance instruments which are randomly assigned to each student. The performance instrument score was continuous on a scale of zero (minimum score) to 29 (maximum score). Rubrics outline depth of knowledge, student demonstrated proficiency definitions, and possible responses (see Appendix F). The performance instrument includes five tasks (i.e., items) with each task including multiple items. The five main items within the performance task are measured for difficulty using Webb's Depth of Knowledge and curricular elements (i.e., assessment components) are categorized based on cognitive demands reflecting a depth of knowledge required to correctly solve the item (Mississippi Department of Education, 2009). Four of the five items (i.e., items one, three, four, five) within the mathematics performance task include a DOK score of three; therefore, items are ranked in difficulty by specific requirements to solve each item, as classified by NWEA™ senior mathematics content specialists. The current mathematics performance instrument includes items with a depth of knowledge of three, as ranked by NWEA™ senior mathematics content specialists.

As mentioned earlier, items are polytomously scored with item one worth the highest score with eight points, item three and five worth seven points, item four worth six points, and item two worth one point (Appendix F; Table 1.1), as indicated by NWEA™ senior mathematics content specialists. Item one is the most difficult of the items because it requires a full solution. Then, item five is ranked as the next difficult item because of the number of different variables required to solve, including like terms followed by items three and four because they break down components of item one with item four including more variables than previous items. Item two is ranked as least difficult because it requires basic strategy identification. Table 1.1, from Chapter

I, outlines the items by DOK, difficulty, and points.

MAP® scores. As mentioned in Chapter I, the MAP® assessment is a low-stakes interim assessment that is administered seasonally (fall, winter, spring). MAP® is a multiple choice, computerized adaptive test (CAT) constructed based on a students' unique performance in response to items constrained in content by standards (Thum & Hauser, 2015). MAP® scores were collected for the purposive subset of students ($N = 7$). This subsample of students completed the MAP® test in mathematics for fall 2016, the same season they completed the mathematics performance task for the current study. The relationship between student MAP® scores ($N = 7$) and student scores on the mathematics performance task were explored. Additionally, the relationship between students' total engagement score on the CE-S-DSP and performance task performance were examined.

Qualitative measure - student interview. The current study implemented a qualitative interview aimed to obtain additional information about the features of the technology-enhanced mathematics performance instrument (see Appendix G). The student interview included four questions collecting open ended (qualitative) responses about the features of the technology-enhanced performance instrument modality as well as overall interest in technology-based assessments. A purposive subset of participants ($N = 7$) using maximum variation sampling was selected to participate in the interview based on variation on dimensions of interest (Patton, 2001) such as student sex and time of completion of the performance tasks (early-finisher, mid-finisher, and late-finisher). Interviews occurred face-to-face following the completion of the three mathematics performance instruments and subsequent CE-S-DSP & SOS surveys. The interviews were audio recorded and transcribed for data analysis. The data transcribed was

assigned with the student identification number in order to analyze themes between performance task achievement, cognitive engagement surveys, and interview data.

The qualitative research plan and data collection was considered in order to mitigate any threats to validity. To begin, the current study utilizes data triangulation by implementing multiple and different methods (qualitative and quantitative data), (Miles, Huberman, & Saldana, 2013; Onwuegbuzie & Leech, 2007). Utilizing a triangulation approach helps to reduce the possibility of chance associations, systematic biases, and providing great confidence in interpretations (Maxwell, 1992). In the current study, quantitative data was used to support the qualitative interpretations (Eisner, 1991). Additionally, the inclusion of specific participant quotations were used to avoid any threats to descriptive validity which may occur in translation of qualitative data (Maxwell, 1992).

Demographic variables. Demographic variables were collected to account for additional variance between students. The demographic variables collected at the student level include: (a) sex, (b) technology use at home (in hours per day), (c) type of technology use at home (e.g. software), (d) student birthday, (e) school, and (f) race/ethnicity. Demographic variables of sex, birthday, school, and race/ethnicity were collected on the initial login page and amount of technology use at home and type of technology use at home was collected at the end of the assessment. Sex included a text box to avoid a gender dichotomy and student birthday was used to determine student exact age at time of test. Race and ethnicity data were collected. Student birthday (i.e. age) and school were not further evaluated within the present study.

Home technology variables were collected to determine amount of technology use at home per day, modality, and type of technology use occurring at home. Amount of time spent on technology at home was a categorical variable measured in hours per day with five levels: (a)

none; (b) less than one hour; (c) between one and three hours; (d) between three and five hours; and (e) more than five hours. Modality type was collected and aggregated by device (e.g., desktop computer, laptop computer, tablet, smartphone) allowing for multiple item selection by use of checkboxes, specifics are displayed in Appendix J. Finally, type of technology use was collected by software type and activity type also allowing for multiple item selection by use of checkboxes, as shown in Appendix J. Although many software variables were selected, these variables were condensed into categories. The categorical variable has seven levels and include the categories: (a) typing (comprised of internet use, e-mail, messaging, writing, and presentations); (b) development (comprised of coding, web design, and use of spreadsheets); (c) learning games (comprised of mathematics games and reading games); (d) all other games; (e) entertainment (comprised of music, artwork, movies); (f) reading (e.g. eBooks); and (g) social networking. Appendix J outlines the specific variables collected.

Procedures

School participant selection. NWEA™ made the decision to involve schools in the *NWEA™ Partners in Innovation Program* which was established during the 2014-2015 school year. The *NWEA™ Partners in Innovation Program* is funded through the Advanced Research and Development team within the Research Department at NWEA™ located in Portland, Oregon. The purpose of the *NWEA™ Partners in Innovation Program* is to involve K-12 schools (public or private) in innovative research studies and new product trials. The program creates a partnership between schools and the NWEA™ research team to try out new assessment approaches, proprietary technology, and new educational research; therefore, providing crucial insights to the NWEA™ research team in order to make improvements to products, approaches, and further the mission of the organization (NWEA, 2015c).

The program began recruiting internally within the NWEA™ organization used basic internal communication to highlight key information about the program to NWEA™ staff. The internal communication was used to target NWEA™ staff members who may work closely with a school or district that the staff member knows would be interested. Additionally, many staff members are former district employees who may know of a specific school or district that would benefit from and be interested in the program. After the internal distribution, the Advanced Research and Development team worked with the NWEA™ Marketing department to create a professional flyer to display on the NWEA™ website (see Appendix H). Lastly, the Advanced Research and Development team specifically engaged with Account Managers and relied on their expertise and knowledge to provide recommendations and referrals. Once a school was identified and both parties (NWEA™ and the school or district) agreed in the partnership program, the school or district administrator signed the legal document for the *NWEA™ Partners in Innovation Program* (see Appendix I).

As of the current school year (2016-17), there are five schools participating in the *NWEA™ Partners in Innovation Program*. The schools represent five states and span all grade levels K-12, see Table 2.4. Three of the lab schools are private (religious affiliated) schools, and two schools are a public (one rural and one suburban). School data are displayed in Table 2.4. There is minimal racial and ethnic diversity among the participating schools. A majority of participants were White; non-White participants included 17.8% of the sample which consists of 6.2% two or more races, 4.9% Hispanic/Latino, 2.9% Asian Pacific Islander, 2.4% other, 0.9% Black, and 0.4% Native American.

Table 2.4

Participating Schools in the *NWEA™ Partners in Innovation Program*

State	Type	Number of students	Grades
Florida	Private; religious	291	K-8
Illinois*	Public; suburban	500	3-5
North Carolina	Private; religious	683	K-8
Oregon	Public; rural	564	K-12
Washington*	Private; religious	240	K-8

Note. An asterisk indicates the school is a current partner and user of NWEA™ products.

Cooperation. Schools that agreed to participate in the *NWEA™ Partners in Innovation Program* received many benefits and communications prior to any involvement. Upon contract signature, schools receive half of the annual stipend awarded as a participating school in the program. The annual stipend ranges anywhere from $3,500-$10,000 depending on size, school type, student demographics, and location. Schools receive half of the stipend upon contract start date (September) and half of the stipend at the contract end date (May). Once schools have been identified to participate in a research study, NWEA™ researchers make contact with both the school or district administrator (often the principal) as well as the NWEA™ account manager. This initial contact explains the research study, estimated completion dates, and begins to discuss possible study dates. Once dates for participation have been established, the school receives and distributes information and consent forms for all parties involved including the teachers, parents, and students. Consent forms for the current study are in Appendix K (e.g., parent consent form and student assent form).

As the study start date neared, the school received an information sheet outlining the specifics of the study for both the administrators and the teachers or other staff involved. Information sheets for the current study are in Appendix L (e.g., information documents for school administrators and information documents for school staff). These forms outline the purposes of the study as well as the roles and expectations of the school, location of study, students, and staff. The purpose of the forms is to provide a basic overview of the study prior to researchers' arrival to the school. Lastly, researchers make every attempt to enter the school prior to the study to review space, technology hardware/software, and other study logistics. This is another opportunity to discuss the study purpose and needs with the administrators and participating staff.

Participating schools are considered partners in these research efforts. A school representative may be invited to participate in any publications or conference presentations that develop as a result of the study. For this particular study, participating schools were invited to present at an assessment conference with the NWEA™ researchers; school leaders and representatives were eager to collaborate in these efforts.

Data collection. At the start of the study, students completed demographic information via the introductory screen. Students were then assigned a unique ID number and were randomly assigned to one of six groups (see Figure 2.1). The grouping variable is unknown to the student. Once randomly assigned to a treatment group, students completed the three performance task conditions in a strategic order, see Figure 2.1. The order of the conditions was dependent upon which group they were assigned. After each treatment condition (i.e., A, B, and C), students received a cognitive engagement survey (CE-S-DSP & SOS) consisting of an adapted form of the CES (Miller et al., 1996) and the SOS (Sundre, 1999), see Appendices A-B. At the

conclusion of the study, students completed one final survey that was comprised of technology use at home including: amount of technology use at home (in hours per day), modality, and type of technology use (i.e., software), see Appendix J. Finally, a purposive subset of students (N = 7) using maximum variation sampling was selected at the conclusion of the study based on variation on dimensions of interest (Patton, 2001) in order to gain additional information about the features of the technology-enhanced mathematics performance instrument (see interview protocol located in Appendix G). Purposive elements measured a range of affective and performance traits within the subset of students participating in the interview.

Conditions. The study implemented a single group counterbalanced design with three conditions: (1) paper-and-pencil, (2) technology-enabled, and (3) technology-enhanced. The paper-and-pencil assessment includes the performance task in a traditional paper test format. Performance tasks specifications across all three modes are outlined in Appendix C. The study implemented all three assessments per respondent with each round including one of the three 20-minute (60-minute total) mathematics performance instruments (see Appendix D). Each respondent completed the student engagement survey (CE-S-DSP & SOS) at the completion of each assessment modality.

The performance instrument used the same framework across each mode but the questions differed. Due to possible equivalency issues between three different item sum values, the same value (24) was used for all three modes but shapes, ordering, and configurations were changed (see examples in Appendices D-E). Equivalency issues that may arise due to harder mathematics (e.g. dividing 96 by 2 or 4 is substantially harder than dividing 24 by 2 or 4) as well as additional values (e.g. 36, 96) did not result in additional ways to create a balanced set-up; therefore, the value of 24 was held constant as part of the task across all three modes. Using the

number 48 could have provided an equivalence; however, with three assessment items it would be hard to discern what effect the value of 48 had versus the effect of 24. Using the same value (24) is an attempt to avoid construct irrelevant variance using different values for each modality. This decision was verified with subject-matter experts[12].

Finally, the adapted survey from the CES (Miller et al., 1996), CE-S-DSP, as well as the SOS (Sundre, 1999) (CE-S-DSP & SOS) were administered after each treatment condition (see Appendices A-B). This survey measured cognitive engagement after students complete each performance task instrument. The survey remained the same at each survey event for appropriate comparison purposes.

Due to the design format, the student gained techniques and skills to solve the problems as they progressed through the three performance instrument modes. The student may have also experienced changes in cognitive engagement involved with engaging in a sequence of tasks. The design attempts to offer control and mitigate these concerns by utilizing the counterbalanced approach. Additionally, analysis included the first task and survey completed for each student. Because of the counterbalanced design, the first task that was completed included similar numbers of students per task (technology-enhanced n = 152; technology-enabled task n = 150; paper-and-pencil task n = 148).

A purposive sample of approximately seven students were interviewed at the conclusion of the study to gain additional information about the features of the technology-enhanced mathematics performance instrument (see Appendix G). The purposive sample of students was selected from the participating Washington school. Purposive elements involved students who demonstrated a range of affective and performance traits. As a result, students were selected to participate based on sex and time of completion of the performance tasks (early-finisher, mid-

[12] Subject-matter experts included NWEA™ senior mathematical content specialists.

finisher, and late-finisher) to select students who may differ on test taking strategies and mathematics abilities using time to completion as a proxy. Interviews were recorded and transcribed in preparation for analysis.

Data Analysis

The study includes two main research questions with each question including multiple parts, as discussed in Chapter I. The following section describes the goal of each research question and the analysis that was conducted to answer each question. Research Question One investigates the performance of the CE-S-DSP & SOS using quantitative data analysis. Research Question Two investigates both quantitative and qualitative relationships between affective measure outcomes and the use of mode types, specifically when evaluating by sex and race/ethnicity. Additional analysis evaluated the effect on cognitive engagement of time spent on home technology use as well as what type of technology use.

Research Question One (RQ1) analysis. RQ1 investigates performance of the instrumentation used. The primary hypothesis for Research Question One is that the CE-S-DSP & SOS is performing as anticipated. Subcomponents of Research Question One investigated the variance, internal consistency, and correlations within the measures. Additionally, to address RQ1, qualitative analysis was employed to investigate the performance of the mathematics performance task from a small, purposive subset of students as well as considered the validity of the qualitative research plan, data collection and analysis.

Variance for affective measures. The first part of Research Question One evaluates the variance of the CE-S-DSP & SOS within the study context. To consider the hypothesized factors of cognitive engagement within the context of this study, RQ1 utilized a confirmatory factor analysis (CFA). Due to item-level categorical data and latent variable summative outcomes, all

models were estimated using weighted least squares means and variance adjusted (WLSMV) estimation. Results from the CFA determined if the items load on the specified subscales of cognitive engagement (shallow processing, deep processing, persistence, importance, and effort), as measured by the CE-S-DSP & SOS.

The CFA was chosen over the principle component analysis (PCA) because the hypothesized factors of cognitive engagement to be employed were already determined based on previous versions of the measures. CFA allows for the analysis to be driven by theoretical relationships among the latent variables (Schreiber, Nora, Stage, Barlow, & King, 2006) of cognitive engagement, as determined by theory about number of factors.

The CFA here (see Figures 2.2 and 2.3) employed the construct of cognitive engagement using five second order constructs and 21 categorical items. The first CFA analyzed the five latent constructs (e.g., deep processing, shallow processing, persistence, importance, effort) across the 21 items, see Figure 2.2. The second CFA analyzed the construct of cognitive engagement using a second order factor analysis (see Figure 2.3). Results from the CFA determined factor loadings and error using WLSMV estimation. Tests included goodness of fit statistics, item loading significance, error variance and covariance significance, and appropriateness of model constraints or additions via tests for model fit changes (Templin, 2011). Acceptable goodness of fit statistics used for evaluation of model fit are displayed in Table 2.5.

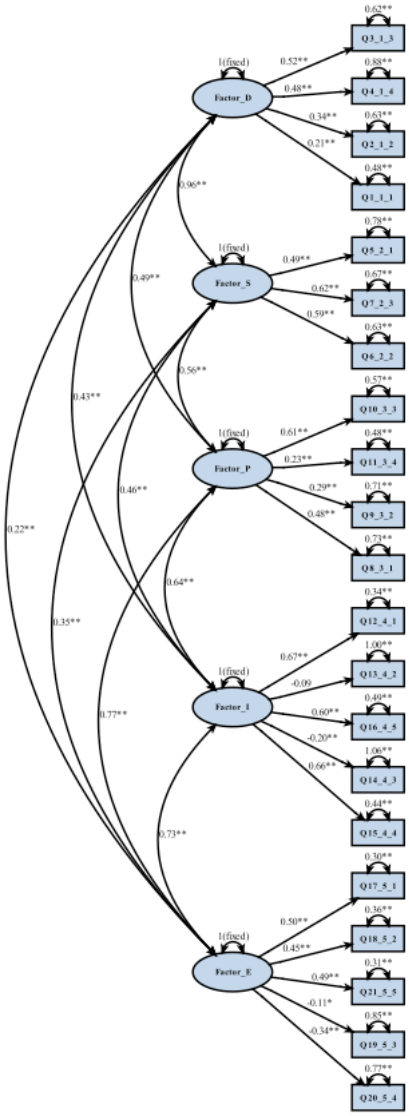


Figure 2.2. Confirmatory factor analysis for the combined CE-S-DSP & SOS (adapted from Miller et al., 1996; Smiley & Anderson, 2011; and Sundre, 1999).

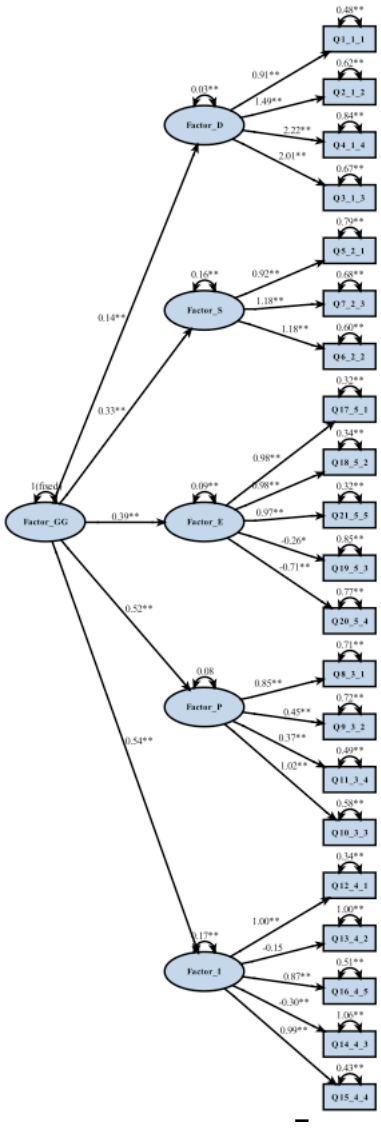


Figure 2.3. Full model - second order confirmatory factor analysis for CE-S-DSP & SOS (adapted from Miller et al., 1996; Smiley & Anderson, 2011; and Sundre, 1999).

Table 2.5

Model Fit Statistics Acceptable Thresholds Used

Fit index	Threshold
χ^2	Compare to other chi-square values in model
p	One-Tailed *t*-test > .10
CFI	> .95 will be considered adequate fit
χ^2 / df	< 2.0 will be considered adequate fit
RMSEA	Upper > .10 will be considered poor fit Lower < .03 will be considered excellent fit
LO 90	close to 0 to accept the close fit hypothesis
HI 90	$\leq$ .10 to reject the poor fit hypothesis
Interaction / *p* value	One-tailed *t*-test > .10 (look at interaction effects; not main effects)
SRMR	$\leq$ .05 indicates good fit

Internal consistency of affective measures. RQ 1 also evaluates the internal consistency of the cognitive engagement measure using an item reliability analysis using SPSS Version 21 (IBM Corp., 2015). Measuring reliability in this way can help ensure that the cognitive engagement measure (CE-S-DSP & SOS) is consistently reflecting the construct it intends to measure (Field, 2013). To measure reliability, the cognitive engagement measure (CE-S-DSP & SOS) was analyzed for internal consistency using Cronbach's Alpha (α).

Research Question Two (RQ 2) analysis. RQ 2 investigates where there is a significant difference between the technology-enhanced performance task modality and the other paper-and-pencil type and technology-enabled mode type (categorical), first overall and then when looking at student cognitive engagement outcome score (continuous) when disaggregating by sex and race/ethnicity. The relationship between the dependent variable of student cognitive engagement and the independent variable of modality type was tested using a one-way ANOVA. Additional evaluation of the relationship between the dependent variable of student cognitive engagement and the independent variable of time spent on technology use at home was evaluated using a one-way ANOVA. Technology use at home is treated categorically in hours spent per day with 5 levels: (a) 0 hours; (b) less than 1 hour; (c) 1-3 hours; (d) 3-5 hours; and (e) more than 5 hours. Lastly, the evaluation of the effect of type of technology use at home on student cognitive engagement was evaluated

One-way ANOVA – modality type on cognitive engagement. First, between-subjects main effects were analyzed to evaluate modality on student cognitive engagement ($Cognitive\ Engagement_i = b_0 + b_1 mode_i + \epsilon_i$).

One-way ANOVA – time spent on technology at home on cognitive engagement. Additional between-subjects main effects were analyzed using a one-way AOV to evaluate the effect of time spent on technology at home (in hours per day) on cognitive engagement ($Cognitive\ Engagement_i = b_0 + b_1 home\ technology_i + \epsilon_i$).

Qualitative Analysis. Qualitative analysis of in-depth interview data, using Harris Cooper methodology, was used to expand on the quantitative analyses conducted to analyze affective outcomes by modality. Methods in the Harris Cooper approach are premised on systematic guidelines for evaluating the validity of synthesis outcomes. In this methodology, concepts

offered to explain a particular phenomenon are collected together and compared in breadth, internal consistency, and the nature of their predictions, from a selected set of materials, which in this case was student interviews. The methodology specifies clear approaches to problem formulation, data collection, data validation stage, data synthesis, and presentation of results, involving five stages of work: (a) establishing the research question or questions that clearly define the scope of the project for problem formulation (described above); (b) utilizing basic tenets of sound data gathering to produce a sufficiently comprehensive integration of relevant materials in a data collection stage (a set of approximately seven students were purposively sampled for interview data collection, see interview questions in Appendix G, for which a semi-structured interviewing protocol was developed, to allow the introduction of all questions to all students, but also the addition of probes to follow-up on questions); (c) implementing a data validation stage, where clear methodology is used to assess and compare the quality of evidence in the materials (see sections below that describe instruments, variables and elements to be collected, validity issues, and some analytic techniques); (d) employing an analysis and interpretation (or synthesis) stage, where the carefully scoped, collected, and evaluated data is triangulated through synthesis techniques as appropriate to the body of work examined (see sections below that describe instruments, variables and elements to be collected, validity issues, and some analytic techniques); and (e) disseminating results designed to share the synthesized research endeavor with policy makers and educators (dissertation manuscript and resulting papers and presentations).

Additional qualitative data displays were employed for data reduction (Miles et al., 2013) and narrative elements of the interview passages were used to illustrate patterns identified during the synthesis of the interviews. Data were coded via NVivo (QSR International, 2014), and

analyzed for networks/relationships between codes. Following this, data were summarized with data reduction into informative patterns and trends, and data displays and other descriptive elements

Qualitative Coding Process

Codes were generated based on the theory of cognitive engagement (e.g. students who express favor or enjoyment towards a specific mode may be more engaged with that specific modality). Initially, codes were created to capture three main components. First, the overall code of *engagement* was used to highlight quotes that may demonstrate enjoyment, excitement, or favor including identification of a favorite task, enjoyment of the ability to write, and favor toward the movement, while the overall code of *disengagement* was used to highlight quotes that may demonstrate frustration, least favor, or confusion. Transcripts were also evaluated as *context-dependent* signifying a mention of a specific component that may infer either engagement or disengagement, depending on the context of the discussion. Within the context dependent codes, transcripts were also coded for *features of task* which identify specific components of the tasks that were mentioned by the students, see Table 2.6. The *features of task* that were identified are more specific components beyond what is mentioned within the engagement or disengagement codes. The initial codes (Table 2.6) did not include platform specific codes in an attempt to capture platform favorability organically through engagement and disengagement discussions.

Table 2.6

Initial Codes and Subcategories for Qualitative Interviews (N = 7)

Code	Sub categories of code
Engagement	Enjoyment (words of enjoyment, excitement, or favor around using a specific task)
	Favorite (identified as a favorite task)
	Writing (being able to take notes / write information down)
	Movement (mentioned, favorably, the movement of the mobile)
Disengagement	Frustration (words of frustration around using a specific task)
	Confusion (words of confusion around using a specific task)
	Least Favorite (identified as a least favorite task)
Context-dependent	Order (mentioning of task order)
	Similarities (mentioning of task similarities)
	Speed (mentioning of speed to completion
	Features (note features of the task such as theme, avatar, look, drag and drop, set responses, brightness of screen)

CHAPTER III

RESULTS

This chapter presents results of the study described in Chapters I and II, beginning with information about the student participants and a review of the descriptive statistics and student demographics. The chapter then presents results of quantitative and qualitative statistical analyses.

Demographic Data

The sample included 450 students who completed at least one mathematics performance task and cognitive engagement survey. Demographic data for the sample are displayed in Table 3.1.

As shown in Table 3.1, the 450 participants for this study were drawn from grades 6-8 in one public school in Oregon (n = 122), one K-8 private school in Washington (n = 74), and one K-8 private school in North Carolina (n = 254). As described previously, the sample was selected based on availability through the NWEA™ overarching project through which the modality investigation was deployed. The sample consisted of 129 sixth graders, 153 seventh graders, and 168 eighth graders. Students who completed less than the first full mathematics performance task were removed from the dataset using listwise deletion, and are not included in this sample and the subsequent analyses.

At 82.2%, the majority of the sample participants were White. The non-White participants constituted 17.8% of the sample, with 4.9% Hispanic/Latino, 2.9% Asian Pacific Islander, 2.4% other, 0.9% Black, 0.4% Native American, and about 6.2% of two or more races. The sample included 46.0% female, 50.4% male, and 3.6% unidentified.

The analyses throughout the study utilized the first performance task that students

completed. For this, 152 students completed the technology-enhanced performance task for their first performance task, 150 students completed the technology-enabled performance task first, and 148 students completed the paper-and-pencil performance task first. Students were assigned to conditions based on a counter-balanced design (see Figure 2.1).

Table 3.1

Demographic Data n(%)

Variable	Oregon	Washington	North Carolina	Total
Female	55(45.1)	34(45.9)	123(48.4)	207(46.0)
Male	63(51.6)	39(52.7)	125(49.2)	227(50.4)
Unidentified	4(3.3)	1(1.4)	6(2.4)	16(3.6)
White	99(81.1)	52(70.3)	219(86.2)	370(82.2)
Non-White	23(18.9)	22(29.7)	35(13.8)	80(17.8)

Research Question One

RQ1 evaluates the correlations, internal consistency, and variance within the measures. Additionally, RQ1 used qualitative analysis to measure the performance of the of the mathematics performance task from the purposive subset by evaluating MAP® scores, performance task scores, and interview data.

I begin RQ1 with descriptive statistics including means for the cognitive engagement measure, standard deviations, skew, and kurtosis estimates, as shown in Table 3.2. Descriptive statistics throughout were generated through SPSS 21 (IBM Corp., 2015).

Table 3.2

n-sizes, Means for Cognitive Engagement, Standard Deviations, Skew, and Kurtosis (*N* = 450)

Sample	*n*	*M(SE)*	*SD*	*s*	*k*
Overall sample	450	57.57(0.42)	8.84	-0.57	0.92
Sex					
Female	212	58.39(0.57)	8.25	-0.34	0.61
Male	227	56.80(0.62)	9.41	-0.67	0.87
Other	11	57.73(1.92)	6.36	0.75	0.29
Race/ethnicity					
White	370	57.32(0.44)	8.55	-0.46	0.61
Black	4	55.25(2.02)	4.03	1.50	2.01
Hispanic/Latino	22	57.78(2.33)	10.95	-0.70	0.30
Asian Pacific Islander	13	61.85(1.90)	6.87	1.04	1.10
Native American	2	62.50(5.50)	7.78	•	•
Two or more races	28	58.29(1.72)	9.10	-0.61	-0.12
Other	11	58.55(4.60)	15.25	-1.54	3.08

Note. • indicates no measure given due to small sample size

Mean scores for engagement in each modality show in Table 3.3. Please note that inferential comparisons among the modalities will be discussed in RQ 2, see next section; RQ1 continues with examination of the instrumentation itself and characteristics of the data generated.

Regarding instrument performance, cognitive engagement scores for the technology-enhanced task (M = 57.26, SD = 9.34), the technology-enabled task (M = 57.31, SD = 0.71), and the paper-and-pencil task (M = 58.16, SD = 8.44) met assumptions of normality (i.e., skew < |2.0| and kurtosis < |9.0|; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010), see Table 3.3. Assumptions of normality based on total score of engagement on the three performance tasks were analyzed using the Shapiro-Wilk test. Results of the test suggest the data are statistically different from a normal distribution (Field, 2013) when grouped by total engagement (p < .05).

Additionally, Q-Q plots of each task demonstrate slight left skewed data across all three tasks, see Figure 3.1. Data were visually inspected in a histogram for normality and variance, see Figure 3.2. Histograms for all three tasks also show a slight left skew. Using Levene's test to measure homoscedasticity, we can conclude that variances among the three measures are not significantly different $F(2, 447) = 0.59$, $p = .56$; therefore, meeting the homogeneity of variance assumption.

Table 3.3

Descriptive Statistics of Modality ($N = 450$)

Modality	*n*	*M(SE)*	*SD*	Skew	Kurtosis	Shapiro-Wilk (*df*)
Technology-enhanced	152	57.26(0.76)	9.34	-0.53	0.73	.973(152)*
Technology-enabled	150	57.31(0.71)	8.73	-0.49	0.53	.980(150)*
Paper-and-pencil	148	58.16(0.69)	8.44	-0.69	1.78	.973(148)*

* $p < .05$

Note. df = degrees of freedom

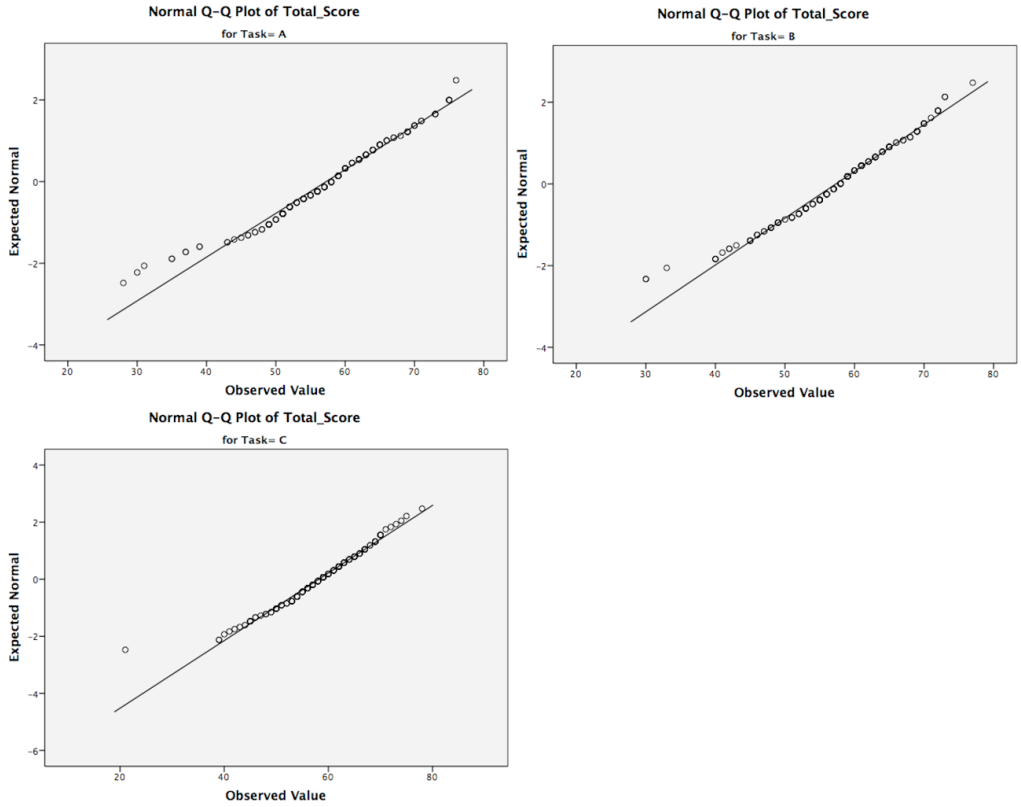


Figure 3.1. Q-Q plots of performance tasks. Task A = technology-enhanced; Task B = technology-enabled; Task C = paper-and-pencil.

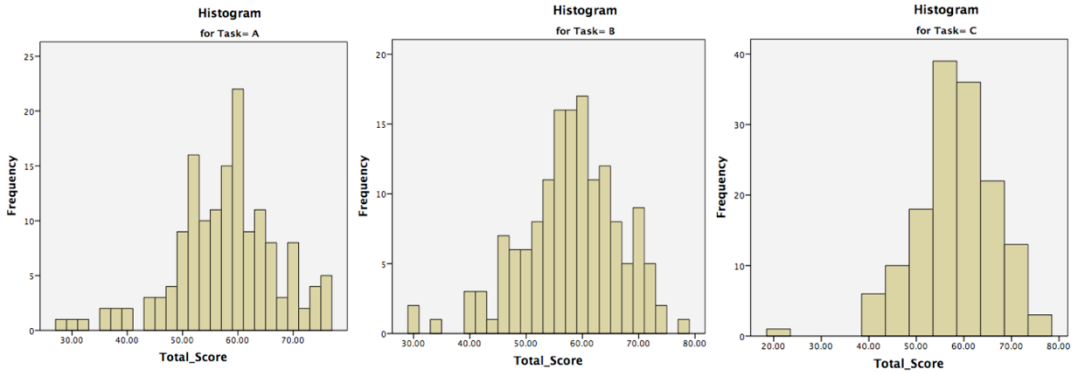


Figure 3.2. Histogram of performance tasks. Task A = technology-enhanced; Task B = technology-enabled; Task C = paper-and-pencil.

To continue with the instrument investigations in RQ1, home technology variables were used to determine amount of technology use at home per day, modality, and type of technology use occurring at home. Because not every student was able to complete the three performance tasks in the allotted class time, not every student reached the usage survey questions. A total of 309 students completed the technology use at home survey. Descriptive statistics for amount of technology use and type of technology use occurring at home are displayed in Table 3.4. Frequencies of how many students reported multiple programs under each category are reported in Table 3.5. Table 3.6 shows the descriptive statistics for amount of time spend on technology at home. Modality use (e.g. laptop computer, desktop computer, tablet, smartphone) also was reported for each student ($N = 309$) across each type of mode.

Amount of time spent on technology at home was a categorical variable measured in hours per day with five levels: (a) none, (b) less than one hour, (c) between one and three hours, (d) between three and five hours, and (e) more than five hours. Modality type was collected and aggregated by device (e.g., desktop computer, laptop computer, tablet, smartphone) allowing for multiple item selection by use of checkboxes, with additional details available in Appendix J. Finally, type of technology use was collected by software type, also as shown in Appendix J. Although many software variables were selected, the variables were condensed into categories. The categorical variable has seven levels and include the categories: (a) typing (comprised of internet use, e-mail, messaging, writing, and presentations); (b) development (comprised of coding, web design, and use of spreadsheets); (c) learning games (comprised of mathematics games and reading games); (d) all other games; (e) entertainment (comprised of music, artwork, movies); (f) reading (e.g. eBooks); and (g) social networking.

The type of technology use at home was further categorized hierarchically as either educationally-based (main categories include typing, development, learning games, reading), n = 1,464, versus entertainment-based (main categories include social networking, entertainment, all other games), n = 1,377. Codes were further split into the two aforementioned categories to examine whether more educationally-focused programs and/or entertainment-focused programs impact student cognitive engagement, see Table 3.7. Despite that only 4 students reported spending no time on a computer at home, 170 students did not report specific program usage, as seen in Table 3.7.

Table 3.4

Descriptive Statistics of Technology Use at Home ($N = 309$)

Software program	*n*
Typing	293
Internet Use	231
E-Mail	191
Messaging	238
Writing	59
Presentations	117
Development	67
Coding / web design	43
Spreadsheets	36
Learning games	59
Mathematic games	53
Reading games	25
Entertainment	277
Music	235
Artwork	100
Movies	253
Reading	52
Social networking	189
All other games	186

Note. Due to the fact that not every student was able to complete the end of user survey, technology use at home variables are only included for a portion of the sample ($N = 309$). All students who participated in the survey ($N = 309$) reported use of each modality use at home (e.g. desktop computer, laptop computer, tablet, smartphone).

Table 3.5

Frequencies of Technology Use at Home Categories, N = 309

Software program	One program	Two programs	Three programs	Four programs	Five programs
Typing	53	54	95	65	26
Development	55	12	-	-	-
Learning games	40	19	-	-	-
Entertainment	56	131	90	-	-
Reading	52	-	-	-	-
Social networking	189	-	-	-	-
All other games	186	-	-	-	-

Table 3.6

Descriptive Statistics for Amount of Time Spent on Technology at Home (in Hours per Day), N = 309

Time	*n*	% of sample
None	4	0.9
Less than 1 hour	44	9.7
Between 1-3 hours	74	16.4
Between 3-5 hours	163	36.1
More than 5 hours	24	5.3

Table 3.7

Frequencies of Educational Computer Use, Non-Educational, Computer Use, Both, and None, N = 452.

Home computer use	*n*
Educational use	294
Non-educational use	292
Both	282
None	170

CFA assumptions. In order to run a structural equation model (e.g. CFA), the expectation is that the data have met certain assumptions. To begin, there must be temporal precedence, that is, the cause must have occurred before the effect (Hoyle, 2011). This is also an assumption that can be used to determine a causal effect of a structural equation model. In the current study, this assumption was met through the theoretical supposition that the performance task occurred prior to the measure of engagement.

The second assumption is that dimensionality is assumed known (ICPSR, 2011). This assumption is met in the current study because the model dimensions are based on previous measures and have a theoretical foundation. Therefore, latent traits were categorized and dimensionality was assumed known.

The third assumption is that the data are continuous and a linear model would be appropriate (citation). This if often questionable for Likert scale data (ICPSR, 2011); however, for the current study, the Likert data were treated as continuous to arrive at a total score for the measure.

The fourth assumption is that the covariance between items will be predicted and, therefore, the basis of model fit (ICPSR, 2011). All model fit statistics will be presented and latent factor variances will be discussed later in the current chapter.

The last assumption is that there is not an item intercept present (ICPSR, 2011). The current study does not utilize an intercept, therefore, meeting the final assumption. Further assumptions can be made on the causal relationship of a structural equation model; however, these relationships beyond the scope of the current study.

Correlation coefficients. Pearson's correlation coefficient (*r*) was used as a standardized measure to determine the strength of association or relationship between the variables (Field, 2013), as shown in Table 3.8. The range of values (-1 to +1) determines the direction and strength of the relationship. Variables included sex race/ethnicity, total score on the measure of cognitive engagement (CE-S-DSP & SOS), and scores on each of the five subsections (deep processing, shallow processing, persistence, importance, and effort). As displayed in Table 3.8, correlation coefficients indicated a significant correlation between sex and the total deep processing (DP) score ($p < .05$). Otherwise, correlations between sex with all other variables were not significant. Also, correlations between race/ethnicity and all variables were not significant.

As expected, because the elements of cognitive engagement were expected to be related as discussed in Chapter I, bivariate correlations among the overall engagement measure and the various subcomponents all showed significance, with correlations ranging from small to moderate and high correlations. These are reported here for completeness of the data for RQ1 but relationships are explored in more depth in upcoming sections when the measures and subcomponents are deconstructed into their theoretical components, and investigated with confirmatory factor analysis according to the theoretical model of the instrument.

Table 3.8

Correlations for the Sample

Variable	1	2	3	4	5	6	7	8
1. Male	1.00							
2. Non-White	.05	1.00						
3. Total score	-.08	-.06	1.00					
4. Total DP	-.11*	-.08	.56**	1.00				
5. Total SP	-.06	-.05	.66**	.38**	1.00			
6. Total P	-.01	-.02	.76**	.31**	.25**	1.00		
7. Total I	-.07	-.04	.76**	.21**	.29*	.42**	1.00	
8. Total E	-.04	-.04	.80**	.28**	.23**	.60**	.49**	1.00

Note. The SOS is a measure developed by Sundre (1999) and the CE-S-DSP is adapted from Miller et al., 1996 and Smiley & Anderson, 2011. Total score is comprised of a summative score on the CE-S-DSP & SOS. Total CE-S-DSP is a summative score on the cognitive engagement scale including shallow processing, deep processing, and persistence. Total DP is the summative score on deep processing. Total SP is the summative score on shallow processing. Total P is the summative score on persistence. Total I is the summative score on importance. Total E is the summative score on effort.
* $p < .05$
** $p < .01$

Internal consistency. Reliability analyses were conducted to evaluate the 21 Likert items for internal consistency as a single scale; please note that this does not reflect on the reliability of the subcomponents. Cronbach's alpha was calculated to determine internal item consistence on the 21 Likert items; or, the degree to which the items measure the construct overall of cognitive engagement. Cronbach's alpha creates two sets of items in every possible way and computes the correlation coefficient for each split (Field, 2013).

The reliability based on this data set for the overall instrument is $\alpha = 0.84,$ which is above the preferred threshold of 0.80 and well above the more minimally acceptable threshold of 0.70 (Tavakol & Dennick, 2011). Reliability of sub measures are not used as reliable sub scores

in this study but will be reported in the CFA discussion later in Chapter III.

Variance. The next component of RQ1 evaluates the variance of the CE-S-DSP & SOS within the context of the technology-enhanced mathematics performance instrument for the sample dataset. Variance of student scores on the CE-S-DSP & SOS (see Figure 3.3) is 78.08 with a standard deviation of 8.84 and a mean of 57.57. This indicates that there were fluctuations in students' self-report of cognitive engagement which indicates there is enough variance in the measure. The variance of the CE-S-DSP & SOS was analyzed using a confirmatory factor analysis. The relationship between the purposive subset of data and MAP® scores are analyzed qualitatively.

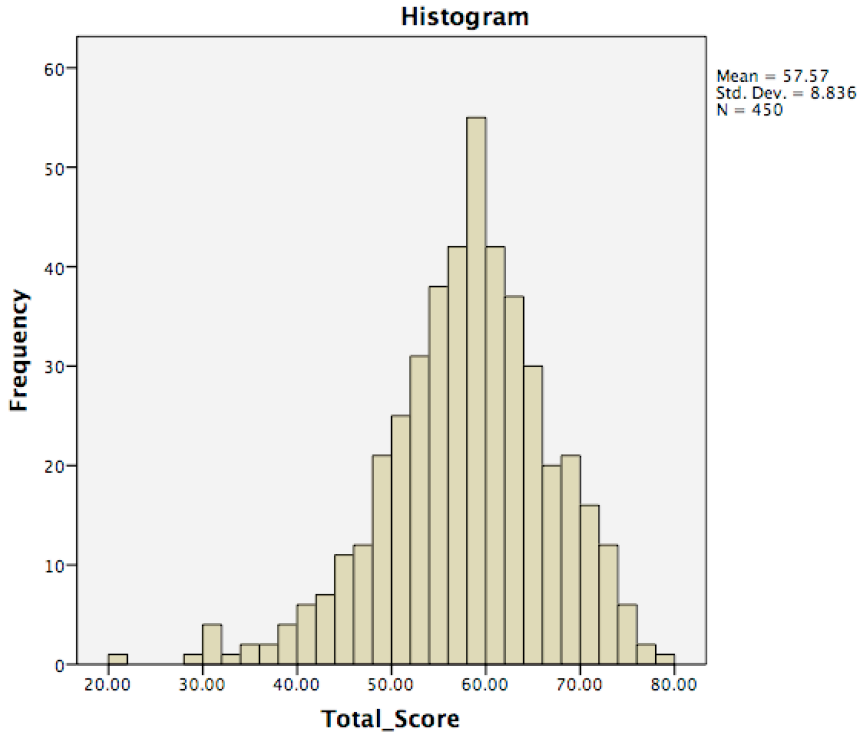


Figure 3.3. Histogram of total score on the CE-S-DSP & SOS.

CFA Results. The SAS (SAS Institute Inc., 2013) PROC CALIS procedure was used to fit the CFA models of the study. Because the CE-S-DSP & SOS measure of cognitive engagement used item-level data, where are categorical and non-normal, all models were estimated using robust weighted least squares estimation. Determinations of model fit were based on the seminal text in this domain (Hu & Bentler, 1999); however, Kline's (Kline, 2013) approach was used in treatment of thresholds discussed by Hu and Bentler.

Although the reporting of sub scores are not discussed in the current study, sub score reliability estimates are displayed in Table 3.9. Three measures (Deep Process, Shallow Processing, Persistence) have reliability estimates below the accepted threshold of $\alpha = 0.70$. Upon further investigation, none of the items for each of the three factors would result in an improved reliability estimate. The additional sub scores of Importance and Effort resulted in reliability estimates above the accepted threshold of $\alpha = 0.70$. The overall reliability estimate for the measure, which is used in the current study, is $\alpha = 0.84$.

Table 3.9

Reliability Estimates for CE-S-DSP & SOS Sub Scales

Sub scale	α
Deep processing	0.56
Shallow processing	0.37
Persistence	0.68
Importance	0.73
Effort	0.81

The first order CFA (see Figure 2.2) specified the CFA model across 5 factors (deep processing, shallow processing, persistence, importance and effort), which included 21 categorical item responses.

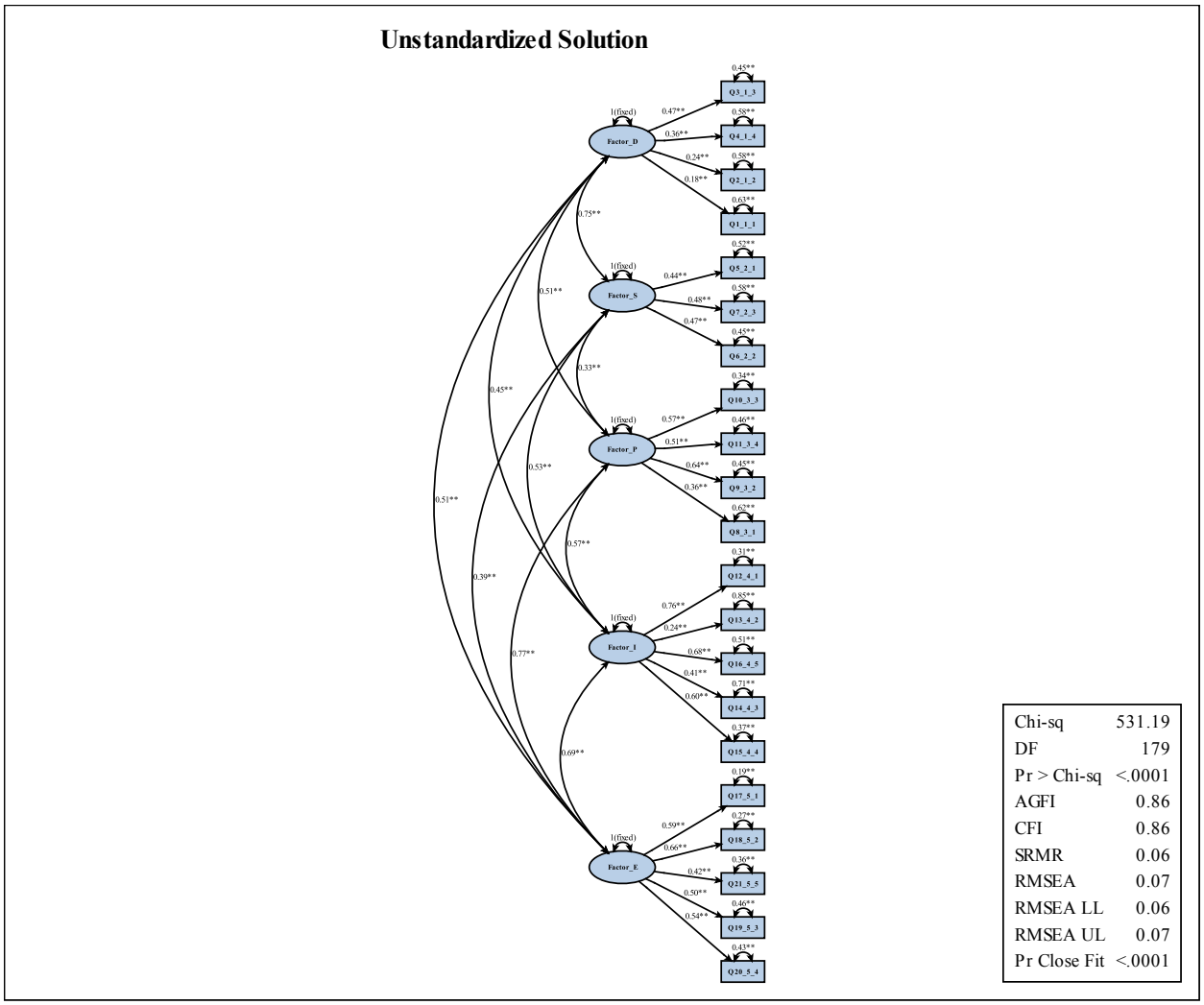


Figure 3.4. Five-factor cognitive engagement confirmatory factor analysis.

As seen in Table 3.10, the model fails the chi-square test, $\chi^2(179) = 531.193, p < .001$. This is expected given the categorical, non-normality of the data and does not necessary mean that the model does not fit the data sufficiently. In order to continue to evaluate model fit, it is important, and often preferred, to evaluate fit based on other fit statistics (Holtzman, 2014).

The Root Mean Square Error of Approximation (RMSEA) of .07 (.06 to .07) is less than 0.08 which indicative of an acceptable model fit (Browne & Cudeck, 1993), rejecting the poor-fit hypothesis. The Comparative Fit Index (CFI) does not exceed the .95 threshold, indicating the model does not fit the data well. Lastly, the Standardized Root Mean Square Residual (SRMR) is .06 which is less than the .08 indicative of good fit (Hu & Bentler, 1999). Given the fact that chi-square can be overly sensitive to sample size and non-normal data and the remaining fit statistics demonstrate reasonable fit across two additional fit measures, it is concluded that the data are not ideal for CFA and show some but not ideal fit to Model 1. So, results of this model will be explored here but should be interpreted cautiously; future work could include the use of latent variable models such as confirmatory IRT for which a data set such as this would be better fitting to the assumptions of the model, including better allowing the employment of categorical and non-normal data, when the correct estimation algorithms are employed within the latent variable setting. However, this extension is outside the scope of this dissertation so will be discussed in Chapter IV on Conclusions and Future Implications.

Table 3.10

Goodness-of-Fit Indices of the Five Factors of Cognitive Engagement (N = 450)

Model	*df*	χ^2	χ^2/df	RMSEA (90% CI)	CFI	SRMR
Five-factor	179	531.19***	2.97	.07 (.06 - .07)	.86	.06

Note. RMSEA= root mean square error of approximation; CFI= comparative fit index; SRMR= standardized root mean square residual.

***$p < .001$

The higher order CFA (see Figure 3.5) specified the CFA model across the second order factor of cognitive engagement and 5 first order factors (deep processing, shallow processing, persistence, importance and effort), including 21 categorical item responses. Figure 3.5 displays the final second order CFA model with factor loadings and unexplained variance that is unaccounted for by the model. As seen in Table 3.11, this model also fails the chi-square test, $\chi^2(179) = 584.75, p < .001$. As with the previous model, a failed chi-square test is expected given the categorical, non-normality of the data and does not necessary mean that the model does not fit the data sufficiently. Additional evaluation of fit statistics indicate reasonable model fit of the Root Mean Square Error of Approximation (RMSEA) of .07 (.06 to .08). The RMSEA is less than 0.08 which is indicative of reasonable model fit (Browne & Cudeck, 1993). As with the first order 5-factor CFA, the Comparative Fit Index (CFI) does not exceed the .95 threshold, indicating the model does not fit the data well; yet, the Standardized Root Mean Square Residual (SRMR) is .07 which is less than the .08 indicative of good fit (Hu & Bentler, 1999). So once again it is concluded that the data are not ideal for CFA and show some but not ideal fit to the theoretical 5-factor model, as explored through CFA. So results of this model will be discussed here later in this chapter and the next, but should be interpreted cautiously; future work could include the use of latent variable models such as confirmatory IRT, as will be discussed in Chapter IV.

Table 3.11

Goodness-of-Fit Indices of the Second Order CFA (N = 450)

Model	*df*	χ^2	χ^2/df	RMSEA (90% CI)	CFI	SRMR
Five-factor	179	584.75***	3.27	.07 (.06 - .08)	.84	.07

Note. RMSEA= root mean square error of approximation; CFI= comparative fit index; SRMR= standardized root mean square residual.

***$p < .001$

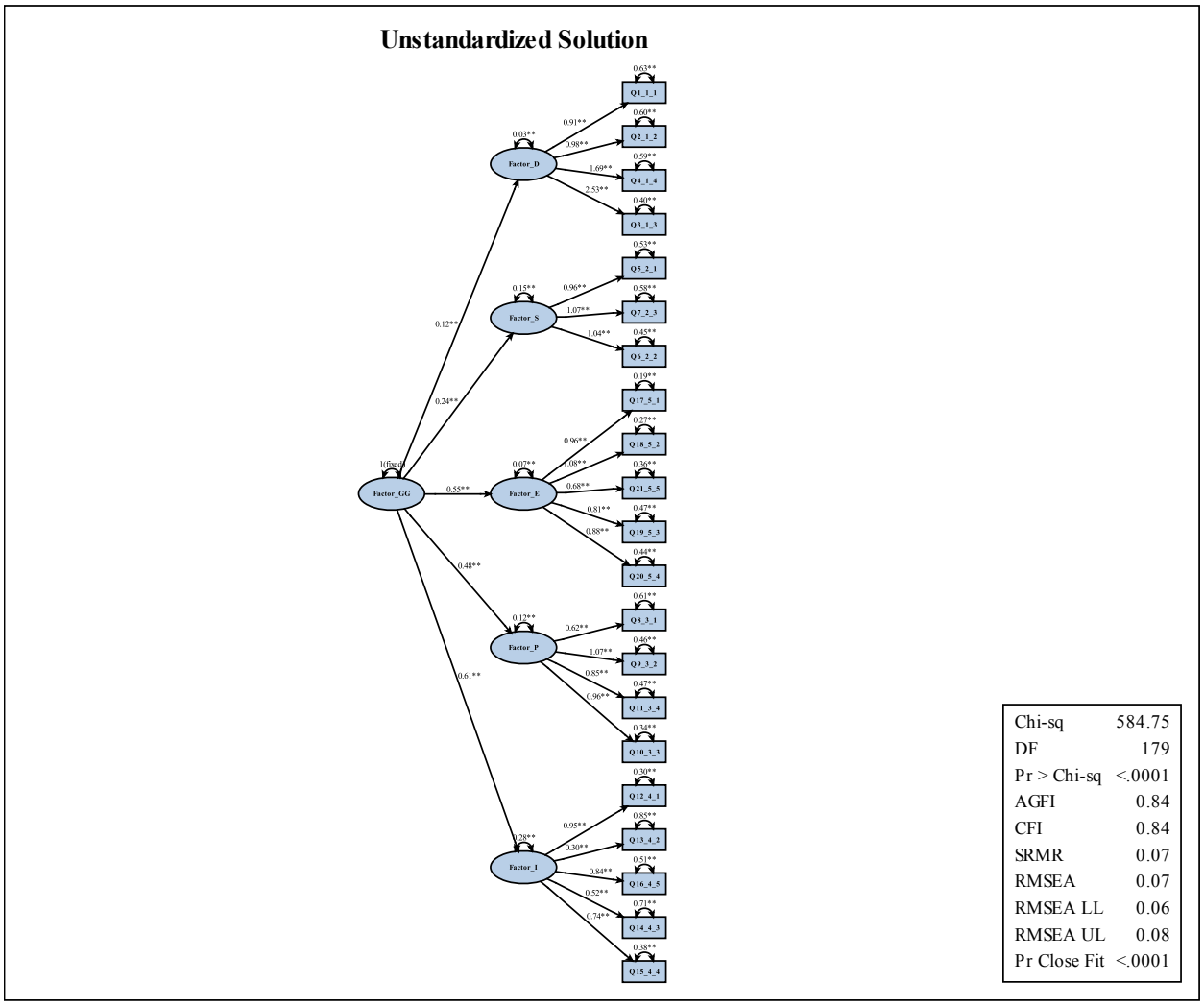


Figure 3.5. Higher order cognitive engagement confirmatory factor analysis.

Qualitative analysis. The following section analyzes the variance through a qualitative approach with a purposive subset of students ($N = 7$). Table 3.12 displays the descriptive statistics for the sub-sample of students who participated in the follow-up interview. The outcomes of the mathematics performance instrument that were scored from the qualitative data ($N = 7$) are reported. Additionally, student scores on the NWEA™ MAP® assessment from the purposive subset were used as a comparison for performance outcomes in mathematics. RQ2 further explored qualitative interview data, themes and pattern analyses.

Table 3.12

Descriptive Statistics for Interview Subsample (N = 7)

ID number	Grade	Sex	Race/ethnicity	First task	Total score
152	8	Male	White	A	46
155	8	Female	Hispanic/Latino	C	70
160	6	Male	White	B	69
169	6	Female	White	A	59
179	6	Female	Asian or Pacific Islander	C	59
180	7	Female	Black	C	61
200	7	Male	White	A	53

Three students completed the performance task with a perfect score (ID 152, ID 160, ID 180), one eighth grader, one seventh grader, and one sixth grader. The other four students completed the task without demonstrating proficiency, see Table 3.13.

Table 3.13

Student Outcome Scores on Paper-and-Pencil Performance Instrument by Item (N = 7)

Student ID	Item 1	Item 2	Item 3	Item 4	Item 5	Total score
152	8(100)	1(100)	7(100)	6(100)	7(100)	29(100)
155	2(25)	0(0)	1(14)	2(33)	2(29)	7(24)
160	8(100)	1(100)	7(100)	6(100)	7(100)	29(100)
169	4(50)	1(100)	3(43)	6(100)	1(14)	15(52)
179	3(38)	0(0)	0(0)	0(0)	0(0)	3(08)
180	8(100)	1(100)	7(100)	6(100)	7(100)	29(100)
200	2(25)	0(0)	1(14)	1(17)	1(14)	5(17)

Note. Points received (percentage received)

MAP® score data for the subsample. The subsample of students who participated in the interview ($N = 7$) also completed the MAP® test in mathematics for fall 2016. Student scores on the MAP® are displayed in Table 3.14 along with outcome scores from the paper-and-pencil performance task completed for this study. In order to compare achievement within MAP®, NWEA™ uses RIT Scale Norms. The Scale Norms (Thum & Hauser, 2015) provides status and growth norms for individual students across MAP® subjects (Reading, Language Usage, Mathematics, and General Science) (NWEA, 2015a). The RIT Scale Norms for grades 6-8 in mathematics are displayed in Table 3.15.

Table 3.14

Student Outcome Scores on Paper-and-Pencil Performance Instrument, Fall 2016 MAP® Test, and CE-S-DSP & SOS Engagement Survey (N = 7)

Student ID	Grade	RIT (*SE*)	PT score	Total engagement score
152	8	246.86(3.08)	100	46
155	8	228.42(2.97)	24	70
160	6	236.81(2.92)	100	69
169	6	222.99(2.95)	52	59
179	6	213.59(2.88)	8	59
180	7	213.10(2.92)	100	61
200	7	206.74(3.10)	17	53

Note. Maximum score of cognitive engagement was 84. The average total score of cognitive engagement among the entire sample (N = 450) is 57.57; the sub-sample is slightly higher than the average score (M = 59.57).

Table 3.15

Fall 2015 NWEA™ Beginning of the Year Norms in Mathematics

Grade	Begin-year mean norm	Begin-year *SD*
6	217.6	15.53
7	222.6	16.59
8	226.3	17.85

Note. Data based on the NWEA™ Measures of Academic Progress Normative Data (NWEA, 2015b)

Relationship between MAP® scores and performance task outcomes for sample of students. The MAP® data for Fall 2016 (Table 3.14) show that all students (N = 7) scored within at least one standard deviation of the mean, with three students scoring within one standard deviation below the mean (IDs 179, 180, 200), two students scoring within one standard deviation above the mean (IDs 155, 169), and two students scoring within two standard

deviations above the mean (IDs 152, 160), based on the NWEA™ RIT Scale Norms (Thum & Hauser, 2015). The paper-and-pencil performance task scores (Table 3.14) were somewhat consistent with students' MAP® outcomes. Four of the student scores demonstrated consistent mathematics performance between the two measures. Two students who scored within two standard deviations above the mean (IDs 152, 160) received a perfect score on the performance task data; therefore, demonstrating consistent mathematics performance between the two measures. Additionally, two of the three students who scored below the mean (IDs 179, 200) received a low score for the performance task (8% and 17%, respectively); therefore, also demonstrating consistent mathematics performance between the two measures.

Despite the performance consistency between four of the students across both measures, there were still three students whose scores did not demonstrate consistency. To begin, two students who scored within on standard deviation above the mean on the MAP® test (IDs 155, 169) received low scores for the mathematics performance instrument (24% and 52%, respectively). This could be a result of low engagement during the performance task; however, total scores of engagement on the CE-S-DSP & SOS measure (Table 3.14) indicates student 155 and 169 both self-reported engagement higher than the mean for the entire sample (M = 57.57; N = 450) with ID 155 self-reporting an engagement score of 70 and ID 169 self-reporting an engagement score of 59.

The final student (ID 180) scored within on standard deviation below the mean on the MAP® test but received a perfect score (100%) on the mathematics performance instrument. This score could indicate that the student demonstrated engagement during the performance instrument (self-reported a higher than average engagement for the performance task) but may have demonstrated a lack of engagement during the MAP® test administration.

Research Question Two

RQ2 first evaluates the relationship of different performance task modalities with engagement then evaluates the relationship when factoring in race/ethnicity. The relationship between the dependent variable of student cognitive engagement and the independent variable of modality type was evaluated using a one-way ANOVA.

The second component of RQ2 evaluated the relationship between home technology use patterns and cognitive engagement. Finally, the final component of RQ2 explores the in-depth qualitative interview data from the purposive sample described above, to further help evaluate and interpret student attitudes towards technology-enhanced assessments.

ANOVA Assumptions. In order to run a one-way ANOVA, the expectation is that the data have met certain assumptions. To begin, the dependent variable in the study is continuous and the independent variable has two or more categorical groups (technology-enhanced, technology-enabled, and paper-and-pencil); therefore, meeting the first two assumptions.

The third assumption is that the observations are independent. Within the current study, the assumption of independence is assumed because the inclusion of one participant is not related to the inclusion of another (Biancarosa, 2015).

The fourth assumption ensures homogeneity of variance. Using Levene's test to measure homoscedasticity, we can conclude that variances between the three measures are not significantly different $F(2, 447) = 0.59$, $p = .56$; therefore, meeting the homogeneity of variance assumption.

The fifth assumption ensures the dependent variable is normally distributed. Normality was analyzed using a Q-Q plot (see Figure 3.1) as well as visually inspected using histograms (see Figure 3.2). The Q-Q plot implies demonstrate slight left skewed data across all three tasks;

the closeness of points on the line is indicative of a normal distribution. Mean scores for cognitive engagement on the technology-enhanced task (M = 57.26, SD = 9.34), the technology-enabled task (M = 57.31, SD = 0.71), and the paper-and-pencil task (M = 58.16, SD = 8.44) met assumptions of normality (i.e., skew < |2.0| and kurtosis < |9.0|; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010), see Table 3.2. The Shapiro-Wilk test further analyzed cognitive engagement scores for each of the tasks, all of which are non-significant suggesting a non-normal distribution: (a) technology-enhanced, p = .90; (b) technology-enhanced, p = .034; and (c) paper-and-pencil, p = .04.

Normality was analyzed for the overall score of engagement using a Q-Q plot (see Figure 3.1) as well as visually inspected with a histogram (see Figure 3.2). The Q-Q plot show a strong left skew across the overall measure of cognitive engagement. The histogram shows a normal distribution with a slight left skew due to outliers (see stem and leaf plot in Figure 3.6). Looking at the data, it is inferred that the extreme outlier (ID 290) is a result of selecting the same response throughout the entire survey (e.g. all of the survey responses were selected as 'strongly disagree', which, in most cases, was a proxy for lack of engagement) and, therefore, should be treated with caution. Assumptions of normality based on total score of engagement on the three performance tasks were analyzed using the Shapiro-Wilk test. Results of the test rejects the null hypothesis; therefore, indicating the data are statistically different from a normal distribution (Field, 2013) when grouped by total engagement (p < .001).

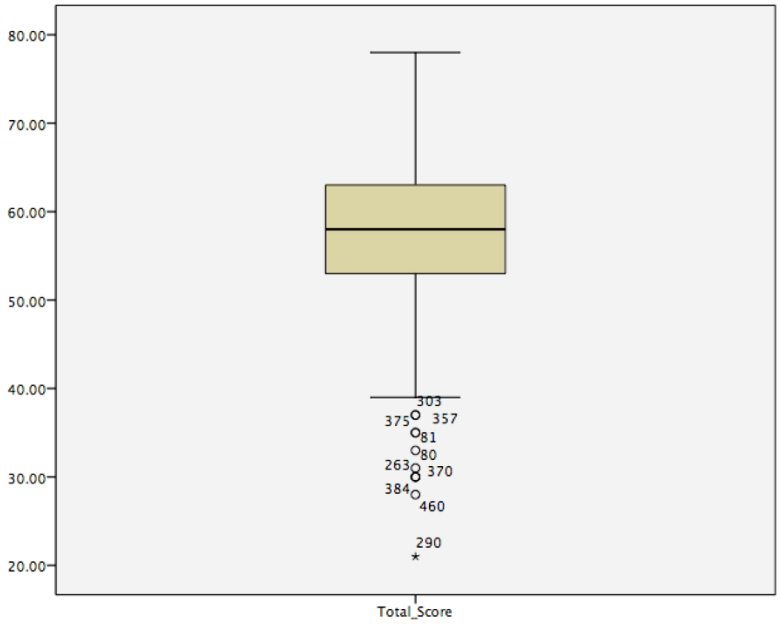


Figure 3.6. Stem and leaf plot of total cognitive engagement score.

Lastly, the assumption that there are no significant outliers was analyzed using boxplots (see Figure 3.7). The boxplots indicate there are eight outliers and one extreme outlier. As mentioned earlier, the extreme outlier should be treated with caution due to the data pattern observed (all survey selections as 'strongly disagree'). It is also worth noting that all of the outliers (when not considering the extreme outlier in the paper-and-pencil modality) appear in the two computer modalities and are all outliers demonstrating low total scores of engagement.

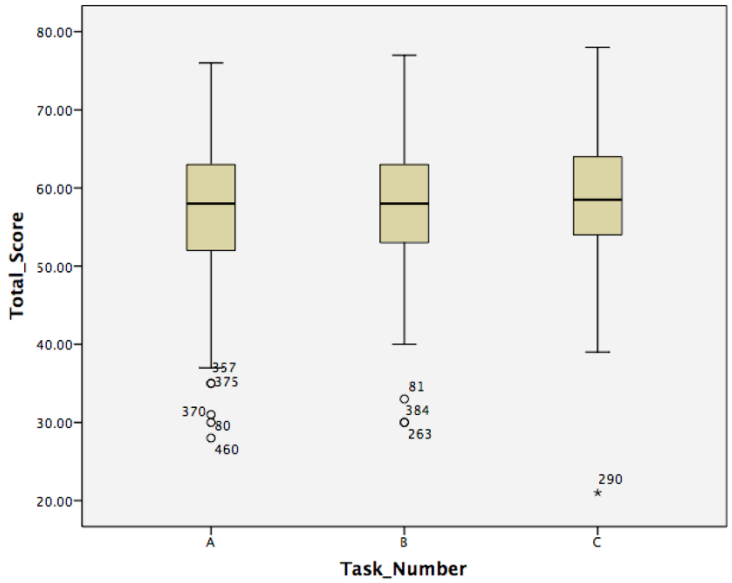


Figure 3.7. Boxplots of total cognitive engagement score across three tasks; A = technology-enhanced, B=technology-enabled, C=paper-and-pencil.

Measures of central tendency across all three platforms are displayed in Table 3.16 and show similar average scores of engagement across all three modes. Again, it is important to note how the extreme outlier in Task C (paper-and-pencil) may impact these scores, specifically the minimum value and range. With the extreme outlier removed, the minimum score for the paper-and-pencil mode (Task C) is 39 with a range of 39. Histograms were used to analyze the shape of each distribution of the independent variable (platform type). The distributions (see Figure 3.8) between the three tasks have the same shape when looking at total score of cognitive engagement.

Table 3.16

Measures of Central Tendency by Platform Type

Task	*M*(*SE*)	*SD*	Median	Min-max	Range
Technology-enhanced	57.26(0.76)	9.34	58.00	28.00-76.00	48.00
Technology-enabled	57.31(0.71)	8.73	58.00	30.00-77.00	47.00
Paper-and-pencil	58.16(0.69)	8.44	58.50	21.00-78.00	57.00

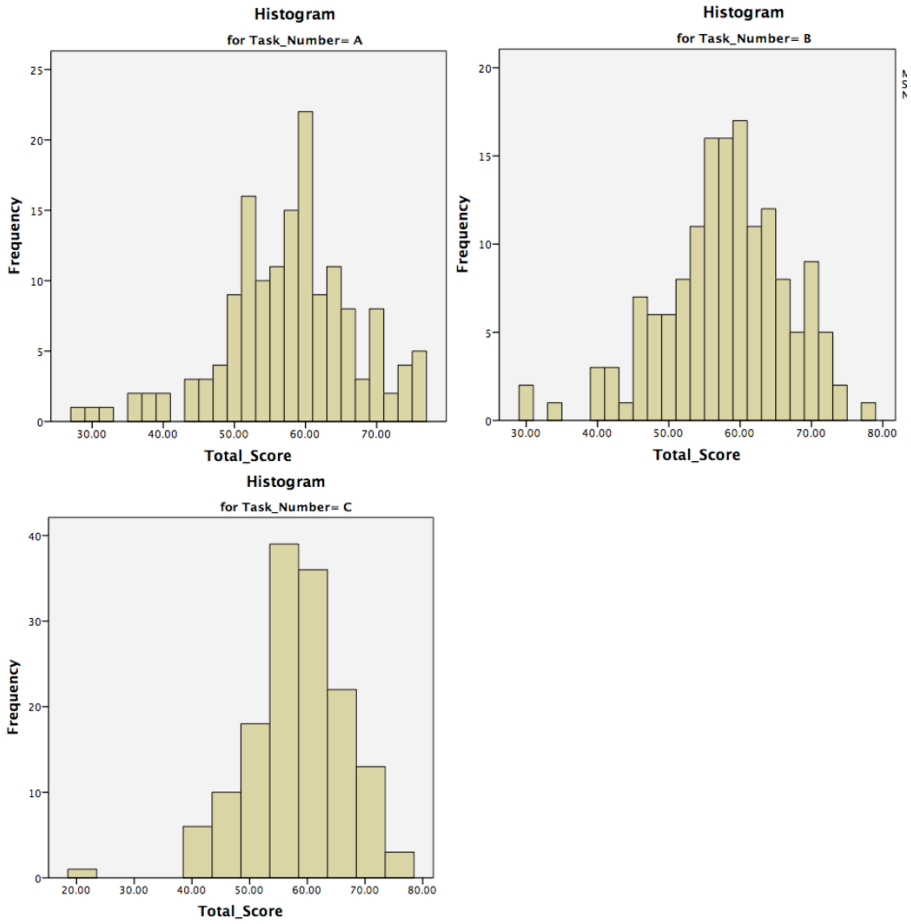


Figure 3.8. Histograms of total cognitive engagement score across three tasks; A = technology-enhanced, B=technology-enabled, C=paper-and-pencil.

The data show a slight left skew across the three tasks. Although this may violate the assumption of normality, Glass, Peckham, and Sanders (Glass, Peckham, & Sanders, 1972) suggest skewed distributions have minimal effect on error rate and power for two-tailed tests due to the robustness of the *F* in an ANOVA. Games and Lucas (Games & Lucas, 1966) also advise that transforming skewed distributions of data help as much as they hinder the accuracy. However, further literature (Levine & Dunlap, 1982) show that data transformations can improve the performance of *F*.

ANOVA Results.

Effect of modality on cognitive engagement. A between-subjects main effects analysis evaluated modality type on student cognitive engagement, see Equation 3.1. Levene's test of equality of error variance was not significantly different $F(2, 447) = 0.47$, $p = .62$; therefore we can conclude that variances between the three measures are not significantly different. Results from the one-way, between-subjects analysis of variance are presented in Table 3.17. The dependent variable was the total score on the cognitive engagement measure (CE-S-DSP & SOS). The independent variable was modality type with three levels: (a) technology-enhanced, (b) technology-enabled, and (c) paper-and-pencil. The main effect of type of modality on cognitive engagement was not significant, $F(2,447) = .48$, $p = .62$, $\eta^2_{partial} = .002$. There was not a significant difference in means of cognitive engagement following the use of the first performance task between self-report of cognitive engagement for students who completed the technology-enhanced task first ($M = 57.26$), students who completed the technology-enabled task first ($M = 57.13$), and students who completed the paper-and-pencil task first ($M = 58.16$).

$$Cognitive\ Engagement_i = b_0 + b_1 mode_i + \epsilon_i \quad (3.1)$$

Table 3.17

One-Way Analysis of Variance Summary Table for the Effect of Modality Type on Cognitive Engagement

Source	*df*	*SS*	*MS*	*F*	*p*
Modality type	2	75.53	37.77	0.48	.62
Error	447	34,982.69	78.26		
Total	450	35,058.22			

An additional between-subjects main effects analysis further evaluated modality type on student cognitive engagement, specifically by evaluating the impact of a computer on cognitive engagement (e.g. technology-enhanced or technology-enabled in comparison to paper-and-pencil), see Equation 3.2. Levene's test of equality of error variance was not significantly different $F(1, 448) = 0.58$, $p = .45$; therefore we can conclude that variances between the measures are not significantly different. Results from the one-way, between-subjects analysis of variance are presented in Table 3.18. The dependent variable was the total score on the cognitive engagement measure (CE-S-DSP & SOS). The independent variable was computer use with two levels: (a) technology-enhanced or technology-enabled, and (b) paper-and-pencil. The main effect of computer use on cognitive engagement was not significant, $F(1,448) = .96$, $p = .33$, $\eta^2_{partial} = .002$. There was not a significant difference between self-report of cognitive engagement for students who completed the performance task using technology ($M = 57.28$) and students who completed the performance task on paper-and-pencil task ($M = 58.16$).

$$Cognitive\ Engagement_i = b_0 + b_1 computer_i + \epsilon_i \quad (3.2)$$

Table 3.18

One-Way Analysis of Variance Summary Table for the Effect of Technology Use on Cognitive Engagement

Source	*df*	*SS*	*MS*	*F*	*p*
Computer	1	75.29	75.29	0.96	.33
Error	448	34,982.94	78.09		
Total	450	1,526,553.00			

Modality use was further investigated by evaluating the two extreme platforms (technology-enhanced in comparison to paper-and-pencil) on cognitive engagement. A between-subjects main effects analysis was run, see Equation 3.3. Levene's test of equality of error variance was not significantly different $F(1, 298) = 0.91$, $p = .34$; therefore we can conclude that variances between the measures are not significantly different. Results from the one-way, between-subjects analysis of variance are presented in Table 3.19. The dependent variable was the total score on the cognitive engagement measure (CE-S-DSP & SOS). The independent variable was platform with two levels: (a) technology-enhanced and (b) paper-and-pencil. The main effect of extreme platform on cognitive engagement was not significant, $F(1,298) = .76$, $p = .38$, $\eta^2_{partial} = .003$. There was not a significant difference between self-report of cognitive engagement for students who completed the performance task using the technology-enhanced platform ($M = 57.26$) and students who completed the performance task on paper-and-pencil task ($M = 58.16$).

$$Cognitive\ Engagement_i = b_0 + b_1 platform_i + \epsilon_i \quad (3.3)$$

Table 3.19

One-Way Analysis of Variance Summary Table for the Effect of Extreme Platform (Technology-Enhanced versus Paper-and-Pencil) on Cognitive Engagement

Source	*df*	*SS*	*MS*	*F*	*p*
Platform	1	60.58	60.58	0.76	.38
Error	298	23,632.42	79.30		
Total	300	1,022,480.00			

Modality use was investigated one more time by evaluating the effect of grade level on cognitive engagement. This was analyzed in order to explain variance that may be due to assessment difficulty level. A between-subjects main effects analysis was run, see Equation 3.4. Levene's test of equality of error variance was not significantly different $F(1, 447) = 2.79$, $p = .06$; therefore we can conclude that variances between the measures are not significantly different. Results from the one-way, between-subjects analysis of variance are presented in Table 3.20. The dependent variable was the total score on the cognitive engagement measure (CE-S-DSP & SOS). The independent variable was grade level with three levels: (a) sixth grade; (b) seventh grade; (c) eighth grade. The main effect of grade level on cognitive engagement was not significant, $F(2,447) = 1.83$, $p = .16$, $\eta^2_{partial} = .008$. There was not a significant difference between self-report of cognitive engagement for students who were in sixth grade ($M = 58.68$), students who were in seventh grade ($M = 57.58$), and students who were in eighth grade ($M = 56.71$).

$$Cognitive\ Engagement_i = b_0 + b_1 grade_i + \epsilon_i \quad (3.4)$$

Table 3.20

One-Way Analysis of Variance Summary Table for the Effect of Grade Level on Cognitive Engagement

Source	*df*	*SS*	*MS*	*F*	*p*
Grade	2	284.32	142.16	1.83	.16
Error	447	34,773.906	77.79		
Total	449	35,058.224			

Effect of type of modality, sex, and race/ethnicity on cognitive engagement. The additional factors of sex and race/ethnicity were added to explore the effect of modality type on cognitive engagement. A three-way, between-subjects analysis of variance was run to explore the effect of race/ethnicity, sex, and modality type on the total score of cognitive engagement. Levene's test of equality of variance was not statistically different $F(15, 434) = 1.27$, $p = .22$. Results from the three-way, between-subjects analysis of variance including the independent variables of race/ethnicity, sex, and modality type on the dependent variable of cognitive engagement are presented in Table 3.21, see Equation 3.5. Results indicated a non-significant main effect of modality type on cognitive engagement, $F(2,434) = .16$, $p = .85$, $\eta^2_{partial} = .001$. There was not a significant difference between self-report of cognitive engagement for students who completed the technology-enhanced task ($M = 58.44$), students who completed the technology-enabled task ($M = 57.59$), and students who completed the paper-and-pencil task ($M = 57.48$). There was also a non-significant main effect of race/ethnicity on cognitive engagement, $F(1,434) = 0.90$, $p = .77$, $\eta^2_{partial} = .000$. Students who were White did not have significantly

different cognitive engagement outcomes (M = 57.20) than students who were non-White (M = 58.74). Lastly, there was also a non-significant main effect of sex on cognitive engagement, $F(2,434) = 2.36$, $p = .09$, $\eta^2_{partial} = .011$. Female students did not have significant cognitive engagement outcomes (M = 59.12) compared to male students (M = 56.67) and students reported as neither male nor female (M = 57.81). In addition to the main effects, the interaction effects were also non-significant, as shown in Table 3.21. This indicates that effects on cognitive engagement were the same regardless of modality type, sex, and/or race/ethnicity.

Table 3.21

Three-Way Analysis of Variance Summary Table for the Effect of Modality Type, Sex, and Race/Ethnicity on Cognitive Engagement

Source	*df*	*SS*	*MS*	*F*	*p*
Modality type	2	24.89	12.45	0.16	.85
Race/ethnicity	1	7.05	7.05	0.09	.77
Sex	2	370.53	185.26	2.36	.10
Modality type x Race/ethnicity	2	3.44	1.72	0.02	.98
Modality type x Sex	4	254.99	63.75	0.81	.52
Race/ethnicity x Sex	2	140.87	70.44	0.90	.41
Modality type x Race/ethnicity x Sex	2	16.37	8.19	0.10	.90
Error	434	34,036.27			
Total	450	1,526,553.00			

Note. Task x Race/ethnicity are the results of the interaction effect between variables

$$Cognitive\ Engagement_i = b_0 + b_1 mode_i + b_2 race/ethnicity_i + b_3 sex_i + \epsilon_i \qquad (3.5)$$

Due to the violation of the assumption of normality based on the Shapiro-Wilk test, we cannot assume the data (scores of cognitive engagement) are normally distributed. Therefore, an

independent samples Kruskal-Wallis nonparametric test was run to evaluate the effect of the independent variable of modality type on cognitive engagement score. The results of the test are nonsignificant (p = .63) indicating there is not a significant difference between modality type on the dependent variable of cognitive engagement score.

Effect of time spent on technology at home on cognitive engagement. A between-subjects main effects analysis evaluated time spent using technology at home (in hours per day) on student cognitive engagement, see Equation 3.6. Levene's test of equality of error variance was not significantly different $F(4, 302) = 1.60$, $p = .17$; therefore we can conclude that variances between time spent on technology at home are not significantly different. Results from the one-way, between-subjects analysis of variance are presented in Table 3.22. The dependent variable was the total score on the cognitive engagement measure (CE-S-DSP & SOS). The independent variance was time spent on technology at home (in hours per day) with five levels: (a) none, (b) less than 1 hour, (c) between 1-3 hours, (d) between 3-5 hours, and (e) more than 5 hours. The main effect of home technology time on cognitive engagement was not significant, $F(4,302) = 2.22$, $p = .07$, $\eta^2_{partial} = .067$. There was not a significant difference between self-report of cognitive engagement for students who reported spending no time on technology at home each day (M = 55.00), students who reported spending less than one hour of time on technology at home each day (M = 57.36), students who reported spending more than one hour but less than three hours of time on technology at home each day (M = 56.99), students who reported spending more than three but less than five hours of time on technology at home each day (M = 57.43), and students who reported spending more than five hours of time on technology at home each day (M = 51.5).

$$Cognitive\ Engagement_i = b_0 + b_1 home\ technology_i + \epsilon_i \quad (3.6)$$

Table 3.22

One-Way Analysis of Variance Summary Table for the Technology Use at Home (in Hours per Day) on Cognitive Engagement

Source	*df*	*SS*	*MS*	*F*	*p*
Time spent on technology at home	4	732.39	183.10	2.22	.67
Error	302	24,922.83	82.53		
Total	306	1,012,411.00			

After descriptively looking at the mean scores on the cognitive engagement measure (CE-S-DSP & SOS; see Table 3.23), it was evident that mean scores are very similar. This may be due to the fact that the majority of the sample (62.4%) uses computers at home both educationally as well as non-educationally, while the remainder of the sample (37.6%) did not report specific computer program use at home. As a result, it would not make sense to further investigate the impact of computer use at home on student report of cognitive engagement.

Table 3.23

Cognitive Engagement Total Score Mean by Type of Computer Use at Home

Home computer use	*M*
Educational use	56.68
Yes	56.68
No	59.21
Non-educational use	
Yes	56.69
No	59.17
Both	
Yes	56.65
No	59.09
None	59.09

Qualitative Analysis. This section will explore the qualitative analyses by providing a review of the interview and coding process, an overview of the data uncovered through the coding process, and a summary of the main themes. This section will also qualitatively discuss modality attitudes described in the interviews, as well as discuss the links between performance task score and reported favorability.

Participant interviews ($N = 7$) were conducted in an attempt to explore the relationships between students' self-report of cognitive engagement, performance on the performance task, and the performance task modality (e.g. technology-enhanced, technology-enabled, paper-and-pencil). Table 3.12, previously described in RQ1, displays the descriptive statistics for the sub-sample of students who participated in the follow-up interview.

As discussed in Chapter II, initial codes were iteratively updated with analysis of the interview data, until final codes and themes were identified. Initial codes were generated based

on the theory of cognitive engagement (e.g. students who expressed favor or enjoyment towards a specific mode may be more engaged with that specific modality). Initially, codes were created to capture three main components. First, the overall code of *engagement* was used to highlight quotes that may demonstrate enjoyment, excitement, or favor including identification of a favorite task, enjoyment of the ability to write, and favor toward the movement, while the overall code of *disengagement* was used to highlight quotes that may demonstrate frustration, least favor, or confusion. Transcripts were also evaluated as *context-dependent* signifying a mention of a specific component that may infer either engagement or disengagement, depending on the context of the discussion. Within the context dependent codes, transcripts were also coded for *features of task* which identify specific components of the tasks that were mentioned by the students, see Table 2.6 from Chapter II. The *features of task* that were identified are more specific components beyond what is mentioned within the engagement or disengagement codes. The initial codes (Table 2.6) did not include platform specific codes in an attempt to capture platform favorability organically through engagement and disengagement discussions.

The final codes were categorized into three major themes, or overarching categories, and are displayed in Tables 3.24-2.26. Table 3.24 highlights the first category and includes codes that discuss positivity and negativity towards computer-based assessments in comparison to paper-and-pencil assessments. Table 3.25 highlights the second category and includes codes that are specific to one of the three performance task modalities. Table 3.26 highlights the final category and includes codes that indicate engagement or disengagement.

After the coding process was underway, it was evident that there were differences between discussions of *positivity* and *negativity* towards both the use of computers as well as specific tasks beyond engagement and disengagement coding. Therefore, coding was revised to

include *positivity* and *negativity* towards computer use as well as the same codes towards paper-and-pencil in order to discern positive and negative feelings towards modality, in general (not specific to the current performance task). This allowed for separation between when comments were made about the performance task, in general versus when comments were made about specific components of each modality. Originally, *indifferent* was included as a code; however, after the coding process was finished, it was clear that the comments made were either positive or negative so *indifferent* was removed from the coding.

The new coding scheme began to take on a layered approach with subcategories and codes under each main level. Subcategories are displayed in Tables 3.24-3.26. Subcategories were created to further delineate between main codes (e.g. positivity expressed by one student towards the animation effects of the technology-enhanced while another student expresses positivity towards the movement of the mobile). The subcategories of each modality are also where favorite tasks and least favorite tasks were specified.

Table 2.6 (see Chapter II), highlights the initial coding which originally specified engagement and disengagement; however, after going through the transcripts and practicing the coding process, it was evident that not all discussions of engagement were modality specific. Additionally, it was hard to discern whether or not a student was actually identifying a time they were engaged just based on a mention of a specific component. Therefore, many of these codes would have been speculative. To avoid this, a similar theoretical approach was taken as the original coding scheme (e.g. students who express favor or enjoyment towards a specific mode may be more engaged with that specific modality); however, these codes were created for the overall experience and not necessarily towards a specific modality (e.g. a student identifying the process as repetitive). Table 3.24 highlights the final codes and subcategories for modality that

are not specific to a performance task modality. Consequently, positivity and negativity coding towards the specific modality aims to help identify themes and explain overall engagement towards a specific task. Table 3.25 highlights final codes and subcategories that are modality specific. Furthermore, since the interview was conducted at the conclusion of the entire assessment (i.e., after a student completed all three tasks), the sentiments expressed about engagement could be confounded with the completion of all three tasks; whereas, the measure of engagement discussed throughout this study evaluates engagement only after the first task was completed.

After the coding process, the final codes and subcategories of codes were found to fall into three categories, or themes. The first category specifies codes discussing positivity and negativity towards computer-based assessments versus paper-and-pencil assessments, see Table 3.24. This first category is not specific to modality of the current tasks; rather, it discusses the use of modality in general. The second category highlights codes that are specific to one of the three modalities specific to the performance tasks, see Table 3.25. The third category highlight codes that may indicate either engagement or disengagement, see Table 3.26. While it is not possible to discern engagement from the interviews, the codes highlighted specify components of the study that may be a contributing factor to more or less engagement throughout the assessment. Additional, this third category of codes was not modality specific and related to the overall experience.

Table 3.24

Final Codes and Subcategories for Modality Not Specific to Performance Task

Code	Sub categories of code
Computer positivity	Avatar
	Can go back to problems (in certain programs)
	Animated
	Cannot look ahead (mitigates anxiety)
	Do not have to go back to problems (in certain programs)
	Drag and drop functionality
	Easier
	Fun
	Typing
	Use of scrap paper
	Visual representations
Computer negativity	Bright screen
	Cannot go back to problems (in certain programs)
	Confusion on how to use program or computer
	Drag and drop functionality
	Causes headache
	Hurts eyes
	Stressful
	Technical difficulties
	Typing functionality
	Cannot write on screen
Paper-and-pencil positivity	Can go back to fix answers
Paper-and-pencil negativity	Plain and boring (accepted on the platform)
	Can see ahead (anxiety provoking)
	Does not allow for movement
	Hand hurts from writing
	Lots of writing

Table 3.25

Final Codes and Subcategories for Modality of Performance Task

Code	Sub categories of code
Computer- enhanced	Positivity
	Favorite task
	Appealing
	Easy to interpret
	Animation effects
	Movement of mobile
	Fun
	Theme
	Visuals
	General interest
	Negativity
	Least favorite task
Technology-enabled	Positivity
	Favorite task
	Simplicity
	Straight Forward
	Negativity
	Least favorite task
	Cannot write
	Confusing
	Harder
	No visual hints
	Simplicity
	Slow
	Do not like
Paper-and-pencil	Positivity
	Favorite task
	Ability to write
	Easier
	Fast
	Negativity
	Least favorite task
	Plain

Table 3.26

Final Codes and Subcategories for Engagement and Disengagement

Code	Sub categories of code
Engagement	Already knew what steps to take to complete the problem
	First task completed
Disengagement	Repetitive

Codes discussing general modality. Students expressed general interest as well as disinterest towards computer-based as well as paper-and-pencil based modalities. Overall, more students expressed interest towards computer-based assessments (n = 7) in comparison to paper-and-pencil assessments (n = 1).

Positivity towards computer-based assessments. Overall, all students expressed general appreciation for using computers to take tests (n = 7). Specifically, students valued the features such as avatars, animation, drag and drop functionality, visual representations of problems, as well as typing abilities that are possible in a computer-based test in comparison to paper-and-pencil tests.

Many students referenced the computer animation (movement) as the main reason why they liked the computer-based assessment, specifically the technology-enhanced performance task. Five students expressed favor towards the tilting of the mobile (Participants 152, 155, 160, 169, 200, Personal conversation, October 13, 2016) with specific references to it being more helpful (Participant 152 and 169). One participant (169) talked about the animation by saying "it explained itself really easily and it moved so that way I could understand if it were correct or if it weren't, basically." Another student mentioned explained how the animation was favored

"because you would put an answer then it would balance it out…so then you actually knew what you were doing" (Participant 200). Other students referenced the movements as making the problem easier (Participant 160 and 169) by stating,

> it was just really easy to see how things were represented…based on the animations and the shapes and things…. it was pretty easy to see what you're doing and that it's…it'll…if it'll tell you if it's wrong because you want it to be all the way balanced out. (Participant 160).

Another student concluded the test was more fun because of the animations (Participant 169).

The use of visuals also made the test appealing for those who seemed to think of themselves as visual learners, "I work better with pictures and visual things and when it evens it out it makes it easier" (Participant 179). Additionally, others took notice of the overall theme (e.g. castle library) of the technology-enhanced performance task and commented how the theme may affect students' overall favorability of the task (Participant 160).

Some students cited computer-based assessments as easier and more fun (n = 3). One student noted a decrease in anxiety using computer-based assessments because she was unable to look ahead and see many problems still to come as well as the fact that she did not have to go back to visit already answered problems (Participant 155). Consequently, other students (Participants 155 and 160) noted their appreciation of the ability to go back to previous problems (if the program allows) as well as the ability to use scrap paper to help solve the items (Participant 179).

Negativity towards computer-based assessments. Although there was some negativity expressed towards using computer-based assessments, much of the negativity was around the use of a screen. Students (n = 3) mentioned some aspects of screen use as a negative feature of

computer-based tests (Participants 152, 160, 169). These features included the brightness of the screen causing headaches. One student noted that the screen "gets really bright…sometimes it gives me headaches because of the bright screen" (Participant 169) while another student noted that "sometimes it's easier to get a headache if you're just staring at a screen for awhile…whereas taking a [paper-and-pencil] test … I just like it better because it's not staring at a bright screen" (Participant 160). Other students mentioned the computer causing their eyes to hurt over time, "it starts to kind of hurt my eyes after a little bit…staring at a screen" (Participant 152).

The theme of concern around bright screen during technology usage was mentioned by three different participants as a negative feature of computer-based tests (Participants 152, 160, 169), despite that the performance task testing was students' typical school-based experience. Additionally, students completed the task on Apple laptop computers (Washington) or Chromebooks (Oregon, North Carolina) and were in computer lab or classroom settings free from direct sunlight. Additionally, students did have access to brightness settings on the computers, which some chose to adjust prior to beginning the tasks.

Additional negative features discussed included confusion that could result from using a new software or hardware program, potential technical difficulties, struggles with dragging and dropping, and keyboarding difficulties. Two students (Participants 155 and 160) discussed the negativity associated with certain programs such as not being able to go back to a previous item, "you can't go back so if later if I was like 'Oh! I did that wrong' I couldn't go back and fix it" (Participant 155) while another student discussed the negativity associated with not being able to write on the screen, "you can't really draw on the computer… I mean some they have a drawing program where you can do that but you can't to that…you can on a paper-and-pencil though"

(Participant 179).

Positivity towards paper-and-pencil based assessments. Overall, students expressed somewhat higher positivity towards computer-based assessments in comparison to paper-and-pencil assessments in the interviews; however, the interview questions did not specifically highlight student attitudes towards general paper-and-pencil based assessments. As a result, the overall favorability (or lack of favorability) towards paper-and-pencil assessment was not as evident. Only one student made a favorable comment towards paper-and-pencil assessments when discussing interest in computer-based assessments, noting the ability to go back and fix answers, "I liked paper-and-pencil because I write a lot to like get all my answers and ideas out there… it just helped me write everything out which would be easier…[it's my most favorite] because I got to like write out…or like…express like all of my ideas on the page…so they would be there even if someone had like no idea what I was doing…" (Participant 180). It is worth mentioning that specifics about paper-and-pencil based assessments were more thoroughly discussed during the questions on specific modality.

Negativity towards paper-and-pencil based assessments. As previously mentioned, interview questions did not specifically address paper-and-pencil based assessments. However, four students made note of specific downfalls to the use of paper-based assessments. Specifically, two student noted the acceptance of the fact that paper-and-pencil tests are plain and boring (Participants 180, 200) while one student noted an unfavorable attitude towards the paper-and-pencil modality's inability to move (e.g. animation). The ability to see ahead to future items within the test was noted multiple times by one student who said that the feature can be anxiety provoking,

> I guess I kind of like tests on the computer a little more because with a test on paper then I have like anxiety stuff so I could look at the next problem and stress about that but on the computer I just take it one problem at a time and I can't get ahead of myself…so I like that… If I'm just looking at them one at a time it's a lot less stressful (Participant 155).

Additionally, two students noted the amount of writing that is associated with paper-based assessments (Participants 179 and 180) with Participant 180 mentioning the pain that can be associated with a lot of writing,

> I like using computers to take tests because I feel like if it's a really long written test, my hand starts to hurt, so it's like better than written … it's just like better than paper because you're not just like sitting at your desk like writing and writing and writing…there are like fun pictures and like you get to like type or drag or something instead of just like just paper-and-pencil (Participant 180).

Codes discussing performance task specific modality. Aside from a general discussion on the use of computers to take tests, a majority of the interview questions were focused on attitudes and opinions towards the individual platforms within the study. This resulted in the coding positive and negative attitudes and opinions based on performance task modality (technology-enhanced, technology-enabled, paper-and-pencil). Additionally, four students identified the technology-enhanced platform as their most favorite, while only two students identified the paper-and-pencil modality as their favorite, and one student identified the technology-enhanced as their favorite (see Table 3.25). Consequently, six students identified the technology-enabled modality as their least favorite, one as the technology-enhanced as their least favorite, and zero for the paper-and-pencil as their least favorite (see Table 3.25). Additional

positive and negative attitudes and opinions across modalities are discussed below.

Technology-enhanced. The technology-enhanced task was the only task that did not have negative coding (other than least favorite task) identified by students while also having the most positive codes (see Table 3.25). The many positive codes highlight features that are specific to the enhancements implemented for this modality. To begin, students identified the animation (*n* = 4), visuals (*n* = 1), theme (*n* = 2) and movements of the mobile (*n* = 7) as enjoyable features. Some even went so far as identifying the mobile movements as attributes that contributed to the difficulty level of the task (*n* = 3), "I like the animated stuff because you would put an answer then it would balance it out…so then you actually knew what you were doing" (Participant 200), further specifying that the movements made the item easy to interpret,

> it was just really easy to see how things were represented…based on the animations and the shapes and thing …. it was just really easy to see how things were represented…based on the animations and the shapes and things.

Another student mentioned the animation helping to figure out how to solve the problem, "It explained itself really easily and it moved so that way I could understand if it were correct or if it weren't" (Participant 200). Although five students identified the technology-enhanced mode as their favorite task, one student changed their mind last minute changing the count from six to five. Two students identified the technology-enhanced modality as fun, specifically with the added features of the technology, "there are like fun pictures and like you get to like type or drag or something instead of just like just paper-and-pencil" (Participant 180).

Technology-enabled. The technology-enabled mode had the most negative codes and the fewest positive codes (see Table 3.25). The positive codes that were highlighted include the modality's simplicity (*n* = 1) and straightforwardness (*n* = 1). There were many negative

attributes mentioned by students, despite the numerous similarities between the technology-enabled task and the technology-enhanced task. Notably, students identified not being able to write (n = 1), confusion of the item (n = 2), and difficulty level (n =2) as downfalls. One student noted the difficulty level in addition to the inability to write by saying, "you can't really write out things on it and it just feels a little bit harder in my opinion because it takes several tries to do it" (Participant 179). Confusion was a theme discussed by Participant 152, "it was a little bit more confusing…it really does not give you much visual hints" and Participant 169 "it went a little bit slower for me because…for me… it was just a little bit more confusing." One student referenced the technology-enabled platform as being outdated,

> that one seemed like a very like… um… like old school… well not really old school because they're computers…but…the most like…something we would use today sort of… (Participant 155).

While two students (Participant 180 and 200) referenced it as being plain,

> I didn't really like it… I'd say it was sort of plain…like…if they're going to put something on the computer, I feel like it's gotta be cool. Like paper is like…alright…it's paper… (Participant 200)
>
> it was like every other test…like our MAP testing… it was like that…just kind of like out there and blank …. like it was just like not like boring but kind of like plain (Participant 180)

Differences between the technology-enabled and technology-enhanced task that were noted as negative features included the lack of visual hints (n = 2), simplicity (n = 3), and slowness of task (n = 1). One student noted that the different pictures may have contributed to the difficulty level, "they don't even out and they have different pictures so it makes it, in my

opinion, a little bit harder" (Participant 169). One student made it very clear this was just not a task that he liked while all but one student (n = 6) identified it as their least favorite task.

Paper-and-pencil. The paper-and-pencil modality is lauded as the second favorite task following the technology-enhanced modality. No students identified the paper-and-pencil task as their least favorite (see Table 3.25). Students only identified plainness of the modality (n = 1) as a negative feature, mentioned in conjunction with simplicity discussions about the technology-enabled task. Students largely appreciated the ability to write on the paper-and-pencil modality (n = 4),

> I liked that one because … I could write my answers…and write them out on the paper instead of just like thinking it in my head…I like to see stuff written down too… (Participant 155).

Students found the task to be easier (n = 1) and faster (n = 1) in comparison to their technology-enhanced and the technology-enabled tasks, "I think it's a bit faster writing than putting up things on the computer" (Participant 179). One student identified the paper-and-pencil task as their favorite:

> I liked paper-and-pencil because I write a lot to like get all my answers and ideas out there… it just helped me write everything out which would be easier…[it's my most favorite] because I got to like write out…or like…express like all of my ideas on the page… so they would be there even if someone had like no idea what I was doing…(Participant 180).

Engagement and disengagement codes. Although it would not be possible to discern engagement by coding interviews, it is worth noting that a few miscellaneous codes were mentioned that may contribute to student engagement. First, four students (mentioned over eight

times) noted the repetitiveness of the tasks. This is a concern for engagement and may have a negative effect on items beyond the initial task. Students also identified the first task as "fun to figure out" since they were "figuring out the problem for the first time" (Participant 155, Personal conversation, October 13, 2016). The newness of the task (difficulty level, excitement) may only apply to the first problem the student encounters and may not be applicable to subsequent tasks. This was also reiterated when a student mentioned already knowing what to do "since it was the second one" which caused a deeper understanding and could impact engagement (Participant 169 Personal conversation, October 14, 2016).

Qualitative themes. After the interview data was transcribed and coded, the data were analyzed to determine emerging themes. The data were based from the interviews conducted and will be used to help triangulate the quantitative data (e.g. engagement scores, performance task score). Table 3.27 highlights the overall themes that were uncovered and the number of quotes that were referenced per theme.

Table 3.27

Themes Uncovered and Quote Count per Theme

Theme	Count
The technology-enhanced task was favored	5
Students appreciated the note taking during the paper-and-pencil task	5
The technology-enabled platform was the least preferred	8
Students liked interacting with the items that animate	7
Students preferred using technology in comparison to paper-and-pencil	3
Students did not always like staring at a screen	3
Students wanted the ability to go back and fix an answer	2

The Technology-Enhanced Performance Task was Favored, Overall. Students seemed to, overall, enjoy the technology-enhanced performance task, with four students reporting the technology-enhanced mode as their most favorite and two students reporting it as their second favorite. Even students who did not mention the technology-enhanced task as a favorite still talked about the many strengths associated with the platform. Overall, students expressed favorability to the movements (Participants 152, 155, 169, 179), specifically it being more fun (Participant 169). Additionally, students reported that the theme made the task more interesting (Participant 160) and the visuals were appealing (Participant 179) and may have made the task easier (Participant 179).

Students Appreciated the Note-Taking Ability while taking the Paper-and-Pencil Task. Students reported appreciation to the paper-and-pencil task, mainly for the familiarity of the platform, the ability to take notes while solving the item, and to be able to go back and forth between items within the test. Two students reported the paper-and-pencil task as their most favorite mode, five students reported it as their second favorite, and no students reported it as their least favorite. Overall, students appreciated the ability to write down answers instead of doing mental mathematics (Participants 155, 180), specifically noting that writing is faster than using a computer (Participant 169) and the ability to draw out an answer and show your work is favorable (Participant 179).

The Technology-Enabled Platform was the Least Preferred. Students were largely unimpressed with the technology-enabled platform. Only one student reported the technology-enabled platform as their most favorite platform. All other students (n = 6) reported the technology-enabled as their least favorite platform. Notably, students expressed that the technology-enabled task was more confusing (Participants 169, 152) perhaps due to lack of

visual hints (Participant 152) which could have made it harder (Participant 169). One student reported the inability to write out ideas causing the test to be harder and take several tries (Participant 179). Students noted that the test was plain (Participant 180) and "old school" (Participant 155).

*Students like Interacting with the Items that Animate***.** Five students mentioned a preference towards the movement of the animated items within the technology-enhanced performance task. Students mentioned that the movements were either enjoyable (Participant 155, 169), helpful (Participant 152) or made the problem easier to understand (Participants 160, 200, 169).

Overall, Students Prefer Using Technology in Comparison to Paper-and-Pencil. Despite the fact that five of the seven students reported one of the two technology modes their favorite, many also mentioned their preference towards using a computer. One student noted a decrease in anxiety when using a computer to answer test items because of the inability to look ahead to problems (Participant 155). Another student noted that a computer-based test may be easier (Participant 179) and save the hand pain from so much writing on a written test (Participants 179 and 180). One student (Participant 180) went so far as to say the computer-based tasks had features that made them more appealing (e.g. drag and drop).

Students Do Not Always Like Staring at a Screen. There were a few students who mentioned screen brightness as a major downfall to using computers with one student mentioning that it hurt their eyes to stare at a screen (Participant 152) and others mentioning that staring at the bright screen may cause a headache (Participant 160 and 169). Some students noted the relaxing look of the animation.

Students Want the Ability to Go Back and Fix an Answer. Some computer programs limit students' ability to navigate back to a previous item. Two students mentioned this as a downfall of computer modalities (Participants 155, 160) specifically mentioning realizing a mistake was made and being unable to return to a previous item to correct the mistake.

Table 3.28

Student Rankings of Modality Favorability (N = 7)

Student ID	Technology-enhanced	Technology-enabled	Paper-and-pencil
152	Most	Least	Middle
155	Middle	Least	Most
160	Least	Most	Middle
169	Most	Least	Middle
179	Most	Least	Middle
180	Middle	Least	Most
200	Most	Least	Middle

Note. Most =ranked favorite mode; Least = ranked least favorite mode;
Middle = middle ranking.

Table 3.29

Total Favorability Rankings by Modality (N = 7)

Modality	Technology-enhanced	Technology-enabled	Paper-and-pencil
Most favorite	4	1	2
Middle	2	0	5
Least favorite	1	6	0

Order of Modality. Since all participants completed the performance tasks in a counterbalanced order (see Figure 1), it is important to consider the platform order when evaluating interview feedback. This helps to evaluate the feedback (positive and negative) by

considering the fact that the task that was completed first may have more favorable opinions expressed in comparison to the task that was completed last. The negativity of the repetitiveness of the items is noted by four students; while, two students mentioned the excitement and unpredictability of the first task completed. The descriptive table displayed in Table 3.30 highlights the order of tasks completed. Additionally, Table 3.30 displays the task preference in addition to first task completed. All students but one (n = 6) indicated preference (favorability) for the same task that was completed as their first task. This suggests that students favored the first task completed; thereby, indicating that students may have been most engaged during the first task completed. This result justifies the evaluation of the first task students completed, which is the framework used for the current study. This limitation is further discussed in Chapter IV.

Table 3.30

Task Preference and Order (N = 7)

ID number	Task preference	First task completed
152	A	A
155	C	C
160	B	B
169	A	A
179	A	C
180	C	C
200	A	A

Note. Task A = technology-enhanced; Task B = technology-enabled; Task C = paper-and-pencil

Connection Between Performance Task Score and Interviews. Based on the outcome scores of the paper-and-pencil performance task (see Table 3.13), there does not seem to be a connection between task preference (Table 3.30) and outcome score on the paper-and-pencil

performance instrument (Table 3.13). Students who reported the paper-and-pencil task (Task C) as their most favorite (IDs 155, 180) scored 24% and 100%, respectively, on the performance task. Additionally, students who received a perfect score on the performance task (IDs 152, 160, 180) each articulated preference of a unique task (Task A, Task B, and Task C, respectively). The interpretation of this will be discussion in Chapter IV.

CHAPTER IV

DISCUSSION

This chapter presents an interpretation of results based on the research questions and literature discussion. The chapter further assesses the significance of the findings and discusses the limitations of the results and interpretation. Recommendations and implications for future research are presented.

Review of Study Components

Findings from the literature review suggested there was a need to evaluate student cognitive engagement within a technology-enhanced performance assessment, specifically looking at type of assessment modality. The current research study examines student self-reported cognitive engagement within a technology-enhanced performance assessment in mathematics for middle school (grades 6-8).

Mathematics, a branch of STEM, was chosen as the subject area of focus for the context of the study performance tasks across modalities due to in part to availability of the project and due to the increase in STEM education efforts in the U.S. described in Chapter I. Despite a somewhat increased nationwide focus on STEM, overall student interest in STEM has been reported as only marginally increasing in recent years, such as by one percent between 2010-2014, as reported by survey research (ACT, 2014b). Evaluating student effort and interest as well as student performance ability within STEM settings may help shed light on STEM interest investigations and help recognize how to encourage students in other STEM efforts. This is particularly true for marginalized groups such as females, students who are non-White, and students who are from a low SES household.

My study aimed to measure student cognitive engagement specifically across an overall composite that was intended to consist of five factors: (a) deep processing, (b) shallow processing, (c) persistence, (d) effort, and (e) importance. No items within factors were dropped in the current study. These five factors were chosen as the theoretical elements that encompass the construct of cognitive engagement, based on prior research findings in the field, as discussed in Chapter I. The application of cognitive engagement as well as perceived effort and perceived importance may particularly key to understand during an assessment in order to understand what students actually know versus the level at which they are willing to perform during a test event (Thelk et al., 2009; Wise & DeMars, 2005b).

Due to numerous technology affordances now available, assessments have started to include technology-enhanced items and tasks to meet CCSS college and career readiness benchmarks such as problem solving and 21st century skills. Current CCSS standardized assessments have started to include technology-enhanced tasks; yet, results from these assessments are still under investigation. Advancements in technology-enhanced assessments call for the need to examine modality differences between paper-and-pencil assessments and computer-based assessments, including different types of computer-based assessments. As a result, the second focus area investigated the quantitative and qualitative relationships between affective measure outcomes (e.g. cognitive engagement) and modality type.

Previous research has mixed outcomes regarding the impact of modality on student achievement and engagement. There were numerous studies that found marginally significant effects (Clariana & Wallace, 2002b; Neuman & Baydoun, 1998; Vispoel et al., 2001), while, other research resulted in no significance differences between modes (Bodmann & Robinson, 2004; Donovan et al., 2000; Finegan & Allen, 1994; Horton & Lovitt, 1994; King & Miles,

1995; Mason et al., 2001; Özalp-Yaman & Çağiltay, 2010). My study aimed to contribute to the literature when evaluating the impact of technology on student engagement.

The research questions for this study included two focus areas. The first focus area aimed to investigate the performance of the cognitive engagement survey (CE-S-DSP & SOS) with students while engaged in responding to a mathematics performance instrument. The CE-S-DSP & SOS was used as a self-report measure of cognitive engagement. The performance instrument was developed based on the CCSS 7.EE.3 standard: *Solve real-life and mathematical problems using numerical and algebraic expressions and equations* (National Governors Association, 2010), which includes the integration of positive and negative numbers, specifically within equations and incorporating other mathematical content strands such as number systems and mathematical practice (Schwols & Dempsey, 2013). My study examined the variance and internal consistency of CE-S-DSP & SOS within the context of a mathematics performance instrument, as well as student scores and group means.

Additionally, my study included interviews with a small sample of students, to evaluate their interview responses along with their cognitive engagement responses, their performance on the mathematics instrument, and the relationship between student scores on an external measure (MAP®) with the performance instrument.

The impact of computer use at home was also explored. Many studies have linked computer use at home with positive impacts on academic proficiency outcomes (Bennett et al., 2010; Casey et al., 2012; Fiorini, 2009; Halldorsson et al., 2009; OECD, 2006a; Tsikalas et al., 2007) while other studies have linked heavy use of home computer to negative impacts on academic outcomes (Fuchs & Woessmann, 2004; Malamud & Pop-Eleches, 2011; Vigdor et al., 2014), specifically in math (Wittwer & Senkbeil, 2008). As a result, my current study aimed to

examine the relationship between engagement and home technology use, by type of program used and time. Additionally, my study also aimed to qualitatively evaluate student attitudes through the interviews mentioned above towards technology-enhanced assessments in comparison to more traditional modalities such as paper-and-pencil or technology-enabled assessments.

Discussion of Findings

Research Question One. RQ1 evaluated the correlations, internal consistency, and variance within the measures. Additionally, RQ1 used qualitative analysis to examine the performance of the mathematics performance task from a purposive subset ($N = 7$) by evaluating MAP® scores, performance task scores, and interview data.

***Correlational Analyses*.** Correlational analyses between variables determined that for the overall data set, prior to investigation by platform modality as described in Research Question 2 below, there was a significant negative relationship in the data between sex and total deep processing score. This negative relationship between sex and total deep processing indicates that as values on one variable increase, values on the other variable tend to decrease; thereby, values on the variables resulting in opposing directions. Otherwise, no statistically significant relationships were seen for this data set between demographic variables and cognitive engagement scores. Notably, there was not a significant relationship between total score on the CE-S-DSP & SOS and either of the two demographic variables (sex and race/ethnicity). Additionally, sub scores were not used for my study; however, specific factors on the CE-S-DSP & SOS (deep processing, shallow processing, persistence, importance, and effort) were included in this analysis for clarity of what was seen in the statistically significant relationships between the variables.

Internal consistency. The CE-S-DSP & SOS was created as an adapted self-report tool to measure cognitive engagement within a formative assessment context. The CE-S-DSP & SOS is a combination of the SOS (Sundre, 1997) and the CES (Miller et al., 1996), later shortened and adapted to CE-S (Smiley & Anderson, 2011). Together, the CE-S-DSP & SOS was comprised of 21 items across five subscales to measure cognitive engagement (deep processing, shallow processing, persistence, importance, and effort). Reliability analyses were conducted to evaluate the 21 Likert items for internal consistency. Cronbach's alpha was calculated to determine the degree to which the items measure the construct of cognitive engagement. The value of alpha, $\alpha = 0.84$, is above the acceptable threshold of 0.80 indicating reasonable reliability and indicating that the survey does an acceptable job of measuring reliably for one overall score over the set of items. This is also consistent with previous reliability of the original measures. Although the reliability of sub measures were not used in the study, only two of the sub measures were above an acceptable threshold. This indicates that the other sub measures would need more items to establish reliable subscores, so subscores are not used in this study.

Variance. Variance of cognitive engagement outcomes showed fluctuation of scores around the mean indicating a range of student performance in the measure. Results of the cognitive engagement measure was analyzed using a CFA, to align with the theoretical five constructs in the research literature. The results were mixed with some fit statistics showing acceptable fit to the model and others not, because at least in part the data were not ideal for fitting to a CFA model (see Chapter III). Because subscores were ultimately not used in analyzing results in this study due to insufficient reliability in the subscore indices, fit of the model to the five subscore factor structure could be examined in future work. Future work could

include not only expanding the subscales to improved reliability but also exploration of latent variable models such as confirmatory IRT that may have better fit for the data.

Validation of the performance task. As mentioned, the performance instrument was developed based on the CCSS 7.EE.3 standard: *Solve real-life and mathematical problems using numerical and algebraic expressions and equations* (National Governors Association, 2010) which includes the integration of positive and negative numbers, specifically within equations and tying in other mathematical content strands such as number systems and mathematical practice (Schwols & Dempsey, 2013). In the purposive sample interview data, student outcomes on the MAP® assessment were evaluated for relationships to outcomes on the performance task used in the study. Four students from the purposive sample ($N = 7$) demonstrated consistent scores between the two measures (MAP® and the performance task); while, three student scores did not demonstrate consistency. Further investigation of the inconsistent outcomes revealed that, in some cases, engagement may have impacted the lack of consistency between the two measures; while, in other cases, engagement may not have contributed to engagement. This should be interpreted cautiously due to the very small sample size. Although this study was limited to only exploring a purposive subsample ($N = 7$) of academic performance task outcomes, future work in this area could explore the performance of all academic student outcome measures. This was not possible for this study because the scoring of the larger data set by NWEA™ has not yet been completed.

Research Question Two. RQ2 evaluated the relationship between different performance task modalities and student self-report of cognitive engagement. This was further investigated by evaluating the relationship for cognitive engagement by performance task modality given demographic characteristics. Additionally, RQ2 examined the relationship between cognitive

engagement and time spent on computer use at home as well as the type of computer use (educational or non-educational) was evaluated. Qualitative analyses were used to further evaluate and interpret student attitudes towards different assessment modalities and features.

Relationship of modality with cognitive engagement. Educational and technology advances, in addition to previous literature, have described a need for a shift in educational assessments to include more technology affordances, as discussed in Chapter I. This shift prompts a need to examine how these technological affordances impact student affective outcomes on assessments. In addition to technological affordances, the modality in which the assessment is delivered (e.g computer-based or paper-and-pencil) may also have an impact on student outcomes. Research has mixed outcomes regarding the impact of modality on student achievement and engagement. As a result, this study aimed to further investigate technology affordances and modality type on student outcomes.

There was not a significant relationship between type of modality (technology-enhanced, technology-enabled, paper-and-pencil) with student self-report of cognitive engagement, for this data set. Additionally, the two extreme platforms (technology-enhanced versus paper-and-pencil) were investigated to determine if there was a significant difference in cognitive engagement. This result also proved to not have a significant difference. The current study results are somewhat consistent with the literature, because there are mixed outcomes regarding the impact of type of assessment modality on student engagement.

Relationship of modality, sex, and race/ethnicity with cognitive engagement. The inclusion of technology enhancements within assessment may lead to academic and cognitive differences between groups of students, particularly gender differences in the STEM fields (Halldorsson et al., 2009; M.-K. Lee, 2009; Ripley, 2009; Sorensen & Andersen, 2009). The

current study examined cognitive engagement differences by demographic characteristics (sex, race/ethnicity).

Results examining the effect of type of modality on cognitive engagement while considering demographic characteristics (sex, race/ethnicity) did not result in a significant difference of student self-report of cognitive engagement. Students who are White did not have significantly different cognitive engagement outcomes, when examined by platform, than students who are non-White. Additionally, under the same conditions, female students did not have significantly different cognitive engagement outcomes in comparison to male students or students who did not report their sex. This indicates that the effects of modality type on cognitive engagement were not significantly different for this data set regardless of modality type, sex, and/or race/ethnicity. This is somewhat inconsistent with the literature combined with theory that indicates there is a relationship of student sex (DeWitt et al., 2013; Elster, 2014) and race/ethnicity (DeWitt et al., 2013; Wang, 2013) with student aspirations in STEM, which has been linked to student engagement (Elster, 2014; Lyons & Quinn, 2010). However, direct research between demographic characteristics and student engagement during a STEM-focused assessment has been largely unexplored. Additionally, as discussed later in Chapter IV, there is a probability of survey validity concerns as well as the probability of a Type II error for this study, so results should be treated with caution.

Relationship between time spent on technology at home on cognitive engagement. A growing body of literature indicates there is relationship between students' technology use in terms of time spent using technology with academic and affective outcomes. For affect, the literature in this area reports mixed outcomes with studies showing computer use at home with positive affective outcomes (Fiorini, 2009). Additional studies further explore the effect of

certain program types on student academic outcomes. The current study aimed to examine the relationship between time previously spent on computer use at home and cognitive engagement when using technology-based assessments.

Results examining the time spent on technology at home (in hours per day) did not result in a significant difference of students' self-report of cognitive engagement. Further descriptive analysis of type of computer program was explored to investigate the relationship between program type (educational versus non-educational) on student self-report score of cognitive engagement. Mean cognitive engagement scores by type of computer program used at home did not result in mean differences worth further exploring. Many students reported computer use that included both educational and non-educational programs. As a result, mean scores of cognitive engagement across program usage were very similar.

Student attitudes towards technology-enhanced assessments. Further qualitative analyses were conducted to examine student attitudes about using technology to take assessments as well as student attitudes about the technology features of assessments. For the current study, a purposive sample of seven students participated in interviews. The interviews were then analyzed as described in Chapter II to consider signs of engagement or disengagement towards a specific modality.

Overall, as discussed in Chapter II, students described that they enjoyed the technology-enhanced performance task and discussed many strengths about the platform, specifically interacting with items that animate. Students reported that item animation helped make the problems more understandable and easier to solve, in comparison to items that did not have animation. Students largely reported favorability towards using technology within an assessment in comparison to paper-and-pencil; however, students reported overall dissatisfaction with the

technology-enabled platform. This suggests that comments regarding computer use within an assessment context may have a relationship with platform; however, the quantitative results from the particular cognitive engagement composite measure used here did not show a relationship with platform between the two technology approaches used, at this sample size and for the questions asked.

Students also reported components of the paper-and-pencil assessment that were favorable. Overall, students appreciated being able to take notes during the paper-and-pencil assessment. Although scrap paper was offered for the computer-based platforms, students rarely used it. Additionally, students valued the ability to go back to a previous item and revise their response, something that was not enabled for the computer-based modalities in this study but was a feature of the paper-and-pencil assessment. This suggests that by implementing at least these two paper-and-pencil type components on a computer-based assessment, students may find themselves interacting with preferred features from both platforms. For instance, response type input capabilities may limit the *scratch* usability of technology platforms. Even on tablet with a touch-responsive screen, for instance, a high-quality stylus is required to replicate paper marking, and not use of finger or more affordable quality stylus often found in schools. The *go back* features are available in some platforms, but require more sophistication of the platform and database to be able to 'repopulate' the prior screen for the individual student's prior response, which would remain intact and present on paper-and-pencil without additional investments. Furthermore, for tasks often found most desirable on technology platform, going back may expose the prior answers, and therefore not be a feature of the task. Paper-and-pencil tasks are often much less interactive due to the modality, and therefore are not designed to take

advantage of connected, synthesized, elaborated performances that have the vulnerability of exposing future answers with earlier responses.

As mentioned in Chapter III, students who received a perfect score on the performance task (IDs 152, 160, 180) each articulated preference of a unique task (Task A, Task B, and Task C, respectively). This could be interpreted as not having a qualitative connection between performance task outcome score and favorability; however, tasks that were reported as students' favorite task (for students who received a perfect score) directly corresponded to the first task the student received, see Table 3.30. This implies that the order of modality could impact favorability.

Limitations

The study has some substantial limitations, especially regarding the data and results. To begin, results gathered are specific to the sample (OR, WA, NC), include two private schools and only students in grades 6-8 within the subject of mathematics; therefore, mathematics may not generalize to other grades, subject areas, measures, schools, and/or states. Since the mathematics performance task focuses on seventh grade content standards, using a sample of grades 6-8 may provide differences in outcomes that are not measured within this study; specifically, the task may be too difficult for sixth grade or too easy for eighth grade, both which could impact engagement and performance outcomes. Lastly, the sample is somewhat homogenous, therefore, demographic factors such as race, ethnicity, English language status, SES, and disability status cannot be appropriately examined for additional variance within the models.

In addition to sampling concerns, the subject of mathematics adds complexity when measuring student engagement and motivation (Miller et al., 1996) due to gender differences prevalent within STEM fields, specifically within mathematics (DeWitt et al., 2013; Elster, 2014;

Wang, 2013). In addition, the study did not control for prior achievement in mathematics which could impact student mathematics performance and subsequent self-report of cognitive engagement. Although the study collects RTE data, further exploration of RTE may contribute to the overall measurement of cognitive engagement due to older grades exhibiting greater rapid-guessing behavior during low stakes assessments than students in lower grades (Ma et al., 2011; Wise et al., 2010). Additionally, *alternate forms* of the same test cannot be determined and are solely based on content expert opinions.

When implementing various modalities, there is always a concern of equality between items and platforms. Naturally, there are different features between computer-based assessments and paper-and-pencil assessments. Platform design attempted to mitigate for differences; however, there were components that remained unaltered (e.g. navigation between items, colors/brightness). These features, as noted through interview data, may have played a role in shaping student attitudes and preferences of modalities. Future modality studies should implement additional controls for these dissimilarities. Additionally, although the current study aimed to developed advanced animations for enhanced items, not all were designed as originally intended. Animations that are subpar may not result in the appropriate enhancements needed to differentiate the platform from the technology-enabled.

The counter-balanced research design (see Figure 2.1) does indicate a potential violation to some degree of independence for the ANOVA tests, since the same students are used in each treatment condition through the rotated design. In an attempt to mitigate the independence violation, only the first assigned condition performance task outcome and cognitive engagement survey outcome were analyzed for the current study.

Self-efficacy and confidence present continuing challenges in STEM (DeWitt et al., 2013; Elster, 2014; Lyons & Quinn, 2010; Wang, 2013) and is particularly prominent in marginalized populations such as females (Louise Archer et al., 2010; Elster, 2014), racial and ethnic minorities (DeWitt et al., 2013; Wang, 2013), and students from low socio-economic backgrounds (Louise Archer et al., 2012; Aschbacher et al., 2010). As a result, outcomes of self-report measures can be directly impacted by these factors; students of marginalized groups (e.g. race/ethnic minorities, language minorities, females) may self-report lower than peers from majority groups which may have impacted self-report scores of cognitive engagement.

As with any tool, self-reporting feature may result in limitations of use. Rios et al. (2014) suggest that self-reported measures may threaten validity due to the exaggeration or underrepresentation of self-reported effort. The exaggeration or reduction of reported error could be due to a lack of ability (Antin & Shaw, 2012), introducing threats to validity within a self-report measure an making accurate estimates difficult. The exaggeration of estimations as well as the underrepresentation of estimations is a common concern among self-report measures. Additionally, motivation within a test event may fluctuate across items (Wise & Kong, 2005) which would be difficult to measure with an summative self-reporting measure. This could be especially true with interactive items, adaptive items, or items that vary in subject matter or type.

Methodologically, the analysis does not account for the nesting of data between states, schools, grades, or classrooms. Additionally, no single approach to measurement of a construct is considered universally acceptable; therefore, there is a possibility researchers will select different behaviors to measure the same construct (Khairani & Razak, 2013), in this case, constructs of both cognitive engagement (and the five factors) as well as mathematics proficiency. Other challenges to measurement include practice effects, lingering effects, and

order effects. In an attempt to mitigate test effect, only the first performance task and subsequent engagement survey was used for analysis. Additionally, time of day assessment occurs could also lead to rapid guessing behavior and lack of motivation. The design utilizes a counter-balancing approach in an attempt to mitigate these concerns.

Using a CFA to fit a latent variable model should be re-evaluated in future studies. Results from the CFA did not show adequate fit across certain indicators and, therefore, results from the model should be interpreted cautiously. Future work could include the use of latent variable models such as confirmatory IRT for which a data set such as this would be better fitting to the assumptions of the model, including better allowing the utilization of categorical and non-normal data, when the correct estimation algorithms are employed within the latent variable setting. Although the mathematics performance task was connected to CCSS and written by content experts, the actual performance of the performance task items were not investigated as a part of this study. As a result, quantitative validity of the performance task items cannot be claimed.

Although standard alpha and beta levels were set for the analyses and appropriate effect sizes were used, there is still a possibility that a more robust finding may have occurred with a larger sample size, resulting in a Type II error. Additionally, there are factors that may cause the findings to be erroneous, specifically, the quality of the cognitive engagement instrument, design and delivery of assessment modalities, and research design. If a Type II error was not committed, results should be treated cautiously due to the study limitations.

Threats to the internal credibility of qualitative research include descriptive validity threats referring to the factual accuracy of the interviews, as documented by the researcher (Maxwell, 1992), although attempts were taken to include specific quotations to avoid spurious

interpretations. Additional research bias occurs when personal biases or a priori assumptions are present and potentially subconsciously transferred to participants (Onwuegbuzie, 2003) and, therefore should be treated as a possible threat to validity. Lastly, the novelty effect poses a threat to internal validity by participants potentially interpreting their participation in the interview as being given special considerations and, therefore, introducing a stimuli into the environment (Onwuegbuzie & Leech, 2007).

Qualitative external threats to validity include investigation validity referring to the personality traits and quality of craftsmanship of the researcher (Kvale, 1995) which can impact participant responses, comfort, demeanor, etc. As a researcher with a fast-paced and extroverted personality, these traits may have impacted participants. Lastly, Reactivity (novelty effect) can also pose a threat to external validity and limit the generalizability of results (Onwuegbuzie & Leech, 2007).

Additional qualitative threats include the order of modality which presents a limitation since all participants completed the performance tasks in a counterbalanced order (see Figure 2.1). Platform order was considered during the qualitative analysis; however, since the qualitative interviews occurred at the end of students completing all three modalities, the task that was completed first may have more favorable opinions expressed in comparison to the task that was completed last. This is particularly true since the repetitiveness of the items and modalities were noted negatively by students; while, students mentioned the excitement and unpredictability of the first task completed. This suggests that students favored the first task completed; thereby, indicating that students may have been most engaged during the first task completed. This result justifies the evaluation of the first task students completed, which is the framework used for the current study.

Lastly, there may be an inconsistency on the construct of measure between the qualitative and quantitative data. Quantitatively, measures aimed to understand cognitive engagement through measurement of deep processing, shallowing process, persistence, importance, and effort. Yet, qualitatively, students were asked questions about their overall attitude and favorability of tasks completed. This may have resulted in an inconsistency between what was measured qualitatively and quantitatively; with the quantitative data measuring cognitive engagement and the qualitative data measuring attitudes and enjoyment.

There are some final limitations that are specific to this study and data collection. To begin, some classrooms allowed students to use a mouse to respond while others used a trackpad. This could add in the extraneous variable of gross motor skill and may impact the way students answer questions or engage with the task. Further, one specific school (North Carolina) scheduled students to complete the performance task during their study hall class. This was seemingly viewed as a punishment by students and may have resulted in a lack of intrinsic motivation and disengagement from the onset of the study. Additionally, technical difficulties such as computer batteries dying or the online performance task experiencing difficulties rare such as missing text boxes or partially loaded may have impacted results (albeit, these difficulties were limited and rare). Finally, there is a concern around the generalizability of results due to controls that may be impossible to meet, specifically the use of intact groups instead of random assignment. The counterbalanced design attempted to control for this by randomizing assignment within existing groups. However, there is still a concern that the sample is still drawn from in-tact groups which may impact interpretation of results.

Validity Concerns

The study includes numerous concerns of threats to validity. To begin, the study introduces extraneous variables including fatigue, testing environment between settings, prior computer use, attitudes towards mathematics, attitudes towards technology, and prior knowledge. These variables are undesirable and could influence the relationship between the variables that the study is investigating (Hall, 1998). Additionally, these variables could introduce error within the experiment.

Internal validity threats. In addition to extraneous variables, the study also introduces internal threats to validity. Threats to internal validity include selection, testing, possible instrumentation, and possible design contamination. Because the sample is non-randomized, selection can be a threat to validity; students were in convenience samples (classrooms) and were not necessarily equivalent at the beginning of the study. Testing is a concern to internal validity because each student completes three assessments that have similar characteristics; therefore, the first assessment could cue the students how to approach and complete subsequent assessments. Lack of perceived novelty as the tasks proceed may undermine cognitive engagement measures. Also, instrumentation could be a threat to internal validity due to multiple proctors in the room, although proctors were prepared to allow students to proceed independently and all proctors received the same documentation on how to navigate and support the study. Design contamination could be a threat to internal validity due to students beginning the test event on various modes (paper-and-pencil, technology-enabled, technology-enhanced). Table 2.7 highlights the identified threats to internal validity.

Table 2.7

Internal Validity Threats

Threat type	Present
History	-
Maturation	-
Statistical regression	-
Selection	+
Experimental mortality	-
Testing	+
Instrumentation	+
Design contamination	+
Compensatory rivalry	-
Resentful demoralization	?

Note. + means threat to validity is present in the study; - means the threat to validity is not present in the study; ? means a possible threat to validity may exist.

External validity threats. The study also includes numerous threats to external validity, or the generalizability of results to other samples. Threats to external validity include reactive effects of experimental arrangements and multiple-treatment interference. The study risks reactive effects of experimental arrangements due to the fact that students know they are participating in an experiment (i.e. Hawthorne Effect). The study could also post threats to external validity due to multiple-treatment interference; students receive multiple treatments and risk a carry-over effect. Table 2.8 highlights the identified treats to external validity.

Table 2.8

External Validity Threats

Threat type	Present
Population	+
Ecological	-
Interaction effect of testing	+
Interaction effects of selection biases and the experimental treatment	+
Reactive effects of experimental arrangements	+
Multiple-treatment interference	+

Note. + means threat to validity is present in the study; - means the threat to validity is not present in the study.

Recommendations for Future Research

The results of this study pave the way for future research opportunities. In combination with previous literature, study results help to shape future work in the area of self-reporting of cognitive engagement within a STEM assessment context. Definitions of how to measure cognitive engagement have minimal consensus among researchers (Appleton et al., 2008). As a result, there are various components that are included as factors within these measures that are believed to contribute to a students' overall measure of cognitive engagement, as described in Chapter I. Better self-reporting tools may be needed to measure cognitive engagement, specifically during a low stakes assessment context for grades K-12. As well, a better understanding of what is being measured by an overall cognitive engagement composite such as shown here, especially in the absence of reliable subscales, is important to consider. Theoretical subscales should be strongly considered for inclusion of factors aimed to measure the construct of cognitive engagement. Future work should aim to improve upon existing tools, such as the CE-S-DSP & SOS, or to create and validate new measures of cognitive engagement.

There are many changes to the study design that could be considered for future research in the technology-enhanced area for assessments. Due to the implications that engagement has on student affective and performance outcomes, it is important to include a sample that is as representative of the population as possible, which was not possible to do fully here given the schools in which the technology-enhancements were being employed. This is particularly important for examination in STEM fields, as considerable literature indicates that STEM is failing to engage students (Aschbacher et al., 2010; Lyons & Quinn, 2010). Additionally, the sample should be representative of marginalized groups such as racial/ethnic diversity, SES, geographic location, inclusion of student who are receiving special education or English as a second language services. This is essential in STEM fields, as marginalized populations are less likely to hold high STEM aspirations and pursue STEM fields.

In addition to an inclusion of a representative sample, study design should also be re-considered. The counter-balanced design of the study was created to randomize the order of condition and to allow all students the chance to complete all conditions and avoid the internal threat of compensatory rivalry. As a result, the study aimed to evaluate within-student cognitive engagement across platforms; however, once the study began it became obvious that engagement was decreasing due to duration and repetitiveness. If evaluated, this could have resulted in threats to internal validity. Consequently, student cognitive engagement was only evaluated for the first performance task modality they were assigned, despite the fact they completed all three modalities. This presented issues when evaluating qualitative data, as interviews occurred at the completion of all three tasks such that parsing out attitudes became difficult. Future studies should include a larger sample and randomly assign students to one condition, if possible, in order to avoid additional threats to validity.

Although the current study aimed to create three different assessment modalities utilizing the same performance task, those differences between platforms may not have been as when delineated in their affordances as originally intended. Future work in this area should implement more advanced technology-enhancements utilizing animation and technology-enhanced items. This will help to establish a strong delineation between the technology-enhanced task and the technology-enabled task. Similarly, extraneous factors between platforms should be controlled, if possible, particularly factors that students reported as favorable. One example would be to allow for navigation within the technology-based platforms, if the item/task design allows for navigation. This will help to ensure that the particular asset is maintained between paper-and-pencil and technology-based modalities. Another example would be to control for screen brightness (or explicitly show students how to change screen brightness) on the computers being used in the study. Changes like these help to ensure the independent variable of measure is modality-type and that the variable is not confounded with extraneous variables that could impact how students self-report cognitive engagement.

Future studies should consider implementing changes to the performance task itself. Item content should be assessed for difficulty level to ensure appropriateness for the target grade. This will help to ensure that student engagement is not negatively impacted because the item is too difficult. Additionally, future work should employ a performance tasks that can be scored for all students in order to better measure the validity and reliability of the items.

Methodological improvements should be factored in to future study considerations. The current study utilized a CFA to measure the latent variable of cognitive engagement across five factors. This was consistent with what done with previous validation studies on self-report cognitive measures (Miller et al., 1996) and more advanced than other validation approaches

(Smiley & Anderson, 2011; Thelk et al., n.d.). However, methodology can continue to be improved in future studies. The CFA did not show adequate fit for the data and, therefore, results from the model warranted cautious interpretation. Future work could include the use of latent variable models such as confirmatory IRT for which a data set such as this would be better fitting to the assumptions of the model, including better allowing the employment of categorical and non-normal data, when the correct estimation algorithms are employed within the latent variable setting. In addition to changes in statistical models, future studies should consider evaluating subscales of the cognitive engagement model in an attempt to improve model fit.

Effect on students' use of technology has been a strong focus in the literature across numerous fields and with mixed outcomes. This is an area that needs to be further explored to better understand how technology use (at home) impacts student outcomes. Future research may consider, theoretically and empirically, how to better collect data about technology use at home (including program usage and amount of time) to further investigate the effect of home computer use on cognitive engagement and other affective and academic student outcomes.

Future research is needed to help better understand the literature on student cognitive engagement within an assessment context, particularly in STEM. The research design could include a quasi-experimental design implementing randomization approaches and taking into consideration the future research needs and limitations discussed in this study. Through careful planning, implementation, and data collection, results may help further inform how technology affordances, computer use at home, and assessment modalities may impact student engagement outcomes.

Conclusions

There are many self-reporting tools that aim to measure student engagement and motivation. Few tools exist to validly and reliably measure cognitive engagement within a formative assessment context. The survey used in this study (CE-S-DSP & SOS) was an amalgamation of specific theoretical factors of previously validated measures. As an overall measure of cognitive engagement, the CE-S-DSP & SOS survey showed acceptable reliability, as evidenced through previous validation of the separate components of the instrument, and empirical results here. However, individual factors of cognitive engagement within the survey did not result in acceptable reliability thresholds in all cases, and some of the sub-strands likely require more information to report valid sub scores, if that were to be the intent. Additionally, analysis using this data set indicated that the survey was not ideal for using a CFA and might be better explored with latent variable models such as confirmatory IRT. Additionally, further exploration of theoretical factors that measure cognitive engagement may help explore additional constructs to consider when developing new self-reporting tools. In doing this, we can begin to create more robust and reliable self-reporting tools for measuring cognitive engagement within K-12 formative assessment, therefore, resulting in a more valid measure. As a result of a valid and reliable tool, we can begin to better measure the variables explored in this study.

Although previous literature suggests there may be differences in engagement by demographic factors such as sex, race, and ethnicity, results from the current study did not indicate a significant difference of engagement by sex or race/ethnicity when measuring cognitive engagement within a mathematics performance task across modalities. Although this is somewhat inconsistent with previous research, the range of possible applications is large and the literature exploring cognitive engagement within a STEM-based formative assessment

remains largely unexplored. Future work would benefit from replication with an extended measure in order to better understand what is happening with the data, and also exploring the degree to which the task itself fully required and afforded the use of the technology enhancements.

The benefits and drawbacks of each modality were discussed within the interview subsample. It is clear there are benefits to each assessment modality. Students enjoy taking assessments on a computer, but the platforms in this case had limitations. Additionally, there are many reported benefits to paper-and-pencil assessments that have the potential to be lost when transferred to technology-based platforms. For example, students reported appreciating the ability to navigate back and forth on a paper-and-pencil assessment but that feature was not employed on the technology assessments. This is an extraneous variable that may not be intended to have an effect on the platform utility; therefore, design of a future technology-based assessment may include the use of backwards navigation in order to keep the feature consistent between platforms. There are some features that will not be possible to maintain consistency (e.g. animation) and this should be noted as a feature of a specific modality. Consideration of these extraneous variables will ensure an equal comparison of modalities without benefits of one modality outweighing the omission of those same benefits on another modality. However, the resources to implement the comparable affordance should also be considered, as they may enter into the question of whether such affordances ultimately are employed in schools for formative assessments or eliminated.

Additionally, the inclusion of computer-based platforms (one technology-enhanced and one technology-enabled) should result in different features that benefit the platform (while controlling for extraneous variable). For example, if the feature of technology-enhanced is to

include animations and technology-enhanced items, those items should perform well and serve the purpose intended. Animations that are less optimal may not result in the appropriate enhancements needed to differentiate the platform from the technology-enabled.

Although technology use at home did not produce meaningful results, it is important to consider this variable for future investigation. The use of technology at home is increasing; students are spending more time on technology than ever before. Understanding how home technology use (time and program usage) impacts student outcomes is another factor that will help shape how parents and educational leaders communicate about technology-usage and implement interventions in the classroom.

In conclusion, it can be said that as technology enhancements advance, the use of technology is becoming more widely used within educational contexts. This use is drastically increasing within educational instruction and assessment; yet, literature on how this increase impacts students' academic and non-academic outcomes remains sparse and somewhat inconsistent. Additionally, there is minimal literature on assessment modality and technology-enhancements. Results from this study imply that modality (type of assessment platform and technology-enhancements employed) as a whole rather than inspected on a detailed-bases for what is actually being supported may not impact student engagement as much as some research suggests. As mentioned earlier, modality type should be further explored to parse out the benefits and drawbacks of each modality in an attempt to control for these differences within the design of each modality. This may help to mitigate the possible effect of extraneous variables in order to evaluate the effect of type of modality. Better understanding how the use of technology (modality and features) impacts student outcomes will help shape how technology is used within education.

The overall results from this study indicate that there is not a significant quantitative relationship between cognitive engagement and modality type, demographic characteristics, grade level, or computer use at home. However, a more in-depth qualitative exploration of students attitudes on using technology to take tests as well as overall favorability of tasks completed indicated that students preferred using technology to take tests in comparison to paper-and-pencil assessments. Additionally, the computer-enhanced task was the most preferred. Despite these technology accolades, however, there were features of the paper-and-pencil task that students appreciate such as the ability to take notes and the ability to navigate back and forth between items. As a result, drawbacks of platforms (e.g. not being able to navigate within the technology-based tasks) may have outweighed the reported benefits.

Results from this study indicate that just putting tests on computers may not be enough. Although students do appreciate using technology to take tests, the technology must be implemented well and purposefully. Additionally, it is important to be thoughtful of the technology affordances that are employed within an assessment. There are features of paper-and-pencil tasks that students appreciate that may be lost when transferred to a computer. It is important to consider inclusion or exclusion of these features and ensure consistency if appropriate. Finally, it is important to slow down and consider how and when to use technology within assessments. Although the market demand may be shifting the focus to technology-based assessments, including advanced scoring algorithms, the fast shift may impact the authenticity of the assessment and limit the ability to measure more performance-based outcomes. It is important to consider whether constructs can be measured using the technology modality and whether features of the paper-and-pencil task are a hindrance to student performance and affective outcomes when transferred to technology.

Footnotes

[1] RIT scale corresponds to Rasch-Unit and is expressed as $P_{ij} = \frac{e^{(\theta_j - \delta_i)}}{1+e^{(\theta_j - \delta_i)}}$.

[2] 40 states include AK, AR, AZ, CA, CO, DE, GA, IA, ID, IL, IN, KS, KY, MA, ME, MI, MN, MO, MT, NC, ND, NE, NH, NJ, NM, NV, NY, OH, OK, OR, PA, RI, SC, TX, UT, VA, VT, WA, WI, and WY.

[3] Spring 2008-Fall 2008 states include all states from Footnote 2 except for ME, MO, NM, RI, UT as well as the addition of HI.

[4] Fall 2008-spring 2009 states include all states from Footnote 2 except for MT and NV as well as the addition of CT, HI, LA, and MS.

[5] Spring 2008-Spring 2009 states include all states from Footnote 2 except MT and NV as well as the addition of CT, LA, and MS.

[6] Spring 2008 to Fall 2009 states include all states from Footnote 2 with the addition of FL, MT, NV, and TN.

[7] States include AZ, CA, CO, DE, GA, IL, IN, KS, MA, ME, MI, MN, MT, ND, NH, NJ, NM, NW, OH, OR, RI, SC, TX, WA, WI, WY.

[8] Twelve states were included in the measurement of concurrent validity including AR, CA, CO, FL, GA, KY, ND, NC, PA, SC, WI, and WY.

[9] Nine states were included in the measurement of predictive validity including AR, CA, CO, FL, GA, NC, ND, PA, and SC.

[10] Seven states were included for the measurement of criterion-related validity for sixth and seventh grade (AR, CA, CO, FL, GA, KY, SC), six states were included for the measurement of criterion-related validity for eighth grade (CA, CO, FL, GA, KY, SC).

[11] Note that the design in Figure 2.1 does indicate a potential violation to some degree of independence for the ANOVA tests, since the same students are used in each treatment condition through the rotated design. The rotated design is typically used in software development to balance treatment conditions. Limitations of the design are discussed at the end of chapter.

[12] Subject-matter experts include NWEA™ senior mathematical content specialists.

APPENDIX A

CE-S-DSP

Subscale 1

Deep strategy (understanding mathematical concepts)

When approaching the questions in (test name here)…

…I drew pictures or diagrams to help me solve some problems

…I considered problems already finished to help me figure out how to solve Similar problems

…I analyzed the problems to see if there was more than one way to get the right answer

…I stopped to ask myself whether or not I understood the items

Shallow processing strategy (rote memorization)

When approaching the questions in (test name here)…

…I considered how those reviewing the answers would want me to respond*

…I looked for clues of how to respond with the test itself*

…I tried to memorize the steps for solving the problem throughout the test

Subscale 2

Persistence

When I ran into problems on items within (test name here)…

…I went over the item(s) until I understood what the question was asking me

…I guessed at the answer rather than working through the problem (R)

…I kept working at it until I thought I solved it

…I gave up and went on to the next problem (R)

4-Item Likert Scale

Strongly Disagree Disagree Agree Strongly Agree

Cognitive Engagement Survey S - DSP. Adapted from Miller, R. B., Greene, B. a., Montalvo, G. P., Ravindran, Bhuvaneswari, & Nichols, J. D. (1996). Engagement in Academic Work : The Role of Learning Goals, Future Consequences, Pleasing Others. *Contemporary Educational Psychology*, *21*(4), 388–422.

APPENDIX B

STUDENT OPINION SCALE (SOS)

Subscale	Items
Importance Definition: How important doing well on the test is to the student	1. Doing well on this test was important to me. 3. I am not curious about how I did on this tests relative to others. 4. I am not concerned about the scores I receive on this test. 5. This was an important test to me. 8. I would like to know how well I did on this test.
Effort Definition: The reported level of effort and persistence expended toward test completion.	2. I engaged in good effort throughout this test. 6. I gave my best effort on this test. 7. While taking this examination, I could've worked harder on it. 9. I did not give this test my full attention while completing it. 10. While taking this test, I was able to persist to complete the tasks.

Student Opinion Survey. Adapted from (Sundre, 1999)

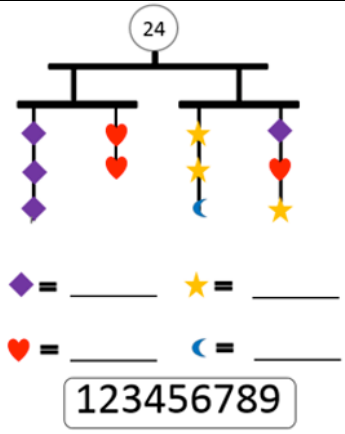
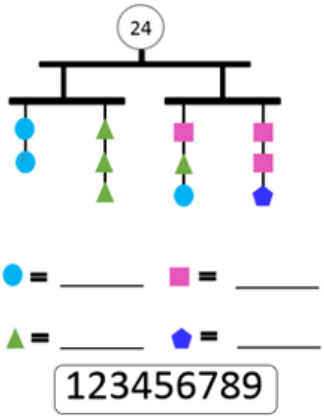
APPENDIX C

MODALITY COMPARISONS

	Paper-and-Pencil	Technology-Enabled	**Technology-Enhanced**
Math Task CCSS 7.EE.3	Yes	Yes	Yes
Avatar	Yes	Yes	Yes
Animated Avatar	No	No	Yes
Interactivity of Items based on student responses	No	No	Yes
Gamified Text Prompts	No	No	Yes
Gamified Environment	No	No	Yes
3-D appearance	No	No	Yes

APPENDIX D

PERFORMANCE INSTRUMENT FOR EACH MODE

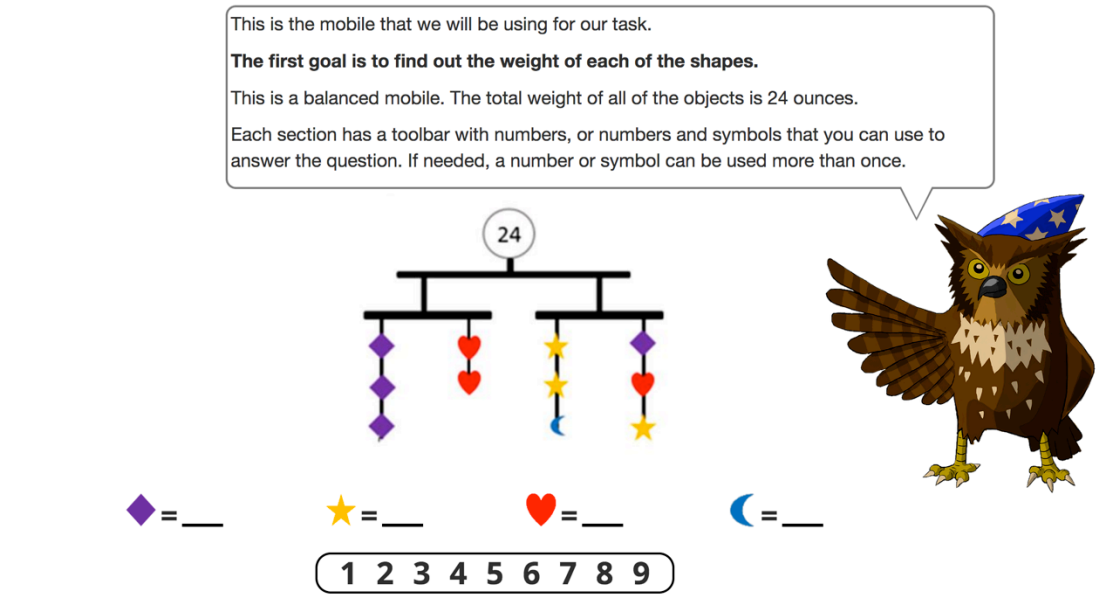
Technology-enhanced	Technology-enabled	Paper-and-pencil
		

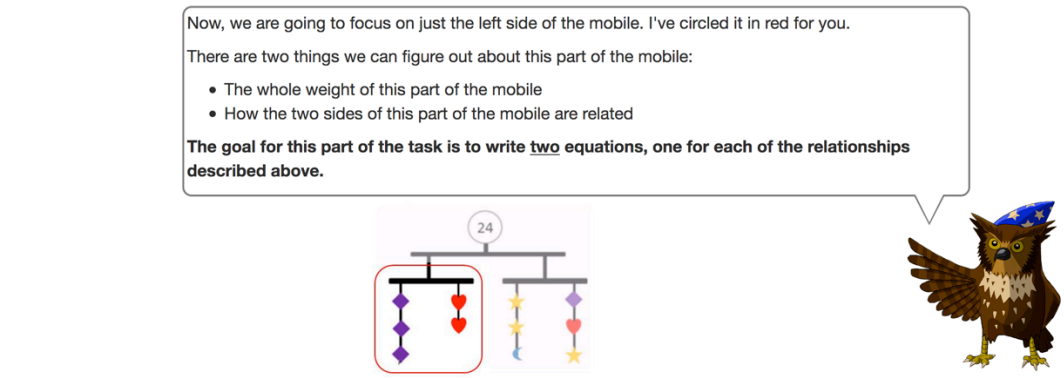
Note: Technology-enabled and paper-and-pencil modes look like the above images. The technology-enhanced mode utilizes a different design. Designs are located in Appendix E.

APPENDIX E

SCREEN SHOTS OF PEROFRMANCE TASK

Technology-enabled design





Technology-enhanced design

APPENDIX F

PERFORMACNE TASK SCORING RUBRICS

Question 1 – Find out the weight of each of the shapes				
Criteria	Points	DOK	This student demonstrates	Possible responses
Full balance: Student gets all 4 values correct	8	3	The student understands the need to balance all parts of the mobile. The student understands all of the pieces sum to 24.	triangle = 2 circle = 3 square = 1 pentagon = 4
Full balance sum NOT 24: The student provides values that result in the left side balancing, the right side balancing, the left and right sides balance to one another BUT the values do NOT sum to 24	4		The student understands the need to balance all parts of the mobile. The student does NOT understand all of the pieces sum to 24.	triangle = 4 circle = 6 square = 2 pentagon = 8
Completely incorrect	0		The student demonstrates guessing behavior.	

Question 2 – Can you tell me a little about what you did to find the answers?				
Criteria	Points	DOK	This student demonstrates	Possible correct responses
Correctly identifies which strategies were used to solve the problem (formal or informal)	1	1	Ability to identify strategies used to solve the problem (formal or informal).	Examples: guess and check, logic, equations
Student does not identify strategies used to solve the problem.	0		Strategies used to solve the problem are not identified.	Examples: blank response, “I’m not sure”, or another answer that does not identify a strategy.

Question 3 – The next goal is to write an equation to describe the relationship between the shapes on the left side of the mobile and the shapes on the right side of the mobile.				
Criteria	Points	DOK	This student demonstrates	Possible correct responses
Equation is correct and combine like terms	7	3	The student demonstrates an understanding of the relationship between a balance and a balanced equation and combines like terms.	3diamond+2heart=3star+1moon+1diamond+1heart Diamond+diamond+diamond+ heart+heart = star+star+star+moon +diamond+heart Notation indicating same thing may be used. Example: 3diamond is the same as 3 x diamond Any variation of 3diamond+2heart on one side of the equation with 3star+1moon+1diamond+1heart on the other side of the equation is acceptable. However indicating variable first followed by the coefficient such as diamond 3 + heart2 = star3+moon1+diamond1+heart1 would **NOT** be correct.
Equation is correct and does NOT combine like terms	6		The student demonstrates an understanding of the relationship between a balance and a balanced equation.	3diamond+2heart=2star+1moon+1diamond+1heart+ 1star Diamond+diamond+diamond+ heart+heart = star+star +moon +diamond+heart+star Notation indicating same thing may be used. Example: 3diamond is the same as 3 x diamond Any variation of 3diamond+2heart on one side of the equation with 2star+1moon+1diamond+1heart+1star on the other side of the equation is acceptable. However indicating variable first followed by the coefficient such as diamond 3 + heart2 = star2+moon1+diamond1+heart1+star1 would **NOT** be correct.
All values correct but operators on one side of the equation incorrect	5		The student demonstrates a partial understanding of the relationship between a balance and a balanced equation by correctly representing the information on one side of the equation	3diamonds+2hearts OR 3stars+1moon+1diamond+1heart
All values correct and NO operators	4		The student demonstrates a partial understanding of the relationship between a balance and a balanced equation, but doesn't communicate the equation formally.	3diamond 2heart= 3star 1moon 1diamond 1heart Diamond diamond diamond heart heart = star star star moon diamond heart Diamond diamond diamond heart heart = star star moon diamond heart star
One side (only) of the equation correct	3		The student demonstrates limited understanding of the relationship between a balance and a balanced equation	
Completely incorrect	0		The student demonstrates guessing behavior.	

Question 4 – Write two equations, one for each of the relationships displayed on the left side				
Criteria	Points	DOK	This student demonstrates	Possible correct responses
Both equations are correct	6	3	The student demonstrates an understanding of both sides being in balance (2 or fewer variables), AND uses that information to deduce that the left side (Balance 2) of the entire mobile must be 12.	3 diamond = 2 heart and 3 diamond + 2 heart = 12 Diamond+diamond+diamond= Heart+heart and Diamond+diamond+diamond + Heart+heart =12 Order of coefficient/variable pair does not matter (e.g. 3 diamond = 2 heart and 2 heart = 3 diamond would both be correct). Nor does it matter which equation is identified first. However indicating variable first followed by the coefficient such as diamond 3 = heart 2 would NOT be correct.
Correctly identifies 3 diamond = 2 heart But sums to 24 (i.e. Identifies 3 diamonds + 2 hearts = 24)	5		The student demonstrates an understanding of both sides being in balance (2 or fewer variables). The student does NOT understand ALL of the variables sum to 24.	
Correctly identifies 3 diamonds = 2 hearts Does NOT identify either 12 or 24 as possible sum Identifies 3 diamonds + 2 hearts = value other than 12 or 24	4		The student demonstrates an understanding of both sides being in balance (2 or fewer variables). Student does not demonstrate understanding the right side (Balance 2) of the mobile must be 12.	
Does not identify 3 diamonds = 2 hearts Does not identify 3 diamonds + 2 hearts = 12 or 24	0		The student does not demonstrate an understanding of how balances represent equality.	

Question 5 – Write two equations, one for each of the relationships displayed on the right side				
Criteria	Points	DOK	This student demonstrates	Possible correct responses
Both equations are correct and combine like terms	7	3	The student demonstrates an understanding of both sides being in balance (3 or more variables), uses that information to deduce that the right side (Balance 3) of the mobile must be 12, and combines like terms.	2 star+1moon = 1 diamond + 1heart + 1star and 3 star + 1 moon + 1 diamond + 1 heart = 12 Star+star+moon= Diamond+heart+star and Star+star+star +moon+diamond+ heart =12 Order of coefficient/variable does not matter (e.g. 2 star+1moon = 1diamond + 1heart + 1star and Star+moon +star = heart+Diamond+ star would both be correct). Nor does it matter which equation identified first. However indicating variable first followed by the coefficient such as star2+moon1 = diamond1+heart1+star1 would **NOT** be correct.
Both equations are correct, like terms NOT combined	6		The student demonstrates an understanding of both sides being in balance (3 or more variables), AND uses that information to deduce that the right side (Balance 3) of the mobile must be 12.	2 star+1moon = 1 diamond + 1heart + 1star and 2 star + 1 moon + 1 diamond + 1 heart + 1star= 12 Star+star+moon= Diamond+heart+star and Star+star +moon+diamond+ heart +star =12 Order of coefficient/variable does not matter (e.g. 2 star+1moon = 1diamond + 1heart + 1star and Star+star+moon= Diamond+heart+star would both be correct). Nor does it matter which equation identified first. However indicating variable first followed by the coefficient such as star2+moon1+ diamond1+heart1+star1=12 would **NOT** be correct.
Correctly identifies 2 stars+1moon = 1 diamond + 1heart + 1star But sums to 24 AND combines like terms Identifies 3 stars + 1 moon + 1 diamond + 1 heart = 24	5		The student demonstrates an understanding of both sides being in balance (3 or more variables) and combines like terms. The student does NOT understand ALL of the variables sum to 24.	

Correctly identifies 2 stars+1moon = 1 diamond + 1heart + 1star But sums to 24 AND does NOT combine like terms Identifies 2 stars + 1 moon + 1 diamond + 1 heart + 1star= 24	4		The student demonstrates an understanding of both sides being in balance (3 or more variables The student does NOT understand ALL of the variables sum to 24	
Correctly identifies 2 stars+1moon = 1 diamond + 1heart + 1star Does NOT identify either 12 or 24 as possible sum Identifies 3 stars + 1 moon + 1 diamond + 1 heart = value other than 12 or 24	3		The student demonstrates an understanding of both sides being in balance (3 or more variables). Student does not demonstrate understanding that this means the right side (Balance 3) of the entire mobile must be 12.	
Does not identify 3 diamonds = 2 hearts Does not identify 3 diamonds + 2 hearts = 12 or 24	0		The student does not demonstrate an understanding of how balance represents equality.	

APPENDIX G

STUDENT INTERVIEW

General Probes:

"what you're saying is...", "anything else you can think of?", "MmmmHmm…", "Go on...", "Interesting, what else can you think of?", "tell me more about that", "thank you for that information, can I now ask you about-------"

1) What are you currently working on in school?

Follow up: Talk to me about how you use computers in school.

2) What is your opinion about using computers to take tests?

Prompt: Tell me about what you like most about using computers to take tests.

Prompt: What do you like least about using computers to take tests?

3) If you explained to a friend about the performance tasks you just took, what would you say to them / what would you tell them?

Follow up: What words would you use to describe this performance task (show technology-enhanced)?

Follow up: What about this performance task (show paper-and-pencil)?

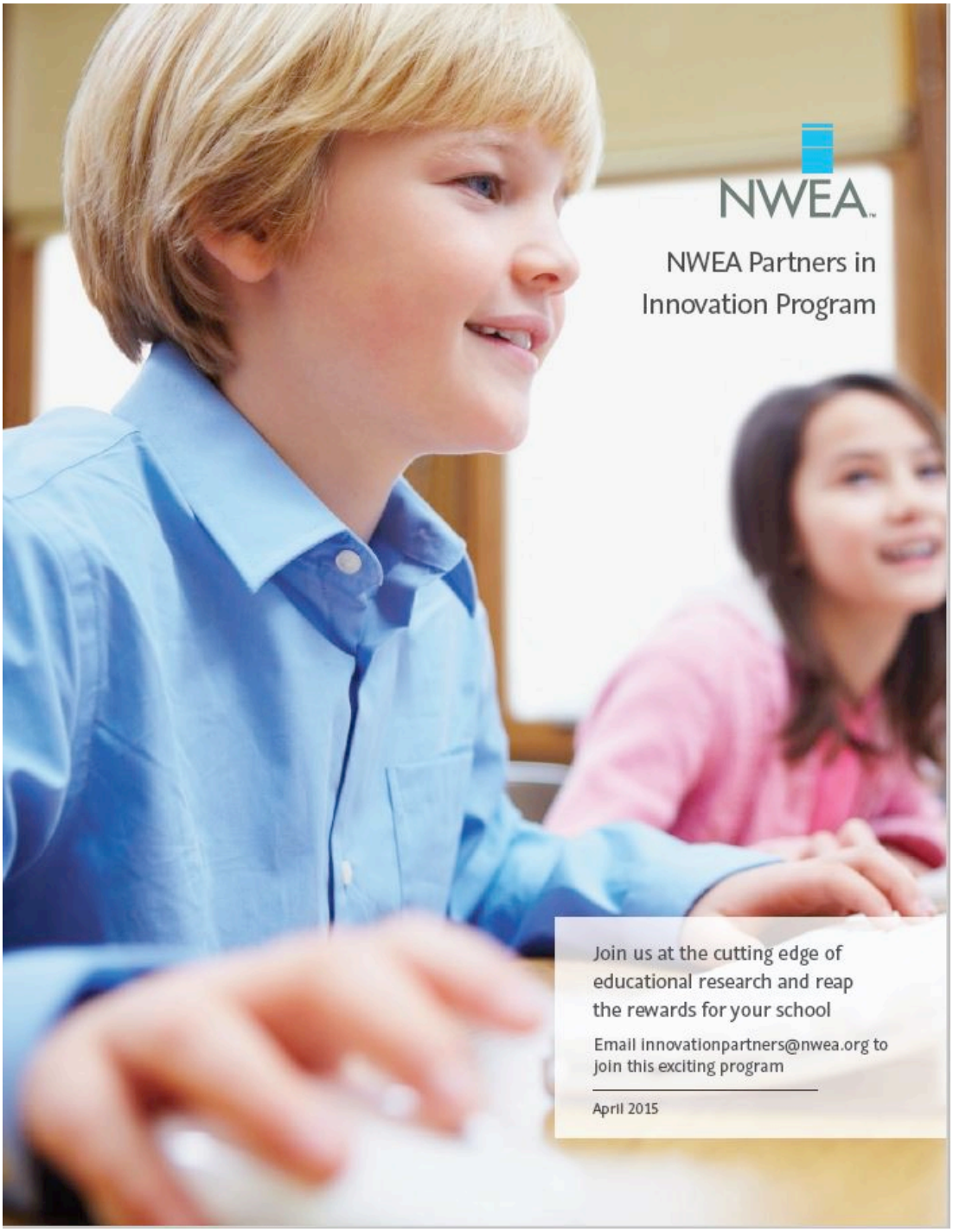
4) Put these three images in order of which you liked using most and which you liked using least and explain to me why you're putting them in that order. (show screen shots)

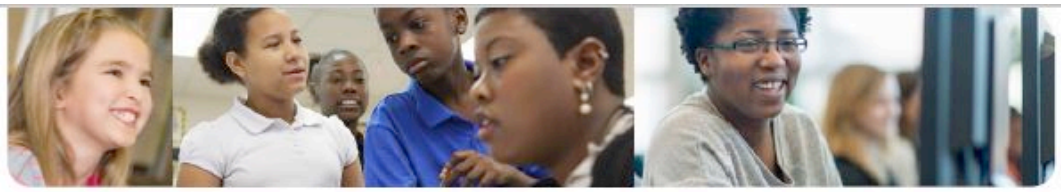
Prompt: Why did you chose this one (point to choice) as your most favorite?

Prompt: Why did you chose this one (point to choice) as your least favorite?

APPENDIX H

PARTNERS IN INNOVATION FLYER





Partners in Innovation

Partnership is at the core of what we do at Northwest Evaluation Association™ (NWEA™). That's why we're inviting schools to join us on exciting new ventures into cutting-edge educational research. The NWEA Partners in Innovation program allows participating schools to join us in testing new assessment approaches, trying out the latest proprietary technology, and exploring new frontiers of educational research. Working with our Partners in Innovation allows us to gain crucial insights, improve our products, develop new approaches, and further our mission of Partnering to help all kids learn®. And you can join this endeavor to meet the evolving demands of educating children!

Partnership has big benefits for your school

Join the Partners in Innovation program and reap the rewards of an exciting research relationship.

- An annual stipend for your school...up to $10,000!
- Gift cards and other rewards for participating educators
- Training and support for teachers, students, and administrators
- Use of NWEA technology for carrying out research initiatives
- The excitement of discovering new educational approaches and assessment solutions
- An ongoing dialogue among NWEA staff and researchers, educators, students, and parents

Making cutting-edge discoveries...together

Work with us to develop innovative approaches to student assessment, professional development, data delivery, and technology implementation. Partners will have an active role in choosing and implementing studies at their schools, making every research effort flexible, collaborative, and valuable for everyone. Together, we'll explore topics like

- new approaches to student engagement in learning and assessment
- novel ways to observe and evaluate student performance
- inventive ideas for assessment tasks
- engaging and effective professional development
- strategies for parental engagement in the educational process

Partnering to Help All Kids Learn® | NWEA.org | 503.624.1951 | 121 NW Everett St., Portland, OR 97209

NWEA

Frequently Asked Questions

1. How does the Partners in Innovation program differ from other NWEA research initiatives, such as field testing?

These studies are much smaller and more targeted than the large-scale field testing studies we perform, in which we ask an entire school population to participate. The Partners in Innovation program is a full collaboration between NWEA and participating schools, with educators taking an active role in choosing and implementing research studies.

2. How long is the commitment, and how many studies are involved?

You will commit to participating in the program for a minimum of one calendar year. Throughout the calendar year, you'll be asked to participate in two to three studies.

3. What types of research studies will I participate in?

We'll collaborate on a wide range of research topics, from assessment tasks and performance observation to professional development and student engagement studies. Whatever we're exploring, it's sure to uncover exciting new ground.

4. How many students and teachers will participate in each research study?

The scope of each study will vary, but only a limited number of your staff and students will be required for any given project.

5. How much time will each research study take?

We anticipate each research study to require no more than four hours of teacher time and no more than three hours of student time total.

6. What happens with the results of the research studies?

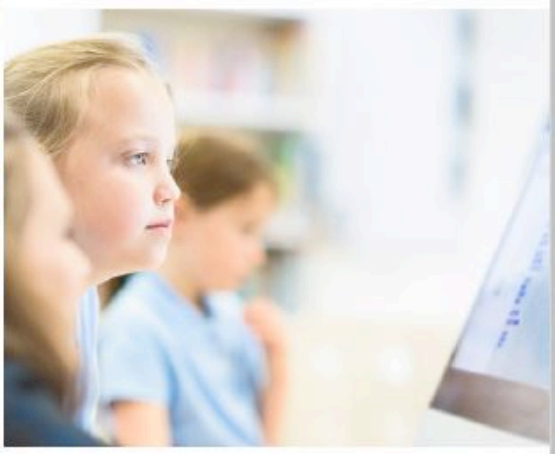
In addition to these results fueling better assessment solutions from NWEA, the results of each study will be shared with your school. Educators, students, and parents will receive reports and presentations detailing the results, which will provide insights for your own education initiatives.

7. How much can my school earn by participating? What can we use the money for?

Annual stipends range from $3,000 to $10,000, depending on the size of your school. The stipend is for your school to spend any way you choose.

8. How can my school become a Partner in Innovation?

We'd love to talk with you about participating in this exciting venture. Email innovationpartners@nwea.org to start the conversation.



NWEA

Join us now for exciting new discoveries. Email innovationpartners@nwea.org to find out how you can become a Partner in Innovation

Founded by educators nearly 40 years ago, NWEA is a global not–for–profit educational services organization known for our flagship interim assessment, Measures of Academic Progress (MAP). More than 7,400 partners in U.S. schools, school districts, education agencies, and international schools trust us to offer pre-kindergarten through grade 12 assessments that accurately measure student growth and learning needs, professional development that fosters educators' abilities to accelerate student learning, and research that supports assessment validity and informed policy. To better inform instruction and maximize every learner's academic growth, educators currently use NWEA assessments with nearly 8 million students.

April 2015

APPENDIX I

PARTNERS IN INNOVATION LEGAL CONTRACTS

[DATE]

[SCHOOL]

Re: Northwest Evaluation Association Research Studies

Dear [SCHOOL CONTACT]:

Northwest Evaluation Association ("NWEA") develops web-based assessment products and instructional material that allow educators to instruct, monitor and assess students and their educational progress. NWEA frequently works with educators and institutions to gain better insight into how its products and services (our "Solutions") are used, with the goal of using information gathered to help improve our Solutions and to develop new ones.

The [SCHOOL] ("School" or "you") has expressed an interest in becoming involved in creating and shaping assessments and curriculum tools that use and leverage cutting edge technology to meet the evolving demands of educating children. To that end, for one year from the date of this letter, we wish to collaborate with you and certain members of your school personnel to conduct research studies involving our Solutions, including field testing of certain test items (the "Studies") as described below and in the attached Project Plan.

NWEA and the School have agreed on the following terms with respect to the Studies:

A. Conducting Certain Studies

In order to compile information for certain Studies, NWEA may record the interactions that users have with the Solutions being evaluated, including through photo, audio and video recordings as well as through interviews, electronic and written surveys, questionnaires, assessments, collected student work, and direct observation of users. NWEA may use this information in internal and public reports, as well as to improve and promote NWEA's products and services.

You have agreed to facilitate reasonable access to the School and to selected School personnel and students to conduct the Studies, and to assist us in distributing and obtaining any requested consent forms or releases (such as the forms attached). In addition to any NWEA personnel assisting with the Studies and any observers invited by the School, you have agreed that we may invite a limited number of observers to be present for portions of the Studies. Further, you acknowledge and agree that no test scores will be generated or shared with you since these are tryout tests and items.

You have further agreed to share with NWEA assessment and demographic data for all your students in the School for the 2014–2015 school year (the "Assessment Data"). It is expected that such data from students not using the Solutions will be de-identified by School prior to providing it to NWEA. We adhere to applicable privacy standards with respect to personally identifiable information ("PII"), as such term is defined under the Family Educational Rights and Privacy Act ("FERPA"). Reports from the Studies will not contain any PII. Studies will be reviewed by NWEA's Institutional Review Board (IRB) to ensure that they meet the standards associated with educational research.
NWEA may also retain third-party vendors to assist with the collection of information and/or facilitate other aspects of the Studies. Such third-party vendors are not parties to this letter agreement. They may be subject to their own respective privacy policies posted on their respective websites, or in the absence of such a policy, NWEA may separately agree with them to treat any PII they may collect in accordance with this agreement.

We appreciate the accommodations you will be making to assist us in conducting the Studies. As an incentive, we will provide a stipend to the School of $[FULL AMOUNT] for the Studies, and we will provide gift cards to teachers and administrators for their participation in the Studies. Upon full execution of this letter agreement, NWEA will pay the School 50% of the stipend $[HALF OF AMOUNT], with the remainder being paid no later than six (6) months from the date of this letter agreement. You have indicated that these contributions and incentives are consistent with applicable policies.

B. General Terms

Each party (for purposes of this paragraph, the "Receiving Party") acknowledges that, in connection with this Studies, the other party (the "Disclosing Party") has provided and will provide to it certain sensitive information that is either subject to the Disclosing Party's confidentiality obligations to third parties or if obtained by competitors of the Disclosing Party's could damage the Disclosing Party's business, including without limitation, inventions, research, designs, methods, processes, customer lists, training materials, documentation, know-how and trade secrets, in whatever form, as well as the terms of this Agreement ("Confidential Information"). NWEA and the School agrees (a) not to use the Confidential Information of the Disclosing Party for any purpose other than in connection with the Studies; and (b) to take all steps reasonably necessary to maintain and protect the Confidential Information of the Disclosing Party in the strictest confidence. Confidential Information shall not include information which: (i) is as of the time of its disclosure or thereafter becomes publicly available through no fault of the Receiving Party; (ii) is rightfully known to the Receiving Party prior to

the time of its disclosure; (iii) has been independently developed by the Receiving Party without use of the Confidential Information; (iv) is subsequently learned from a third party not under a confidentiality obligation; or (v) is required to be disclosed by law or judicial order.

You understand that, because the Studies may involve NWEA's Confidential Information, you will allow for a reasonable period of consultation with us prior to making any public statements or responding to any media inquiries regarding the Studies.

Finally, you understand that either the School or NWEA has the right to terminate the Studies upon 30 days written notice to the other.

If you agree that the foregoing reflects our mutual understanding regarding the Studies, kindly confirm by signing the enclosed copy of this letter and returning it to my attention.

Sincerely,

_____

Jeffrey P. Strickler, Executive Vice President and COO

Enclosures

AGREED AND ACCEPTED:

_____

[CONTACT NAME], [CONTACT TITLE]

APPENXIX I.1

PARENT CONSENT FORM

September 2016

Dear Parent / Guardian,

Your child is invited to participate in a research study conducted by the Advanced Research and Development team from NWEA. We hope to learn how technology enhanced performance tasks impact student engagement. Your child was selected as a possible participant in this study as a result of (school name) lab school agreement with Northwest Evaluation Association (NWEA) to participate in new research projects and studies.

If your child participates, they will complete a series of 3 performance tasks across different modes (paper/pencil, technology-enabled, technology-enhanced). They will also be asked to complete a 20-item survey asking questions about their engagement across the three tasks and modes. The complete study should take approximately 60 minutes. A random sample of students will be selected to participate in a 5-10 minute interview to discuss likes and dislikes about technology in assessment. There are no known risks to your child for participating in the study. Benefits to your child for participating in the study include the opportunity to engage in new technology and assessment design; however, we cannot guarantee that you or your child will personally receive any benefits from this research.

Any information that is obtained in connection with this study and that can be identified with your child will remain confidential and will be disclosed only with your permission. Your child's identity will be kept confidential by utilizing a unique ID number instead of student names. All data will be stored on a secure server and NWEA researchers directly involved in the study will have access to only de-identified data. The de-identified data may also be available to other researchers.

Your child's participation is voluntary. Your decision whether or not to let your child participate will not affect your relationship with (school name) or NWEA. If you decide to allow your child to participate, you are free to withdraw your consent and discontinue your child's participation at any time without penalty.

If you have any questions, please feel free to contact Meg Guerreiro, Sr. Research Associate at NWEA, at (503) 444-6435 or meg.guerreiro@nwea.org

Please contact Meg Guerreiro to opt your child out of participation in the study.

Thank you

APPENDIX I.2

STUDENT ASSENT FORM

Performance Task Study

Your Mother/Father/Guardian has given permission for you to participate in a performance task study. This means you are able to:

Be observed at school during the school day (in your classroom, computer lab, library, etc.).

Complete performance tasks using different forms of technology.

Answer survey questions about how much you enjoyed / did not enjoy the performance task.

Do an interview about what you liked or did not like about the different forms of technology.

Your Mother/Father/Guardian said you can be in this performance task study, but if you do NOT want to be in the study, you do not have to participate. If you decide later to not be in the study, you can also choose to stop.

APPENDIX J

TECHNOLOGY USE AT HOME VARIABLES SURVEY

Which types of technology platforms do you use at **home**? Select all that apply.

Desktop computer	Tablet (iPad, Android, Kindle)
Laptop computer	Smartphone

How do you use technology at **home?** Check all that apply.

Surfing the internet	E-books
E-mail	Instant messaging / texting
Coding / Web Design	Downloading or listening to music
Writing (for example journal / blog)	Creating artwork (for example drawing, photography, music)
Spreadsheets	Watching movies (for example Netflix, YouTube)
Math games	Social Networking (for example Snapchat, Instagram)
Reading games	I do not use technology at home
Other types of gaming (for example Minecraft, Pokemon, and others)	Other:
Creating presentations (for example PowerPoint / Prezi)	

Approximately, how many combined hours (on average) do you spend on technology devices **at home** per day?

- Less than 1 hour per day
- Between 1-3 hours per day
- Between 3-5 hours per day
- More than 5 hours per day

APPENDIX K

CONSENT FORMS

Parent Consent Form

May 2016

Dear Parent / Guardian,

Your child is invited to participate in a research study conducted by the Advanced Research and Development team from NWEA. We hope to learn how technology enhanced performance tasks impact student engagement. Your child was selected as a possible participant in this study as a result of (school name here) lab school agreement with Northwest Evaluation Association (NWEA) to participate in new research projects and studies.

If your child participates, they will complete a series of 3 performance tasks across different platforms (paper/pencil, technology-enabled, technology-enhanced). They will also be asked to complete a 20-item survey asking questions about their engagement across the three tasks and platforms. The complete study should take approximately 60 minutes. A random sample of students will be selected to participate in a 5-10 minute interview to discuss likes and dislikes about technology in assessment. There are no known risks to your child for participating in the study. Benefits to your child for participating in the study include the opportunity to engage in new technology and assessment design; however, we cannot guarantee that you or your child will personally receive any benefits from this research.

Any information that is obtained in connection with this study and that can be identified with your child will remain confidential and will be disclosed only with your permission. Your child's identity will be kept confidential by utilizing a unique ID number instead of student names. All data will be stored on a secure server and NWEA researchers directly involved in the study will have access to only de-identified data. The de-identified data may also be available to other researchers.

Your child's participation is voluntary. Your decision whether or not to let your child participate will not affect your relationship with (school name here) or NWEA. If you decide to allow your child to participate, you are free to withdraw your consent and discontinue your child's participation at any time without penalty.

If you have any questions, please feel free to contact Meg Guerreiro, Senior Research Associate at NWEA, at (503) 444-6435 or meg.guerreiro@nwea.org.

Please contact Meg Guerreiro to opt your child out of participation in the study.

Thank you

Student Assent Form

Performance Task Study

Your Mother/Father/Guardian has given permission for you to participate in a performance task study. This means you are able to:
Be observed at school during the school day (in your classroom, computer lab, library, etc.).
Complete performance tasks using different forms of technology.
Answer survey questions about how much you enjoyed / did not enjoy the performance task.
Do an interview about what you liked or did not like about the different forms of technology.
Your Mother/Father/Guardian said you can be in this performance task study, but if you do NOT want to be in the study, you do not have to participate. If you decide later to not be in the study, you can also choose to stop.

APPENDIX L

SITE DOCUMENTS

Site Research Information Document – School Administrator

Dear Administrator,

Thank you for your support in NWEA's lab school program. Within the next few weeks, NWEA researchers will be implementing a performance task study in your school. As a site for this study, we require the following components to be in place:

- The study requires the use of a quiet space (preferably the library, computer lab, or classroom) with computers supplied by your school (laptop, desktop, or Chromebook). One computer (with internet connection) is required per student. The space used for any given day must remain consistent and **needs to be reasonably free from distractions**.
- **Tablets or mobile devices cannot be used for this study**
- All students in grade 6, 7, and 8 who have not opted out of the study will be participating. The list of students who opted out will be shared between the school and researchers prior to the start of the study.
- The grouping of students will be determined between school and researchers prior to beginning the study. It is preferable to have an entire grade complete the study on the same day. Our goal is to complete the study in two days, in order to ensure we are not impinging on academic experiences within your school.
- A staff member (teacher or administrator) must be present in the room during the study in order to mitigate any inappropriate student behavior. The staff present may be asked to assist the researchers in handing out supplies or directing students between study components.
- A random sample of students (approximately 7-15 over the course of the study) will be selected to participate in a 5-minute audio recorded follow up interview with a researcher involved in the study. A space reasonably free from distractions will be needed in order to ensure sound quality on the recording (this could be the back of the same room or the hallway adjacent to the room).

If you have any questions or require clarification for the expectations please contact Meg Guerreiro (meg.guerreiro@nwea.org). Additionally, if your school is unable to meet any of the above expectations, please let us know so that we can make alternate arrangements to support the study in your school.

Site Research Information Document – School Staff Member

Dear teacher/staff,

Your students will be participating in a research study involving the completion of a math performance task. There is a total of three math performance tasks (one paper-and-pencil and two using a computer). The order of assessments is randomized. All students will begin the study on a computer. The computer will prompt each student to start one of the three assessments. At the conclusion of each assessment, the student will respond to a short survey. Each test event (assessment and survey) should take about 10-20 minutes to complete (approximately 30-60 minutes total).

A random sample of students (approximately 7-15) will be selected to participate in a 5 minute follow up interview with a researcher involved in the study. The interview will occur after the student completes all three of the performance task assessments. The interview will be recorded.

Teacher/Staff guidelines for participation

- DO bring any technical difficulties to the attention of the researchers
- DO monitor students for talking
- DO direct student questions about the study or design to the researchers
- Please **DO NOT** do any of the following, as to not bias with the study design:
    - Help students solve problems
    - Support students in transitioning between test modes
    - Answer questions about the performance task questions
    - Provide hints for solving the problem
    - Provide encouragement for staying on task (this is part of the study)
    - Help students navigate through the technology within the assessment

If you have any questions or require clarification for the expectations, please contact Meg Guerreiro (meg.guerreiro@nwea.org)

REFERENCES CITED

ACT. (2014a). *The Condition of College & Career Readiness 2013*. *The Condition of College & Career Readiness 2013 Report*. Retrieved from http://www.act.org/research/policymakers/cccr13/pdf/CCCR13-NationalReadinessRpt.pdf

ACT. (2014b). *The Condition of STEM 2014*.

Adair-Hauck, B., Gilsan, E. W., Koda, K., Swender, E. B., & Sandrock, P. (2006). The integrated performance assessment (IPA): connecting assessment to instruction and learning. *Foreign Language Annals*, *39*(3), 359–382. http://doi.org/10.1111/j.1944-9720.2006.tb02894.x

Alfieri, L., Brooks, P., Aldrich, N. J., & Tenenbaum, H. R. (2007). Does discovery-based instruction enhance learning? A meta-analysis. *Journal of Educational Psychology*, *103*(1), 1–18. http://doi.org/10.1037/a0021017

Antin, J., & Shaw, A. (2012). Social desirability bias and self-reports of motivation: A study of Amazon Mechanical Turk in the US and India. *Proceedings Fo the SIGCHI Conference on Human Factors in Computing Systems*, 2925–2934.

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: critical conceptual and methodological issues of the construct. *Psychology in the Schools*, *45*(5), 369–386. http://doi.org/10.1002/pits

Archer, L., Dewitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2010). "Doing" science versus "being" a scientist: Examining 10/11-year-old schoolchildren's constructions of science through the lens of identity. *Science Education*, *94*(4), 617–639. http://doi.org/10.1002/sce.20399

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2012). Science

aspirations, capital, and family habitus: how families shape children's engagement and identification with ccience. *American Educational Research Journal*, *49*(5), 881–908. http://doi.org/10.3102/0002831211433290

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2012). Science aspirations, capital, and family habitus: how families shape children's engagement and identification with ccience. *American Educational Research Journal*, *49*(5), 881–908. http://doi.org/10.3102/0002831211433290

Aschbacher, P. R., Li, E., & Roth, E. J. (2010). Is science me? High school students' identities, participation and aspirations in science, engineering, and medicine. *Journal of Research in Science Teaching*, *47*(5), 564–582. http://doi.org/10.1002/tea.20353

Attewell, P., & Battle, J. (1999). Home computers and school performance. *The Information Society*, *15*(November 2014), 1–10. http://doi.org/10.1080/019722499128628

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children's aspirations and career trajectories. *Child Development*, *72*(1), 187–206. http://doi.org/10.1111/1467-8624.00273

Barker, E. G. (2015). *To what extent do early literacy skills predict growth in mathematics for students with reading difficulties (Unpublished doctoral dissertation)*. Eugene, Oregon.

Bennett, R. E., Persky, H., Weiss, A., Jenkins, F., & Russell, E. M. (2010). Measuring problem solving with technology : A demonstration study for NAEP. *The Journal of Technology, Learning, and Assessment*, *8*(8), 1–45.

Biancarosa, G. (2015). Applied Statistical Design and Analysis: Class 6.

Bloom, B. S. (1956). *Taxonomy of educational objectives handbookI: cognitive domain*. New York: David McKay Co.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and Performance Differences Among Computer-Based and Paper-Pencil Tests. *Journal of Educational Computing Research*, *31*(1), 51–60. http://doi.org/10.2190/GRQQ-YT0F-7LKB-F033

Bressler, D. M., & Bodzin, A. M. (2013). A mixed methods assessment of students' flow experiences during a mobile augmented reality science game. *Journal of Computer Assisted Learning*, *29*(6), 505–517. http://doi.org/10.1111/jcal.12008

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In B. K. A. & J. S. Long (Eds.), *Testing structural equation models* (pp. 136 – 162). Newbury Park, CA: SAGE Publications.

Buchanan, T. (2002). Online assessment: desirable or dangerous?, *33*(2), 148–154. http://doi.org/10.1037/0735-7028.33.2.148

Casey, A., Layte, R., Lyons, S., & Silles, M. (2012). Home computer use and academic performance of nine-year-olds. *Oxford Review of Education*, *38*(September 2014), 617–634. http://doi.org/10.1080/03054985.2012.731207

Clariana, R., & Wallace, P. (2002a). Key Factors Associated With the Test Mode Effect, *33*(5), 593–602.

Clariana, R., & Wallace, P. (2002b). Paper-based versus computer-based assessment: key factors associated With the test mode effect. *British Journal of Educational Technology*, *33*(5), 593–602.

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *JRTE*, *42*(3), 309–328.

Csapo, B., Molnar, G., & Toth, K. R. (2009). Comparing paper-and-pencil and online assessment of reasoning skills: A pilot study for introducing eletronic testing in large-scale assessment

in Hungary. In *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 120–125).

DeWitt, J., Osborne, J., Archer, L., Dillon, J., Willis, B., & Wong, B. (2013). Young Children's Aspirations in Science: The unequivocal, the uncertain and the unthinkable. *International Journal of Science Education*, *25*(2013), 1037–1063. http://doi.org/10.1080/09500693.2011.608197

Donovan, M. a, Drasgow, F., & Probst, T. M. (2000). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *The Journal of Applied Psychology*, *85*(2), 305–313. http://doi.org/10.1037/0021-9010.85.2.305

Eccles, J. F., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and Achievement Motives* (pp. 75–146). San Francisco, CA, CA: W.H. Freeman.

Eisner, E. W. (1991). *The Enlightened Eye: Qualitative Inquiry and the Enhancement of Educational Practice*. New York: Macmillan.

Elster, D. (2014). First- Year Students ' Priorities and Choices in STEM Studies – IRIS Findings from Germany and Austria. *Science Education International*, *25*(1), 52–59.

Fairlie, R. W. (2012). Academic achievement, technology and race: Experimental evidence. *Economics of Education Review*, *31*(5), 663–679. http://doi.org/10.1016/j.econedurev.2012.04.003

Fairlie, R. W. (2015). Do boys and girls use computers differently, and does it contribute to why boys do worse in school than girls? *The B.E. Journal of Economic Analysis & Policy*, *0*(0). http://doi.org/10.1515/bejeap-2015-0094

Fairlie, R. W., & London, R. A. (2011). The effects of home computers on educational outcomes: evidence from a field experiment with community college students*. *Home Computers and Educational Outcomes*, *122*(561), 727–753. http://doi.org/10.1111/j.1468-0297.2011.02484.x.

Fairlie, R. W., & Robinson, J. (2013). Experimental evidence on the effects of home computers on academic achievement among schoolchildren. *American Economic Journal: Applied Economics*, *5*(3), 211–240.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. http://doi.org/10.3758/BF03193146

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2009). G*Power 3.1.9.2.

Feurzeig, W., & Roberts, N. (1999). *Modeling and simulation in science and mathematics education*. New York: Springer.

Field, A. (2013). *Discovering Statistics Using IMB SPSS Statistics* (3rd Ed.). SAGE Publications.

Finegan, J. E., & Allen, N. J. (1994). Computerized and written questionnaires: Are they equivalent? *Computers in Human Behavior*, *10*(4), 483–496. http://doi.org/10.1016/0747-5632(94)90042-6

Finn, J. D. (1989). Withdrawing From School. *Review of Educational Research*, *59*(2), 117–142. http://doi.org/10.3102/00346543059002117

Fiorini, M. (2009). The effect of home computer use on children's cognitive and non-cognitive skills. *Economics of Education Review*, *29*(1), 55–72. http://doi.org/10.1016/j.econedurev.2009.06.006

Ford, B. D., Vitelli, R., & Stuckless, N. (1996). The effects of computer versus paper-and-pencil

administration on measures of anger and revenge with an inmate population. *Computers in Human Behavior*, *12*(1), 159–166. http://doi.org/10.1016/0747-5632(95)00026-7

Fuchs, T., & Woessmann, L. (2004). Computer and student learning: bivariate and multivariate evidence on the availability and use of computers at home and at school. *CESifo Working Paper No. 1321*.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and Quasi-Experimental Studies of Inquiry-Based Science Teaching: A Meta-Analysis. *Review of Educational Research*, *82*(3), 300–329. http://doi.org/10.3102/0034654312457206

Gagnon, L. (2010). *Ready for the future? The role of performance assessments in shaping graduates' acadedmic, professional, and personal lives*. *Center for Collaboroative Education*.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The Effect of Computer-Based Tests on Racial-Ethnic and Gender Groups. *Journal of Educational Measurement*, *39*(2), 133–47. http://doi.org/10.1111/j.1745-3984.2002.tb01139.x

Games, P. A., & Lucas, P. A. (1966). Power of the analysis of variance of independent groups on non-normal and normally transformed data. *Educational and Psychological Measurement*, *26*, 311–327.

Garmire, E. (2005). *Tech tally: approaches to assessing technological literacy*. Washington: Pearson.

Garrison, C., & Ehringhaus, M. (n.d.). Formative and Summative Assessments in the Classroom. Dover, NH: Measured Progress.

Gerjets, P., Scheiter, K., & Schuh, J. (2007). Information comparisons in example-based hypermedia environments: supporting learners with processing prompts and an interactive

comparison tool. *Educational Technology Research and Development*, *56*(1), 73–92. http://doi.org/10.1007/s11423-007-9068-z

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failture to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review Fo Educational Research*, *42*(3), 237–288.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347–360.

Greene, B. A., & Miller, R. B. (1996). Influences on Achievement : Goals , Perceived Ability , and Cognitive Engagement. *Contemporary Educational Psychology*, *21*, 181–192.

Hall, R. (1998). Extraneous and Confounding Variables and Systematic vs Non-Systematic Error. Retrieved December 30, 2015, from https://web.mst.edu/~psyworld/extraneous.htm

Halldorsson, A. M., McKelvia, P., & Bjornsson, J. K. (2009). Are Icelandic boys really better on computerized tests than conventional ones? In *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 178–193).

Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? That is the question: does the medium in which assessment questions are presented affect children's performance in mathematics? *Educational Research*, *46*(1), 29–42. http://doi.org/10.1080/0013188042000178809

Hidden Curriculum. (2014). The glossary of education reform. Retrieved from http://edglossary.org/high-stakes-testing/

Holtzman, S. (2014). Confirmatory Factor Analysis and Structural Equation Modeling of

Noncognitive Assessments using PRO CALIS.

Horton, S. V., & Lovitt, T. C. (1994). A Comparison of Two Methods of Administering Group Reading Inventories to Diverse Learners: Computer Versus Pencil and Paper. *Remedial and Special Education*, *15*(6), 378–390. http://doi.org/10.1177/074193259401500606

Hoyle, R. H. (2011). *Handbook of Structural Equation Modeling*. The Guilford Press.

Hu, L., & Bentler, P. . (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

IBM Corp. (2015). IBM SPSS Statistics for Macintosh, Version 21.0. Armonk, NY, NY: IBM Corp.

ICPSR. (2011). Introduction to Confirmatory Factor Analysis and Structural Equation Modeling. Retrieved March 28, 2017, from http://jonathantemplin.com/files/multivariate/mv11icpsr/mv11icpsr_lecture12.pdf

Jong, T. de. (2006). Technological advances in inquirylLearning. *Science*, *312*(5773), 532–533.

Kaiser Family Foundation. (2010). *Generation M2: Media in the lives of 8- to 18-year olds*.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. http://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kardash, C. M., & Amlund, J. T. (1991). Self-reported learning strategies and learning from expository text. *Contemporary Educational Psychology*, *16*(2), 117–138.

Khairani, A. Z. Bin, & Razak, N. B. A. (2013). Advance in educational measurement: Rasch Model analysis of mathematics proficiency test. *International Journal of Social Science and Humanity*, *2*(3), 248–251. http://doi.org/10.7763/IJSSH.2012.V2.104

Kikis-Papadakis, K., & Kollias, A. (2009). Relfections on paper-and-pencil tests to

eAssessments: narrow and broadband paths to 21st century challenges. In *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 99–103).

King, J. W. C., & Miles, E. W. (1995). A Quast-Experimental Assessment of the Effect of Computerizing Noncognitive Paper-and-Pencil Measurements: A Test of Measurement Equivalence. *Journal of Applied Psychology*, *80*(6), 643–651.

Kline, R. (2013). Assessing statistical aspects of test fairness in structural equation modeling. *Educational Research and Evaluation, 19*, 204–222.

Kvale, S. (1995). The social construction of validity. *Qualitative Inquiry*, *1*, 19–40.

Kyllonen, P. C. (2009). New constructs, methods, and directions for computer-based assessment. In *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 151–156).

Lamb, R., Akmal, T., & Petrie, K. (2015). Development of a cognition-priming model describing learning in a STEM classroom. *Journal of Research in Science Teaching*, *52*(3), n/a-n/a. http://doi.org/10.1002/tea.21200

Langdon, D., Beede, D., & Doms, M. (2011). *STEM: Good Jobs Now and for the Future*.

Lankford, J. S. S., Bell, R. W., & Elias, J. W. (1994). Computerized versus standard personality measures: Equivalency, computer anxiety, and gender differences. *Computers in Human Behavior*, *10*(4), 497–510. http://doi.org/10.1016/0747-5632(94)90043-4

Lee, H. K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing*, *9*(1), 4–26. http://doi.org/10.1016/j.asw.2004.01.001

Lee, M.-K. (2009). CBAS in Korea: experiences, results and challenges. In *The transition to*

computer-based assessment: New approaches to skills assessment and implications for large-scale testing (pp. 194–200).

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, *2*(1), 8. http://doi.org/10.1186/s40536-014-0008-1

Levine, D. W., & Dunlap, W. . (1982). Power of the F test with skewed data: Should one transform or not? *Psychological Bulletin*, *92*(1), 272–280.

Levy, D. (2013). How Dynamic Visualization Technology can Support Molecular Reasoning. *Journal of Science Education and Technology*, *22*(5), 702–717. http://doi.org/10.1007/s10956-012-9424-6

Lin, H., & Dwyer, F. (2006). The Fingertip Effects of Computer-based Assessment in Education. *TechTrends*, *50*(6), 27–31. http://doi.org/10.1007/s11528-006-7615-9

Linn, M., & Eylon, B.-S. (2011). *Science learning and instruction: taking advantage of technology to promote knowledge integration*. New York, NY: Routledge.

Lyons, T., & Quinn, F. (2010). *Choosing Science: Understanding the declines in senior high school science enrolments*. Retrieved from http://www.une.edu.au/simerr/pages/projects/131choosingscience.pdf

Ma, L., Wise, S. L., Thum, Y., & Kingsbury, G. (2011). Detecting response time threshold under the computer adaptive testing environment. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment. *Phi Delta Kappan*, *80*(9), 688–695.

Maehr, M. L., & Meyer, H. A. (1997). Understanding motivation and schooling: Where we've

been, where we are, and where we need to go. *Educational Psychology Review*. http://doi.org/10.1023/A:1024750807365

Malamud, O., & Pop-Eleches, C. (2011). Home computer use and the development of human capital. *The Quarterly Journal of Economics*, *126*(2), 987–1027. http://doi.org/10.1093/qje/qjr008

Marks, H. M. (2000). Student Engagement in Instructional Activity: Patterns in the Elementary, Middle, and High School Years. *American Educational Research Journal*, *37*(1), 153–184. http://doi.org/10.3102/00028312037001153

Martin, R. (2009). Utilising the potential of computer-delivered surveys in assessing scientific literacy. In *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing2* (pp. 172–177).

Mason, J., Patry, M., & Bernstein, D. (2001). An Examination of the Equivalence between Non-Adaptive Computer-Based and Traditional Testing. *Journal of Educational Computing Research*, *24*(1), 29–39. http://doi.org/10.2190/9EPM-B14R-XQWT-WVNL

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review*, *62*, 279–299.

McClarty, K. L., & Gaertner, M. N. (2015). *Measuring Mastery, Best Practices for Assessment in Competency-Based Education*.

McGaw, B. (2006). Assessment fit for purpose. *The International Association for Educational Assessment*.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, *23*(2), 13–23. http://doi.org/10.3102/0013189X023002013

Miles, M. B., Huberman, A. M., & Saldana, J. (2013). *Qualitative Data Analysis: A Methods Sourcebook* (Third Edit). SAGE Publications.

Miller, R. B., Behrens, J. T., Greene, B. A., & Newman, D. (1993). Goals and Perceived Ability: Impact on Student Valuing, Self-Regulation, and Persistence. *Contemporary Educational Psychology*, *18*(1), 2–14.

Miller, R. B., Greene, B. A., Montalvo, G. P., Ravindran, B., & Nichols, J. D. (1996). Engagement in Academic Work: The Role of Learning Goals, Future Consequences, Pleasing Others, and Perceived Ability. *Contemporary Educational Psychology*, *21*(4), 388–422. http://doi.org/10.1006/ceps.1996.0028

Mississippi Department of Education. (2009). Webb's Depth of Knowledge Guide. Retrieved January 17, 2016, from http://www.aps.edu/re/documents/resources/Webbs_DOK_Guide.pdf

Mitri, M. (2003). Applying tacit knowledge management techniques for performance assessment. *Computers & Education*, *41*(2), 173–189. http://doi.org/10.1016/S0360-1315(03)00034-4

Naevdal, F. (2007). Home-PC usage and achievement in English. *Computers and Education*, *49*(4), 1112–1121. http://doi.org/10.1016/j.compedu.2006.01.003

National Governors Association. (2010). *Common Core State Standards for Mathematics*. Washington, DC.

Neuman, G., & Baydoun, R. (1998). Computerization of Paper-and-Pencil Tests: When Are They Equivalent? *Applied Psychological Measurement*, *22*(1), 71–83. http://doi.org/0803973233

Newhouse, C. P. (2011). Using IT to assess IT: Towards greater authenticity in summative

performance assessment. *Computers and Education*, *56*(2), 388–402. http://doi.org/10.1016/j.compedu.2010.08.023

Newmann, F. M., Wehlage, G. G., & Lamborn, S. D. (1992). The significance and sources of student engagement. In *Student engagement and achievement in American secondary schools* (pp. 11–39).

Northern Illinois University. (n.d.). Formative and Summative Assessment. Faculty Development and Instructional Design Center. Retrieved from http://www.niu.edu/facdev/_pdf/guide/assessment/formative and_summative_assessment.pdf

Northwest Evaluation Association. (2011). *Technical Manual for Measures of Academic Progress (MAP) and Measures of Academic Progress for Primary Grades (MPG)*. Portland, OR.

Noyes, J., Garland, K., & Robbins, L. (2004). Paper-based versus computer-based assessment: is workload another test mode effect?, *35*(1). http://doi.org/10.1111/j.1467-8535.2004.00373.x

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: are they equivalent? *Ergonomics*, *51*(9), 1352–1375. http://doi.org/10.1080/00140130802170387

NWEA. (2015a). *2015 NWEA Measures of Academic Progress Normative Data*. Portland, OR.

NWEA. (2015b). *2015 NWEA Measures of Academic Progress Normative Data*. Portland, OR.

NWEA. (2015c). NWEA Partners in Innovation Program. OR.

OECD. (2006a). *No TitlAre students ready for a technology-rich world? What PISA studies tell us*. Paris.

OECD. (2006b). *The Programme for International Student Assessment (PISA)*. *Science*. Paris. Retrieved from

http://www.namsmat.is/vefur/rannsoknir/PISA_skyrslur_almennt/1_skyrslur_OECD_PISA/PISA_2006_science_competencies_summary.pdf

Onwuegbuzie, A. J. (2003). Expanding the framework of internal and external validity in quantitative research. *Research in the Schools*, *10*, 71–90.

Onwuegbuzie, A. J., & Leech, N. L. (2007). Validity and qualitative research: an oxymoron? *Quality & Quantity*, *41*, 233–249.

Oregon Department of Education. (2014). Common Core State Standards - communication. Retrieved from http://www.ode.state.or.us/search/page/?id=3265

Özalp-Yaman, Ş. ̧ & Çağiltay, N. E. (2010). Paper-based versus computer-based testing in engineering education. *2010 IEEE Education Engineering Conference, EDUCON 2010*, 1631–1637. http://doi.org/10.1109/EDUCON.2010.5492397

Papanastasiou, E. C., Zembylas, M., & Vrasidas, C. (2003). Can computer use hurt science achievement? The USA results from PISA. *Journal of Science Education and Technology*, *12*(3), 325–332.

Patton, M. Q. (2001). *Qualitative Research and Evaluation Methods* (Third Edit). Thousand Oaks, CA: SAGE Publications.

Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. *Advanced in Motivation and Achievement: Motivation Enhancing Enrionments*, *6*, 117–160.

QSR International. (2014). NVivo qualitative data analysis software.

Quay, L. (2010). *Higher Standards for All: Implications of the Common Core for Equity in Education. Research Brief. Chief Justice Earl Warren Institute on Race, Ethnicity …*. Retrieved from http://files.eric.ed.gov/fulltext/ED536694.pdf

Reder, L. M. (1979). The role of elaborations in memory for prose. *Cognitive Psychology*, *11*, 221–234.

Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, *28*(2), 147–169. http://doi.org/10.1023/B:MOEM.0000032312.95499.6f

Reid, H. M. (2013). *Introduction to Statistics: Fundamental Concepts and Procedures of Data Analysis* (1st Ed). SAGE Publications.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: a comparison of two approaches. *New Directions for Institutional Research*, *161*, 69–82. http://doi.org/10.1002/ir.20068

Ripley, M. (2009). Transformational Computer-based Testing. In *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 92–103). http://doi.org/10.2788/60083

Rosenbaum, E., Klopfer, E., & Perry, J. (2007). On location learning: authentic lapplied science with networked augmented realities. *Journal of Science Education and Technology*, *16*(1), 31–45. http://doi.org/10.1007/s10956-006-9036-0

Rudestam, K. E., & Newton, R. R. (2007). *The method chapter: describing your research plan*. *Surviving your dissertation: a comprehensive guide to content and process*. SAGE Publications. Retrieved from http://books.google.com/books?id=vmWdAAAAMAAJ&pgis=1

Saeed, S., & Zyngier, D. (2012). How Motivation Influences Student Engagement: A Qualitative Case Study. *Journal of Education and Learning*, *1*(2), 252–267. http://doi.org/10.5539/jel.v1n2p252

SAS Institute Inc. (2013). SAS.

Scalise, K. (2012). *Using technology to assess hard-to-measure constructs in the common core state standards and to expand accessibility*. Retrieved from http://k12center.org/rsc/pdf/session1-scalise-paper-2012.pdf

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is It Really Robust? Reinvestigating the Robustness of ANOVA Against Violations of the Normal Distribution Assumption. *European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(4), 147–151.

Schmitt, J., & Wadsworth, J. (2006). Is there an impact of household computer ownership on children's educational attainment in Britain? *Economics of Education Review*, *25*(6), 659–673. http://doi.org/10.1016/j.econedurev.2005.06.001

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *Journal of Educational Research*, *99*(6), 323–337. http://doi.org/10.3200/JOER.99.6.323-338

Schwols, A., & Dempsey, K. (2013). *Common core standards for middle school mathematics*. Denver, CO, CO: ASCD.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215–232. http://doi.org/10.1111/j.1745-3984.1993.tb00424.x

Shilling, R. (2015). Keynote Address. Presentation presented as the annual CRESST meeting,

Redondo Beach, California.

Shuttleworth, M. (2009). Counterbalanced Measures Design. Retrieved September 7, 2015, from https://explorable.com/counterbalanced-measures-design

Silm, G., Must, O., & Täht, K. (2013). Test-taking effort as a predictor of rerformance in low-stakes tests. *Trames. Journal of the Humanities and Social Sciences*, *17*(4), 433–448. http://doi.org/10.3176/tr.2013.4.08

Slavin, R. E., Lake, C., Hanley, P., & Thurston, A. (2014). Experimental evaluations of elementary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching*, *51*(7), 870–901. http://doi.org/10.1002/tea.21139

Smiley, W., & Anderson, R. (2011). Measuring Students ' Cognitive Engagement On Assessment Tests : A Confirmatory Factor Analysis Of The Short Form Of The Cognitive Engagement Scale. *Research & Practice in Assessment*, *6*, 12.

Smither, J. W., Walker, A. G., & Yap, M. K. T. (2004). An Examination of the Equivalence of Web-Based Versus Paper-and-Pencil Upward Feedback Ratings: Rater- and Ratee-Level Analyses. *Educational and Psychological Measurement*, *64*(1), 40–61. http://doi.org/10.1177/0013164403258429

Sorensen, H., & Andersen, A. M. (2009). How did Danish students solve the PSA CBAS items? In *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 201–208).

Sundre, D. L. (1997). Differential examinee motivation and validity: A dangerous combination. *Annual Meeting of the American Educational Research Association*.

Sundre, D. L. (1999). Does examinee motivation moderate the relationship between test consequences and test performance? *Annual Meeting of the American Educational Research*

Association.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162–188. http://doi.org/10.1080/08957347.2011.555217

Tai, R. H., Qi Liu, C., Maltese, A. V, & Fan, X. (2006). Career choice. Planning early for careers in science. *American Association for the Advancement of Science*, *312*(5777), 1143–1144. http://doi.org/10.1126/science.1128690

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. http://doi.org/10.5116/ijme.4dfb.8dfd

Templin, J. (2011). Introduction to Confirmatory Factor Analysis and Structural Equation Modeling. *Advanced Multivariate Statistical Methods. ICPSR Summer Session #2*. Retrieved from http://jonathantemplin.com/files/multivariate/mv11icpsr/mv11icpsr_lecture12.pdf

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (n.d.). Examining Inferences about test-taking motivation: The student opinion scale (SOS). *Journal of General Education*.

Thelk, A. D., Sundre, D. L., Horst, S. J., Finney, S. J., Thelk, A. D., Sundre, D. L., … Finney, S. J. (2009). Penn State University Press Motivation Matters : Using the Student Opinion Scale to Make Valid Inferences About Student Performance All use subject to JSTOR Terms and Conditions Motivation Matters : Using the Student Opinion Scale to Make Valid Inferences. *The Journal Of General Education*, *58*(3), 129–151.

Thomas, A. E. (2017). Gender Differences in Students' Physical Science Motivation: Are Teachers' Implicit Cognitions Another Piece of the Puzzle? *2American Educational Research Journal*, *4*(1), 35–58.

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP Norms for Student and School Achievement Status and Growth*. Portland, OR, OR.

Tsikalas, K., Lee, J., & Newkirk, C. (2007). *Home computing , school engagement and academic achievement of low-income adolescents*. *Computer for Youth*. New York.

U.S. Department of Labor. (2007). *The STEM Workforce Challenge: The Role of the Public Workforce System in a National Solution for a Competitive Science, Technology, Engineering, and Mathematics (STEM) Workforce*. *Workforce*. Washington, D.C.

Vigdor, J., Ladd, H., & Martinez, E. (2014). Scaling the digital divide: home computer technology and student achievement. *Economic Inquiry*, *52*(3), 1103–1119. http://doi.org/10.1111/ecin.12089

Vispoel, W. P., Boo, J., & Bleiler, T. (2001). Computerized and Paper-and-Pencil Versions of the Rosenberg Self-Esteem Scale: A Comparison of Psychometric Features and Respondent Preferences. *Educational and Psychological Measurement*, *61*(3), 461–474. http://doi.org/10.1177/00131640121971329

Walker, C. O., Greene, B. A., & Mansell, R. A. (2006). Identification with academics, intrinsic/extrinsic motivation, and self-efficacy as predictors of cognitive engagement. *Learning and Individual Differences*, *16*(1), 1–12. http://doi.org/10.1016/j.lindif.2005.06.004

Wang, X. (2013). Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, *50*(5), 1081–1121. http://doi.org/10.3102/0002831213488622

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. http://doi.org/10.1017/CBO9781107415324.004

Wenglinsky, H. (2006). Technology and achievement: the bottom line. *Educational Leadership*, *63*(4).

Wigfield, A., & Eccles, J. S. (2000). Expectancy–Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. http://doi.org/10.1006/ceps.1999.1015

Wiggins, G. (1998). *Educative assessment*. San Francisco: Jossey-Bass.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. http://doi.org/DOI 10.1207/s15324818ame1902_2

Wise, S. L., & DeMars, C. E. (2005a). An application of item response time: The effort-moderated IRT model, *43*(1), 19–38.

Wise, S. L., & DeMars, C. E. (2005b). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment*, *10*(1), 1–17. http://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., Kingsbury, G., Thomason, J., & Kong, X. (2004). An investigation of motivation filtering in a statewide achievement testing program. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation inf computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. http://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., Ma, L., Kingsbury, G. G., & Houser, C. (2010). An investivation of the relationship between time and testing and test-taking effort. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Wittwer, J., & Senkbeil, M. (2008). Is students' computer use at home related to their mathematical performance at school? *Computers and Education*, *50*(4), 1558–1571. http://doi.org/10.1016/j.compedu.2007.03.001

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequences of performance, test motivation, and mentally taxing items. *Applied Measurement in Education*, *8*(4), 341–351.

Zacharia, Z. C., Manoli, C., Xenofontos, N., de Jong, T., Pedaste, M., van Riesen, S. A. N., … Tsourlidaki, E. (2015). Identifying potential types of guidance for supporting student inquiry when using virtual and remote labs in science: a literature review. *Educational Technology Research and Development*, *63*(2), 257–302. http://doi.org/10.1007/s11423-015-9370-0