INTEREST BASED ASSESSMENT: THE IMPACT OF CHOICE IN MEASURES OF ASSESSMENT ON PERFORMANCE AND ENGAGEMENT

by

DEBORAH E. ADKINS

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

March 2020

DISSERTATION APPROVAL PAGE

Student: Deborah E. Adkins

Title: Interest Based Assessment: The Impact of Choice in Measures of Assessment on Performance and Engagement

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

Dr. Kathleen Scalise	Chairperson
Dr. Gina Biancarosa	Core Member
Dr. Keith Hollenbeck	Core member
Dr. Sylvia Thompson	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

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

Degree awarded March 2020

DISSERTATION ABSTRACT

Deborah E. Adkins

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

March 2020

Title: Interest Based Assessment: The Impact of Choice in Measures of Assessment on Performance and Engagement

The role of interest in assessment on student performance and engagement is an area of research that has yet to be fully examined. I explored models of assessment grounded in Item Response Theory (IRT) to identify the model that best fit an assessment that utilized interest-based context when administered to middle school students (grades 6–8) in Washington, Oregon, Illinois, and North Carolina ($N = 517$). I examined test properties (i.e., measurement precision & differential item functioning) across test forms. I evaluated the impact of context (i.e., context matched in contrast to not matched to student interest) on systematic measures of student engagement. Measures included response time fidelity (RTF) and response time engagement (RTE) which measure item and test level engagement respectively. Results from IRT modeling showed the unidimensional 1PL model to exhibit reasonable fit for the purpose of this assessment. Results for RTF suggested solution behavior on individual items was similar across students irrespective of context. Similarly, the three-way, between-subjects analysis of variance suggested that RTE did not differ by context, grade, nor gender. Although significant differences were not identified in achievement nor engagement with this reading assessment, it opens the door for further research across multiple subject areas.

CURRICULUM VITAE

NAME OF AUTHOR: Deborah E. Adkins

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Portland State University, Portland, OR
Western Oregon University, Monmouth, OR
Mt. Hood Community College, Gresham, OR

DEGREES AWARDED:

Doctor of Philosophy, Educational Leadership, 2020, University of Oregon
Master of Science, Systems Science, 2011, Portland State University
Bachelor of Science, Computer Science, 1997, Western Oregon University
Associate of Arts, General Studies, 1995, Mt. Hood Community College

AREAS OF SPECIAL INTEREST:

Technology in Education
Measurement and Assessment
Educational Leadership

PROFESSIONAL EXPERIENCE:

Research Scientist II , NWEA, 2017-Present

Research Scientist I, NWEA, 2011-2017

Senior Research Associate, NWEA, 2009-2011

Research Associate, NWEA, 2007-2009

Data Analyst, NWEA, 2005-2007

Technical Support Specialist, NWEA, 2003-2005

Product Development Manager, Kewill, 2000-2001

Software Engineer, Kewill, 1999-2000

Technical Support Specialist, Aristo Computers, 1997-1999

GRANTS, AWARDS, AND HONORS:

Cum Laude, Western Oregon University, 1997

Presidents List, Mt. Hood Community College, 1995

National Dean's List, Mt. Hood Community College, 1995

PUBLICATIONS:

Anderson, R.C., Porter, L., & Adkins, D. (2019). A Dramatic Confrontation of Frames: Arts-Integration Teacher Development, Organizational Learning, and School Change. *Leadership and Policy in Schools.*

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

Cronin, J., Dahlin, M., Adkins, D., & Kingsbury, G.G. (2007). *The Proficiency Illusion.* Thomas B. Fordham Institute. Available from http://edex.s3-us-west-2.amazonaws.com/publication/pdfs/Proficiency_Illusion_092707_7.pdf

Cronin, J., Dahlin, M., Adkins, D., & Kingsbury, G.G. (2007). The Proficiency Illusion. *American Educator*, 23-28.

ACKNOWLEDGMENTS

This research was supported in part by NWEA; the findings and conclusions expressed do not necessarily represent the point of view nor opinions of NWEA. NWEA funded the development and implementation of the prototyped assessment used within this research. Special thanks to Mike Nesterak, Senior Director of the Product Innovation Center, for his support and encouragement in the completion of this project.

I would like to express my deep appreciation for my advisor and mentor, Kathleen Scalise, for her expertise, encouragement, and grace. I have learned much over the past three years and feel truly blessed to have been afforded the opportunity to work with a professional of her stature who is so a highly regarded in the field of measurement. Additionally, I would like to thank my committee members, Gina Biancarosa, Keith Hollenbeck and Sylvia Thompson, for their time, support, and encouragement in preparation of this manuscript.

A very heartfelt and personal thanks goes to my husband Matt and son Hunter. After a cancer scare and throughout his treatment and recovery Matt held steadfast in his love and support as I worked towards completion of my PhD program. Hunter shown his support in his own way as well, such as not asking for extra family time when he knew I just couldn't squeeze in another minute and helping out with household chores. Now that he is off at college and furthering his own education, I feel proud knowing that I had a part in that and look forward to summer break in hopes that he will be able to squeeze a few spare minutes of his time. In closing, I'd like to thank friends, family, and colleagues who have given me the time and space to follow my dream and look forward to getting reacquainted with all of them.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Table Page

Table Page

CHAPTER I

INTRODUCTION

Motivated students are more likely to engage while taking an educational assessment and therefore the research literature has found they may demonstrate greater effort and produce achievement scores that better represent their true ability (Wise, Ma, Kingsbury & Hauser, 2010). Allowing students to have choice has been associated in some studies with attitude and interest in education, and therefore choice has been one attribute of motivation that has been shown to influence achievement (Guthrie et al., 2007; Ivey & Broaddus, 2001; Logan, Medford, & Hughes, 2011; McKenna, Conradi, Lawrence, Jang, & Meyer, 2012).

Purpose of Study

The purpose of this study was to examine the impact of allowing students choice over the topic of interest as the context for assessment. This research sought to examine improvement in measurement characteristics such as model fit, score estimation, precision, and accuracy, all of which may relate to the validity and utility of assessments. In the context of one reading assessment, I examined how student choice, related to selecting topics of interest in a reading passage, interacted with the formal characteristics of how the assessment performed. The intent of this study was to contribute to research on personalized measures of assessment. While numerous subject matter areas might have lent themselves to choice components in assessment, reading was selected as the subject matter area focus here because of the availability of an appropriate data set to apply formal measurement models. The study was intended to represent a contribution to the research literature exploring the relationship between personalized assessment context

using topics of interest selected by the student, and the resulting formal measurement characteristics of the outcome measure(s). In this case, the outcome measures involved (a) an estimate regarding the subject matter area, see below and (b) an estimate of assessment engagement based on behavioral timing data collected during the assessment.

This study used extant data from a reading assessment prototype, the Context Personalization Assessment Instrument (Product Innovation Center, 2018). Data were collected through the Product Innovation Center, a branch of the Research division at the Northwest Evaluation Association (NWEA) in Portland, Oregon. The reading assessment was based on three NWEA *goals* for reading assessment: (a) Literature, (b) Informational Text, and (c) Vocabulary Use and Acquisition. The prototyped assessment, including the choice component, was constructed using existing content from NWEA's item bank. Items were cloned by substituting the context with topics rated of most interest to middle school students as obtained through response data previously collected via survey conducted by NWEA (D. Adkins, personal communication, April 17, 2017).

Contexts selected by middle school students indicated they would prefer reading choices that focused on animals, fantasy, or sports. The choice element in this study employed a data set previously collected through a field test of the prototyped assessment that assigned students to one of three conditions: (a) *Choice Condition*, students chose their preference from among these three contexts and received their preference, and also included two comparison groups with random assignment to passage context (b) *Comparison Condition 1*, students chose their preference from among these three contexts and did not receive their preferred context (i.e., they received the original

context); (c) *Comparison Condition 2*, students indicated no preference followed by random assignment to a context, which is discussed in more detail the Methods section.

In the next section, I present the two research questions that I explored in this study, each of which have several subparts.

Research Questions 1 (RQ1)

Using formal measurement models, I investigated some key measurement properties of the resulting subject matter assessment when context choice was employed, for the reading assessment and data set described above.

RQ1a. To what extent did an item response model from among some operational models (i.e. Rasch, 1PL, 2PL, and 3PL) fit the subject matter data set as a single dimension? A sequence of unidimensional models were fit beginning with the Rasch model as it was the most parsimonious. Subsequent models were compared to previous models to identify the best fit. Data were linked through a common item set (anchor set).

RQ1b. Based on the unidimensional model, identified in 1a, to what extent did the corresponding three-dimensional model, established through assignment of items to the three *goals* identified in the NWEA assessment, show improved fit for the instrument?

RQ1c. As it pertained to the model that exhibited the most reasonable fit for the purposes of this study, to what extent did anchor items exhibit differential item functioning (DIF) for students in the *Choice Condition* as compared to those in *Comparison Condition 1* (choice followed by random assignment to a context) and *Comparison Condition 2* (students indicated no preference and were randomly assigned to a context) groups?

RQ1d. To what extent were distributions of proficiency estimates, of the subject matter assessed, between the *Choice* group and the group of students in *Comparison Condition 1,* the same or different for this data set? Although comparison scores on an alternate instrument were not available for this data set, the condition of randomization for the *Choice* Condition as compared to *Comparison Condition 1* might reasonably support the claim that the two groups could be considered statistically equivalent subsets, and therefore should show similar central tendency and variation for distributions of proficiency estimates if choice were not a statistically significant factor.

Research Questions 2 (RQ2)

To what extent did providing students with the context they chose (versus not providing their choice) impact engagement as measured by response time effort, in the reading assessment for the data set here?

RQ2a. To what extent did patterns of average engagement vary across the three groups (Choice and the two comparison conditions), as captured by NWEA's measure of Response Time Fidelity (RTF), for this data set?

RQ2b. To what extent did the impact of choice differ by gender, as students progress through middle school, as a function of Response Time Effort (RTE), for this data set?

In the remainder of this dissertation, I first provide a review of the literature. Next, I discuss the methodology I used for addressing each of my research questions using the extant data set described above. Then I present the results of my findings. Finally, I close with a discussion of my findings, how findings fared relative to prior research, and how my investigation has extended the prior research.

Literature Review

In educational assessment, interest-based interventions and learning have been an increasingly apparent research topic since at least the late 1980s (Anand & Ross, 1987). Early research focused on a single type of personalization, a fill-in-the-blank item type (Anand & Ross, 1987; Cordova & Lepper, 1996). More recent studies have incorporated a thematic approach including:

- personalizing learning based on individual topics of interest (Bernacki & Walkington, 2014; Walkington, 2013)
- depth (i.e., whether personalized by substituting topic name or by introducing topic specific details) (Walkington & Leigh, 2015)
- difficulty (Bernacki & Walkington, 2018; Walkington, 2013)
- the grain size (i.e., broadly personalized to members of a larger group) (Walkington & Leigh, 2015)
- agency or ownership (i.e., whether or not the student selects the personalization) (Walkington & Leigh, 2015).

The research literature described that specifically in reading for intrinsic motivation, students should be encouraged to read personally interesting materials and feel they have some control over what they read (Brozo et al., 2014; Ivey & Broaddus, 2001). Additionally, Walkington, Petrosino, and Sherman (2013) suggested that personalized context evoked interest and acted as a catalyst for motivation and improved achievement based on their research using context personalization in middle and high school mathematics. In a single fill-in-the-blank question survey asking, *When taking a reading test, I would prefer if the reading were about. . .*, middle school students

demonstrated that choice in reading context mattered to most students, as only 3 of 333 indicated they had no preference in context (D. Adkins, personal communication, April 17, 2017).

Furthermore, Guthrie et al. (2007) demonstrated that internal motivation, consisting of interest, involvement, choice, efficacy, and social, of primary grade students (4th grade) was predictive of their reading achievement measured as a growth score after controlling for prior reading scores. They found that interest explained 12% of the variance in reading scores, choice explained an additional 22%, and involvement another 12%, all of which were statistically significant, however, neither efficacy nor social were significant predictors of achievement.

Similarly, Logan et al., (2011) found that intrinsic motivation differentially impacted reading performance of low and high achievers for students in years five and six from the UK (4th & 5th grades in the US). They found that intrinsic reading motivation explained significant variance in reading growth ($R^2 = .67$, $p < .01$) for the low ability reading group. However, intrinsic reading motivation explained no variance for the high ability reading group.

Test effort engagement. More broadly across many assessment contexts, the level of student engagement has been implicated as an important factor in assessment outcomes. Sundre (1999) examined the relationship between student engagement effort and achievement in assessment with and without consequences. While Sundre found no significant correlation with achievement using assessments for which there were consequences, a significant correlation ($r = .38$) was found with the no consequences assessments that accounted for 14% of the variance in test score performance. Wise and

Kong (2005) found that test validity improved when using a measure of test effort engagement as a mechanism to identify and remove non-effortful test events. Subsequent studies provide additional support of the importance of using measures of effort for assessment (Setzer, Wise, van den Heuvel, & Ling, 2013; Wise, 2006)

Test effort in assessment is most commonly measured through self-report and response time-based measures. Conversely, systematic observational measures are used as a supplementary assessment of students' on-task and off-task behavior (Hintze & Matthews, 2004). Each type of measure has associated advantages and disadvantages and should be considered based on its intended use.

Systematic observational measures. Observational measures such as the Code for Instructional Structure and Student Academic Response-Mainstream Version (MS-CISSAR) is one such instrument that has been used by trained observers to examine the academic engagement of students (Greenwood, Horton, & Utley, 2002). Similarly, the Behavior Observation of Students in Schools (BOSS), is another observational measure of individual student engagement that has been used to supplement individual student assessment procedures (Hintze & Matthews, 2004). The advantage of systematic observational measures is that when used with fidelity they are reliable and valid measures. As such the recommendation for the use of direct observation as a measure of student engagement is that observers are highly trained individuals (Greenwood, Carta, Kamps, Terry, & Delquadri, 1994; Shapiro, 2004). Therein lies one of the disadvantages of using observation measures, the time required for training an observer. Observer training required for the Behavior Observation of Students in Schools is 10-15 hours (Fredricks et al., 2011). Additionally, such measures are labor intensive with

recommendations by developers of three, 20-30 minute observations over the course of at least 2-3 days per individual student (Fredricks et al., 2011; Shapiro, 2004). Lastly, the intended use of systematic observational measures is in assessing students engagement in academics overall. Therefore, using information garnered to predict achievement may call into question the validity of such predictions without further corroborating evidence.

Self-reported measures. One primary advantage of self-reported measures is their ease of use. Additionally, self-reported measures may not require many items to obtain satisfactory reliability (Sundre, 1999). In the Student Opinion Scale (SOS) students indicate their level of effort and importance within an assessment activity through a 10-item Likert scale questionnaire that has reported values for reliability consistently in the .80s (Sundre & Moore, 2002). There are however disadvantages to using self-reported measures of engagement. Disadvantages include when students are untruthful in their responses based on perceived performance (i.e., they do not do well on the test and attempt to justify poor performance by indicating a lack of effort) (Pintrich & Schunk, 2002). Similarly, students may be untruthful in their responses due to perceived consequences in providing a truthful response (i.e., students who do not put forth their best effort may be required to re-test). A third disadvantage is that self-reported measures tend to pose statements of effort related to the entire test as opposed to effort at different times within a test. For example, from the Student Opinion Survey (Sundre, 1999) students respond to statements such as, *while taking this test, I could have worked harder on it* and *I gave my best effort on this test*.

Response time-based measures. As a measure of engagement, response time-based measures have the advantage of being automatically collected through the use of

technology (e.g., a tablet, computer, smartphone, etc.) freeing the measure from bias introduced by either a trained observer or by a student self-reporting her engagement (Wise & Kong, 2005). Additionally, engagement can be measured at both the item and test levels. The added advantage of measuring engagement for each item is that it allows for observing patterns of both solution behavior and rapid guessing behavior over the course of the test. Solution behavior is when a student actively attempts to determine the solution to each item (Schnipke & Scrams, 1997). Rapid guessing behavior is the opposite of solution behavior in a student does not attempt to determine the correct answer, instead the time spent on the item response is likely too brief for the student to have read and responded to the item in a thoughtful manner (Schnipke & Scrams, 1997). The disadvantage of response time-based measures is that the test must be administered through technology (e.g., a tablet, computer, smartphone, etc.) to allow for the precise and unbiased measure of time expended.

Common three second threshold. As the name indicates, the common three second threshold identifies item responses of three seconds or less as non-effortful. This threshold was initially used with computer adaptive tests (CAT) that utilized item pools containing thousands of items (Wise, Kingsbury, Thomason, & Kong, 2004).

Normative threshold measures. Normative threshold measures include the NT10, NT15, and NT20. Each are computed similarly differing only on the percent of the average overall time spent on an item by a group of examinees used as the threshold. For example, a test item that takes students an average of 40 seconds to complete would have a threshold of 4 seconds using the NT10, 6 seconds for the NT15, and 8 seconds using the NT20.

Achievement

In order to assess student ability in any subject, necessary criteria should be considered (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education [NCME], 2014). Standards include that students should have had the opportunity to learn the content on which they will be assessed. Second, the measure used to assess students must adhere to standards of assessment for validity, reliability, and fairness. An additional criterion is that students should be motivated to engage with an assessment by providing effortful responses.

In the United States, as of 2013, 41 states including the District of Columbia (Achieve, 2013) use the Common Core State Standards (CCSS) to reflect the content standards for which instruction should be guided and against which students should be assessed (National Governors Association Center for Best Practices [NGA] & Council of Chief State School Officers [CCSSO], 2010). Currently, Alaska, Florida, Indiana, Nebraska, Oklahoma, South Carolina, Texas, and Virginia, have not adopted the shared standards and Minnesota has adopted only the English Language Arts (ELA) Standards (Achieve, 2013).

In addition to content standards, test developers must adhere to a set of standards to insure the validity of the interpretation of resulting test scores (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education [NCME], 2014). However, when students are not motivated and fail to put forth their best effort and thoughtfully engage with the measure used to assess them, the valid interpretation of those test scores are called into question

and are unlikely to represent what students know (Schnipke, 1996; Setzer, Wise, van den Heuvel, & Ling, 2013; Wise, Kingsbury, Thomason, & Kong, 2004; Wise, 2006; Wise & DeMars, 2005).

Furthermore, effort is used as a proxy for engagement and is measured through response time effort (Wise & Kong, 2005). The normative threshold measure, NT10, has been employed in this study to compute average response times for each item and the threshold against which determination of an effortful response (i.e. solution behavior) is identified. Solution behavior is a dichotomously scored index given an examinee *j*'s response time, RT_{ij}, to item *i* and is computed as:

$$SB_{ij} = \begin{cases} 1 & if\ RT_{ij} \geq T_i, \\ 0 & otherwise. \end{cases}$$

Subsequently, a response time effort score is computed for each test event as a function of student solution behavior for each item within the test. The score has a range from 0 to 1 with scores nearer to 1 indicating more effortful test taking behavior. The index of response time effort for examinee *j* to the test is computed as:

$$RTE_j = \frac{\sum_{j=1}^{n} SB_{ij}}{k},$$

where *k* = the number of items in the test.

In addition to response time effort which is a measure of a student's overall effort on a test, a similar measure, response time fidelity (RTF), is an index of effort across examinees for a particular item (Wise, 2006). The score has a range from 0 to 1 with scores nearer to 1 indicating more examinees exhibited effortful test taking behavior for an item. The index of response time fidelity for item *i* of the test is computed as:

$$RTF_i = \frac{\sum_{j=1}^{n} SB_{ij}}{N}$$

where N = the number of examinees in the sample.

Although interest-based assessments could apply to a variety of domains such as reading, mathematics or science, the domain used in this study was reading. In what follows I have provided the theoretical framework that has served as the basis for and guided this scholarship. I have concluded by extricating applicable components of the framework employed in my current study.

Theoretical Framework

This study draws on both motivation and choice theory. I begin with the history of motivation theory, transition to the grand theories of motivation and then touch on the mini theories of motivation calling out which of those theories best explain motivation as it relates to education. I conclude with a brief discussion on choice theory and describe the aspects of choice theory similar to motivation that are pertinent to education.

History of motivation theory. The history of motivation can be distilled into four somewhat overlapping categories. They are, (a) theories which focus on a hierarchically arranged psyche, (b) those which examined the mind versus body dualism, (c) grand theories, and (d) mini theories. Figure 1 depicts the progression of motivation theory demarcating an approximated timeline based on seminal papers.

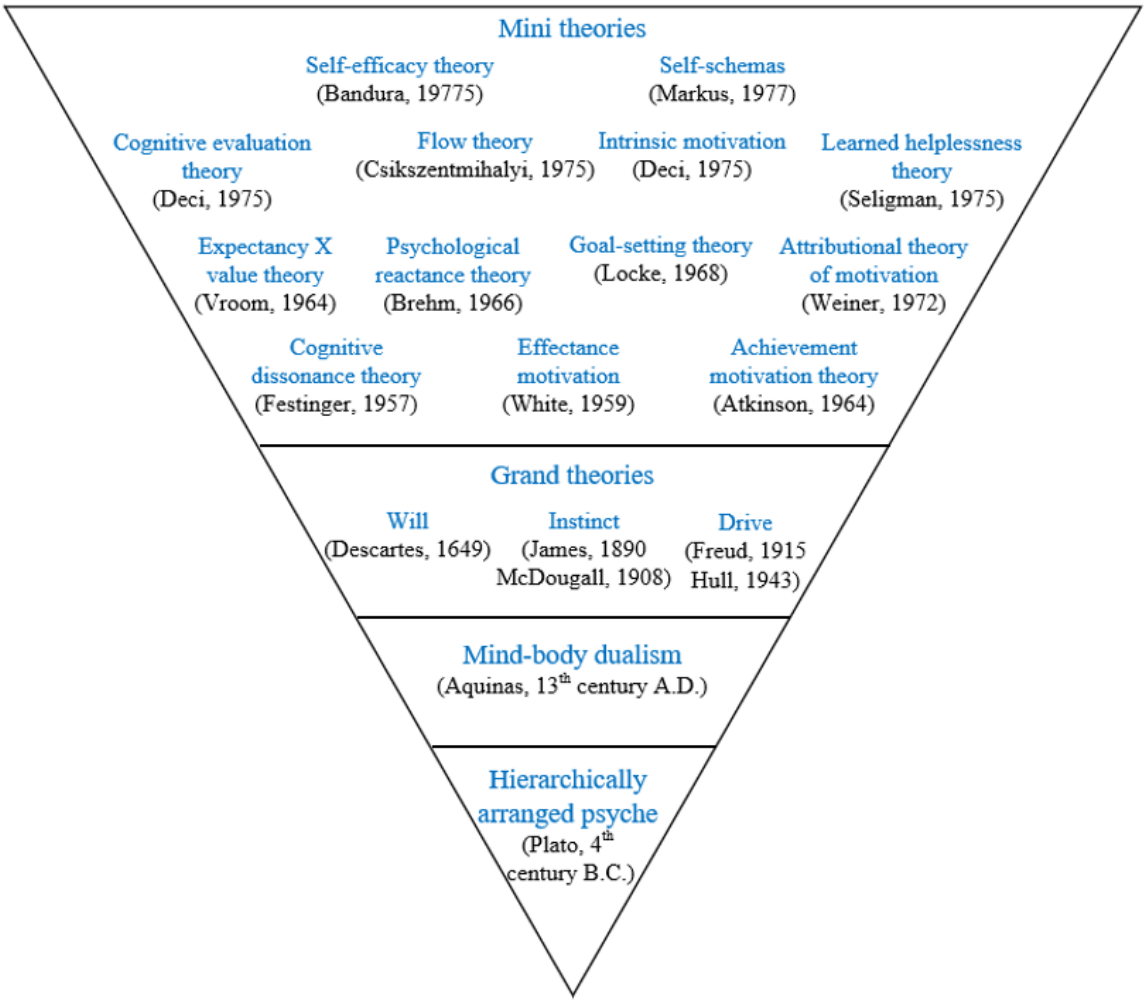


Figure 1. Historical progression of motivation theory.

Hierarchically arranged psyche. The theory of human motivation can be traced back to Plato (4th century B.C.). Plato suggested a hierarchically arranged *psyche*. He believed that motivation was influenced by physiological needs, a set of socially accepted standards, and governed by personal will. Plato identified these elements as, (a) appetite (e.g., hunger, thirst, sex), (b) competition (i.e., social standards), and (c) calculation (i.e., reason and choosing) (Reeve, 2015). Plato's theory remained relatively intact for centuries. Although, Aristotle's theory of motivation (5th century B.C.) substituted *nutrition, perception* and *ration* for *appetite*, *competition* and *calculation* respectively (Reeve, 2015).

Since then motivation theory was distilled further to two major strands, physiology (bodily needs and impulses) and philosophy (will), or a mind-body dualism. In the 13th century A.D. Thomas Aquinas described motivation as *reason* comprised of both cognitive and appetitive powers. Cognitive power was further described as intellect, which aided in knowing and understanding. Aquinas described appetitive power as human will. Another 400 years would pass before will, as motivation, would reemerge (Reeve, 2015).

Grand theories. Between 1649 and 1943 three grand theories of motivation emerged, (a) will, (b) instinct, and (c) drive. Grand theories are those that are all encompassing and through which the full range of human action, in all circumstances could be explained.

Will. The first grand theory understood motivation within two themes, bodily desires and the mind. In 1641, Descartes extended the mind-body dualism, first introduced by Aquinas, in his writing of the meditations (Descartes, 1911). However, it was not until his final writing completed in 1649, *Les Passions de l'âme* (The Passions of the Soul), that Descartes indicated the passions (i.e., emotions or bodily desires) were motivational states that were controlled by the soul (i.e., will) (Descartes, 1975). The crux of which was, if you could philosophically explain the will you would be able to account for all of human motivation (Reeve, 2015). However, the theory of will as motivation fell out of favor as it became evident that philosophy alone could not provide a clear understanding of will. This lead to the exploration by others for precursors to will as the guiding force behind motivation.

Instinct. The second grand theory, guided by Darwin's theory of evolution (1859)

suggested instincts (i.e., evolved impulses) guided our behavior as the source of motivation. The focus of this new theory of motivation was on the, "mechanistic, genetically endowed concept of the instinct" (Reeve, 2015, p. 49). The appeal of Darwin's motivational concept was that it provided the origin of motivation that was lacking in the first grand theory, *will* (Reeve, 2015).

James was one of the pioneers of the instinct theory of motivation (1890). James adopted much from Darwin to which he bestowed upon human beings numerous physical and mental instincts. The pertinent stimulus was all that was necessary to convert an instinct into motivation for action (Reeve, 2015).

McDougall was another of the forerunners of instinct theories of motivation (1908). Detailed in his theory instinct must be unlearned, uniform in expression, and universal in a species. Additionally, instinct was composed of three elements, (a) perception, (b) behavior, and (c) emotion. McDougall (1908) believed instinct was an innate predisposition that guided human perception of an object, attached an emotion to it, and behaved or acted upon it based on that perception as the source of motivation as opposed to will. The undoing of instinct as an all-encompassing theory was that it became tautological and ultimately everything became an instinct (Reeve, 2015).

Drive. Finally, the third grand theory, introduced first by Freud in 1915 and further expanded on by Hull (1943) focused on how motivation originated from biological needs such that behavior was motivated to reduce drive and meet the needs of the body. Freud's view was formed through case studies of his psychoanalytic sessions with patients in which he asserted that biological urges in the body built up energy in the nervous system when not addressed resulted in psychological discomfort and produced

anxiety (Reeve, 2015). Freud's (1915) drive theory was summarized by four components, (a) source, (b) impetus, (c) object, and (d) aim. For example, if the source of the biological need was lack of nourishment/hunger, the impetus would be hunger pangs, the object of which would be food, and finally the aim would be finding and consuming the food to relieve the source of discomfort.

Hull (1943), like Freud, subscribed to a physiological basis with bodily deficit as the source for motivation. However, Hull used scientific methods such as experimental designs and random assignment as opposed to case study analyses to form the basis of his drive theory. Additionally, Hull's drive theory consisted of three main principles, (a) drive emerges from bodily needs, (b) drive energizes behavior, and (c) drive reduction is self-reinforcing and produces learning (i.e. habits). Although the first two principles of Hull's drive theory are similar to the first two components of Freud's (1915) theory, Freud specifically called out both the object that fulfilled the biological need and the fulfillment of the need, whereas Hull (1943 & 1952) combined the two in a single element and emphasized the habit/learning that resulted in the successful fulfillment of the biological need. Furthermore, Reeve (2015) indicated that what separated Hull (1943 & 1952) from other theorists was his attempt to quantitatively account for his theory by providing a formula for predicting when motivation was likely to occur.

$$_{s}E_{r} = {}_{s}H_{r} \times D \times K$$

Where E is the strength of the energized behavior, which is embedded in the pairing of stimulus/response (i.e. little s and r), H is the strength of the habit or learning (also embedded in the pairing of stimulus/response), D is the internal drive motivation, and K is the environmental incentive motivation. Worth noting is that K

was added to the formula by Hull in 1952 in an effort to begin to account for circumstances that were beyond the bounds of the theory and is contrary to his second premise.

Ultimately drive theory fell out of favor as a grand theory because there were circumstances that challenged each of the aforementioned premises. For instance, anorexia does not emerge from bodily needs. Additionally, external sources of motivation such as advertising (e.g., a potato chip commercial) may prompt someone who was not hungry to suddenly feel hungry, again this is contrary to Hull's second premise. Finally, learning can occur without drive reduction. A teacher may provide candy in the classroom to motivate students to learn but doing so does not provide nutritional benefit to them and therefore, does not reduce the need for nutrition (Reeve, 2015). The fall of the grand theories gave way to a series of mini theories.

Mini theories. The appeal of mini theories was the limited scope in which they explained motivation. Mini theories sought to understand or examine a single motivational phenomenon, circumstance or theoretical question. The following list outlines some of the mini theories that emerged from the late 1950s to the 1970s:

- Cognitive dissonance theory (Festinger, 1957)
- Effectance motivation (White 1959)
- Achievement motivation theory (Atkinson, 1964)
- Expectancy X value theory (Vroom, 1964)
- Psychological reactance theory (Brehm, 1966)
- Goal-setting theory (Locke, 1968)
- Attributional theory of achievement motivation (Weiner, 1972)

- Cognitive evaluation theory (Deci, 1975)
- Flow theory (Csikszentmihalyi, 1975)
- Intrinsic motivation (Deci, 1975)
- Learned helplessness theory (Seligman, 1975)
- Self-efficacy theory (Bandura, 1977)
- Self-schemas (Markus, 1977)

Cognitive dissonance theory. Festinger (1957) hypothesized dissonance in two distinct ways. One which borrows from that of drive theory in that a component of dissonance involves an inner drive to be met; that which seeks to hold our attitudes and behaviors in agreement with one another. The other component of dissonance was that of psychological discomfort when attitudes and behaviors were inconsistent (Festinger, 1957). Cognitive dissonance theory, therefore, refers to that which compels one to achieve a necessary state of internal consistency between ones attitudes, beliefs or behaviors. When any are in conflict with one another a feeling of mental discomfort results requiring an adjustment in one of the attitudes, beliefs or behaviors to decrease the discomfort and restore an internally consistent state (McLeod, 2018).

Festinger (1957) further suggested that the magnitude of dissonance depends on, and is positively related to, the importance or value between the two cognitive elements. He (Festinger, 1957) proposed three ways in which dissonance could be reduced, (a) a change in attitudes, beliefs or behaviors, (b) a change in the environmental cognitive element, and (c) attainment of new information that balances the dissonant beliefs. However, reducing dissonance may be met with resistance due to a variety of factors such as pain or loss resulting from a change in behavior, the feeling of satisfaction from

the current behavior, or a change may not be possible (Festinger, 1957). In closing Festinger (1957) maintained that the chief factor in the attempt to reduce dissonance is the amount of resistance to change rather than the source of resistance. In contrast to cognitive dissonance theory, effectance motivation is not derived from drive theory, rather it shared biological roots in the nervous system.

Effectance motivation. In This proposed theory of motivation, White (1959) focused on addressing the gaps for which he perceived in both instinct and drive as the basis for motivation. In his seminal paper, White (1959) described an organism's capacity to interact effectively with its environment as one which has *competence*. White (1959) proposed competence as that which is obtained through activities exhibiting direction, selectivity, and tenacity in interacting with one's environment. Furthermore, he suggested competence had a motivational aspect, and that the motivation needed to attain competence could be derived neither through drive nor instinct as its sole source of energy.

White (1959) deemed this motivational aspect of competence as *effectance,* or that which produces a feeling of efficacy. He postulated effectance not as a deficit motive, rather as one which draws on the nervous system as its source of energy and for which environmental stimulation is secondary. In contrast to drive theory where response or behavior occur as a result of homeostatic crisis, effectance motivation occupies the waking time between such events (White, 1959).

Effectance motivation was aroused by stimulus conditions that varied somewhat from the original stimulus, peaks when novelty was at the fore and diminished when conditions did not vary sufficiently to produce new effects or possibilities (White, 1959).

For example, a teenager may be motivated to play a new video game because he had played a similar game in the past and enjoyed doing so. If there were no variation between the new game and the previously mastered game, he would be less likely to continue play. However, if there were new challenges, tools, or rules of play that effected the behavior or outcome of the game, he would be more likely to persist in this play. While White focused on addressing the gaps he perceived in instinct and drive theories for explaining the basis for motivation, achievement motivation theory sought to examine a person's tendency to achieve success and avoid failure as a vehicle of motivation (Atkinson, 1964).

Achievement motivation theory. Achievement motivation theory attempted "to account for the determinants of the direction, magnitude, and persistence of behavior" (Atkinson, 1964, p. 240) of human activities when individuals were aware their performance was being evaluated and for which consequences would yield favorable or unfavorable (success or failure) results. Atkinson (1964) asserted that ultimately achievement motivation was the difference between an individual's tendency to achieve success and avoid failure (i.e., $T_S - T_{AF}$).

According to this theory, the tendency to achieve success can be calculated as:

$T_S = M_S \times P_S \times I_S$

where S is success, *M* is the motive to achieve, *P* is the strength of the individual's subjective probability (expectancy) and *I* is the strength of the individual's incentive value.

An individual's motive to achieve success was seen by Atkinson (1964) to be a relatively stable personality disposition. Conversely, the strength of an individual's

probability and incentive values of success are deemed situational influences that are dependent on past experiences in the context of similar situations. These situational influences are represented by a continuous range from 0 to 1 and have values complementary to one another.

For example, when an individual's strength for the probability of success on a task is very high (e.g., .90) the respective strength of the incentive value for success in very low (e.g., .10) or $I_S = 1 - P_S$. Atkinson (1964) argued this is because when individuals attempt tasks for which they are highly likely to achieve success, they will feel less pride because the task was too easy. Contrariwise, when an individual's strength for the probability of success on a task is very low (e.g., .10) the respective strength of the incentive value for success is very high (e.g., .90) thus leading to a greater sense of pride when success is achieved.

Through examination of computed values for the tendency to achieve success it is shown that the highest tendency to achieve success is when the strength of the probability and incentive values of success are equal (see Table 1). It followed then that when motive

Table 1

Achievement Motivation Theory: Tendency to Achieve Success as a Function of Motive to Achieve, Expectancy of Success, and Incentive Value of Success

Task	P_s	I_s	T_s when $M_s = 1$	T_s when $M_s = 10$
1	.90	.10	.09	.90
2	.70	.30	.21	2.10
3	.50	.50	.25	2.50
4	.30	.70	.21	2.10
5	.10	.90	.09	.90

to achieve is strong (i.e., 10 as opposed to 1) the tendencies to achieve are more pronounced between tasks of varied probability of success.

Similar to the computation for the tendency to achieve success, the tendency to avoid failure can be computed by:

$T_{-f} = M_{AF} \times P_f \times I_f$

where M_{AF} is the motive to avoid failure, P_f is the strength of the individual's subjective probability (expectancy) of failure, and I_f is the strength of the individual's incentive value of failure.

An individuals' tendency to avoid failure has an inhibiting effect on achievement motivation as it represents the avoidance of shame or embarrassment should failure of the task occur and is strongest when the expectancy of success and failure are comparable. Similar to M_s, M_{AF} is a relatively stable personality disposition. However, P_f is complimentary to P_s in that if an individual feels he has a low probability of success on a task, it followed that he would feel a high probability of failure. Additionally, Atkinson (1964) accounted for the inhibiting effect of T_{-f} on achievement motivation by assigning the value of $I_f = -P_s$, therefore, when the perceived probability of success on a task is high ($P_s = .90$) then the incentive value for failure is very high as well (i.e., this task is easy therefore I should succeed, however, if I don't, I'll be extremely embarrassed). It follows then that when the perceived probability of success on a task is very low ($P_s = .10$) then the incentive value for failure would be low (i.e., this task is difficult therefore I would expect to fail, if instead I succeed then I will feel proud rather than embarrassed).

Somewhat like the tendency to achieve, where the highest tendency to achieve success is when the strength of the probability and incentive values of success are equal,

the tendency to avoid failure is most detrimental to achievement motivation when the strength of the probability and incentive values of failure are equal (i.e., has the highest negative value) (see Table 2). Additionally, when motive to avoid failure is strong (i.e., 10 as opposed to 1) the tendencies to avoid failure are more pronounced between tasks of varied probability of success.

Table 2

Achievement Motivation Theory: Tendency to Avoid Failure as a Function of Motive to Avoid Failure, Expectancy of Failure, and Incentive Value of Failure

Task	P_f	I_f	T_{-f} when $M_{AF} = 1$	T_{-f} when $M_{AF} = 10$
1	.90	-.10	-.09	-.90
2	.70	-.30	-.21	-2.10
3	.50	-.50	-.25	-2.50
4	.30	-.70	-.21	-2.10
5	.10	-.90	-.09	-.90

When examining the impact of each, the tendency to achieve and the tendency to avoid failure, it is evident that achievement motivation is most negatively impacted when the motive to avoid failure outweighs the motive to succeed. It is most positively impacted when the motive to succeed outweighs the motive to avoid failure. However, there is no impact on achievement when both the motive to succeed and to avoid failure are equally weighted (see Tables 3 - 5 respectively).

Table 3

Achievement Motivation Theory: The Motive to Avoid Failure Outweighs the Motive to Succeed

	When $M_s = 1$ and $M_{AF} = 10$						
Task	P_s	I_s	T_s	P_f	I_f	T_{-f}	$T_s + T_{-f}$
1	.90	.10	.09	.90	-.10	-.90	-.81
2	.70	.30	.21	.70	-.30	-2.10	-1.89
3	.50	.50	.25	.50	-.50	-2.50	-2.25
4	.30	.70	.21	.30	-.70	-2.10	-1.89
5	.10	.90	.09	.10	-.90	-.90	-.81

Table 4

Achievement Motivation Theory: The Motive to Succeed Outweighs the Motive to Avoid Failure

	When $M_s = 10$ and $M_{AF} = 1$						
Task	P_s	I_s	T_s	P_f	I_f	T_{-f}	$T_s + T_{-f}$
1	.90	.10	.90	.90	-.10	-.09	.81
2	.70	.30	2.10	.70	-.30	-.21	1.89
3	.50	.50	2.50	.50	-.50	-.25	2.25
4	.30	.70	2.10	.30	-.70	-.21	1.89
5	.10	.90	.90	.10	-.90	-.09	.81

Table 5

Achievement Motivation Theory: The Motive to Succeed and to Avoid Failure are Equally Weighted

	When $M_s = 1$ and $M_{AF} = 1$						
Task	P_s	I_s	T_s	P_f	I_f	T_{-f}	$T_s + T_{-f}$
1	.90	.10	.09	.90	-.10	-.09	0
2	.70	.30	.21	.70	-.30	-.21	0
3	.50	.50	.25	.50	-.50	-.25	0
4	.30	.70	.21	.30	-.70	-.21	0
5	.10	.90	.09	.10	-.90	-.09	0

Expectancy value theory. It is not surprising to find commonalities between achievement motivation and expectancy value theories given the similar time frame in theory development. The notion of an individual's incentive values and subjective probability of success (Atkinson, 1964) and valence and expectancies (Vroom, 1964) portray similar concepts. The development of Expectancy Theory (ET), however, represented a more narrowed view with a focus on motivation related to employees in the workplace (Vroom, 1964).

Vroom's theory was based on three beliefs, (a) valence, (b) expectancy, and (c) instrumentality, that were necessary factors of motivation ("Vroom's expectancy theory," n.d.). Valence referred to that which was of extrinsic (e.g., money) and intrinsic (e.g., fulfillment) value to the employee (Mullder, 2018; "Vroom's expectancy theory," n.d.). Expectancy was centered around the employees' level of confidence in their capabilities and expectations relative to a job well done (Mullder, 2018; "Vroom's expectancy

theory," n.d.). Finally, instrumentality referred to the extent to which the employees performance in the workplace was rewarded as expected (Mullder, 2018; "Vroom's expectancy theory," n.d.).

Dissimilar to achievement motivation theory, Vroom (1964) suggested that motivation was a product of its contributing factors, whereas, Atkinson (1964) described motivation as a factor of achievement motivation theory. Specifically, it was Vroom's (1964) assumption that motivational force was the product of the interaction between valence, expectancy and instrumentality. As we transition to the next mini theory, yet another factor, autonomy, is considered as contributing to motivation.

Psychological reactance theory. In 1966, psychologist Brehm introduced his theory of psychological reactance. He argued that when people are free to engage in a given behavior and when that freedom is subsequently threatened or eliminated, the result is psychological reactance. Brehm defined psychological reactance as, "a motivational state directed toward the reestablishment of whatever freedom has been threatened or eliminated" (p. 15). Additionally, the larger the perceived threat to a behavioral freedom, the greater the resistance against that threat (Brehm & Sensenig, 1966). Furthermore, the magnitude of reactance and consequent amount of resistance is a direct function of the possible implication of threats to further behavioral freedoms (Brehm & Sensenig, 1966).

Brehm and Brehm (1981) initially indicated there was no direct measure of psychological reactance. However, as the theory advanced subsequent research in both neuroscience (Mühlberger, Klackl, Sittenthaler, & Jonas, 2019; Steindl, Jonas, Sittenthaler, Traut-Mattausch, & Greenberg, 2015; Steindl, Klackl, & Jonas, 2016) and psychology (Dillard & Shen, 2005; Hong & Faedda, 1996) contributed evidence to the

contrary. Mühlberger et al. (2019) used electroencephalography (EEG) to examine left frontal asymmetry, a recognized indicator of approach motivation, to identify reactance by manipulating different kinds of freedom restrictions through subject readings of various scenarios. Similarly, functional magnetic resonance imaging (fMRI) was used in the same manner (Steindl et al., 2015, 2016). Self-reported measures have also been validated in measurement of reactance. Hong and Faedda (1996) constructed a unidimensional model comprised of 11 items assessing emotional response, reactance to compliance, resisting influence and reactance to avoidance. Correspondingly, Dillard and Shen (2005) distilled reactance down to self-reported indices of anger and negative cognitions.

Tangential to Hull's drive theory (1943), psychological reactance theory sought to restore one to a sense of equilibrium. However, the homeostatic state one sought to restore was that of autonomous equilibrium. At the core of reactance theory is the reinstatement of a prior freedom.

Goal-setting theory. The goal-setting theory of motivation asserted that goals and intentions regulated performance (Locke, 1968). Whereby, the degree to which one subscribed to or took ownership of the goal was the key factor in creating motivation (Locke, 1968). Locke and Latham (2002) outlined four mechanisms through which goals affected performance, (a) as a directive function, (b) as an energizing function, (c) persistence, and (d) action.

Goals provided direction by guiding behavior towards goal-focused activities and away from activities that were not perceived to provide support in obtaining ones goals (Locke & Latham, 2002). As an energizing function, goals allocated less energy to

low/easy goals (e.g., 10% increase in sales/mastering single digit addition) and more energy towards high/hard goals (e.g., 50% increase in sales/using probability to evaluate outcomes of decisions) (Locke & Latham, 2002). Similarly, less persistence was shown for low/easy goals while greater persistence was shown for high/hard goals when participants controlled the time spent on tasks (LaPorte & Nath, 1976). Lastly, goals could indirectly affect action by prompting the discovery and/or use of task-relevant knowledge and tactics (Locke & Latham, 2002). While goal-setting theory attributed goals and intentions to performance (Locke, 1968), within the attributional theory of motivation, effort and ability are examined as sources of motivating performance (Weiner, 1972).

Attributional theory of achievement motivation. Attribution in this sense relates to the action of regarding something as being caused by a person or a thing (Attribution, 2019). Therefore, the attributional theory of achievement motivation referred to actions that are motivated based on effort and/or ability as perceived causes of demonstrated performance where effort is an unstable attribute (i.e., under personal control) and ability is a stable attribute (Weiner, 1972). Moreover, disparities in perceived causation tended to vary by individual differences in achievement needs as well as by those who evaluated achievement (Weiner, 1972; Weiner & Kukla, 1970)

Weiner and others (Kukla, 1972; Weiner, Heckhausen, Meyer, & Cook, 1972; Weiner & Potepan, 1970) found that students with high achievement motivation tended to attribute failure to a lack of effort (i.e., *I did not do well on the math test this time, so if I put more effort into strategies I use to prepare for testing, I will do better next time*). Whereas those with low achievement motivation tended to attribute failure to lack of

ability (i.e., *I did not do well on the math test because I'm not good at math and nothing I do will raise my score*) (Kukla, 1972; Weiner et al., 1972; Weiner & Potepan, 1970). With this in mind, Weiner (1972) suggested that the progress of achievement motivation is dependent upon, "the learning of cognitive structures which represent the causal importance of effort" (p.209).

Conversely, when an external entity perceives an individual to have the ability to perform well at a task (even when the individual does not) and the individual fails, the external entity attributes the failure to lack of effort rather than lack of ability (Weiner & Kukla, 1970). As such Weiner (1972) suggested that attributions not only affect achievement but rewards and punishments by external entities as well. In his closing words Weiner (1972) broached the issue of teacher training and the inclusion of an introduction of causal perception to raise greater awareness of the attributional process.

Cognitive evaluation theory. Cognitive evaluation theory deals with the effect of extrinsic rewards on intrinsic motivation (Deci, 1975). The working definition of intrinsic motivation upon which Deci described his theory was behavior that, "a person engages in so that he may feel competent and self-determining in relation to his environment" (p.v). Deci proposed three propositions that could affect intrinsic motivation. Proposition one stated that a change in perceived locus of causality from internal to external could affect intrinsic motivation. Proposition two stated that a change in feelings of competence and self-determination could affect intrinsic motivation. Finally, proposition three stated that the nature of rewards, including feedback, determined whether perceived locus of causality or feelings of competence and self-determination affected intrinsic motivation.

In perceived locus of causality, Deci (1975) asserted that a decrease in intrinsic motivation would occur when someone received an extrinsic reward for engaging in intrinsically motivated activities. However, the loss in intrinsic motivation only occurred when the perception of the individual was that the extrinsic reward was the sole purpose of behavior. It is the perception of being controlled that diminished the intrinsic motivation.

Deci (1975) stated that a change in feelings of competence and self-determination would also affect intrinsic motivation. Extrinsic rewards in this case would provide information that either heightened or diminished a person's feelings of competence and self-determination. Whereby, heightened feelings of competence and self-determination increased intrinsic motivation (e.g., when receiving a promotion), while a diminished sense of competence and self-determination will have the opposite affect (e.g., being passed over for a promotion).

It is the relative salience of these two characteristics of extrinsic rewards (i.e. controlling or informational) that drove which process occurred (Deci, 1975). A change in perceived locus of causality will be triggered when the more salient aspect of the reward is control. However, Deci stated that a change in feelings of competence and self-determination will be triggered when the informational aspect of the reward is more salient.

Flow theory. "The holistic sensation that people feel when they act with total involvement" (p.36) defined Csikszentmihalyi's (1975) flow theory. Csikszentmihalyi identified five conditions/components through which flow tends to occur. Together, the following conditions persist flow, (a) the activity was comparable in difficulty to a

person's skills, (b) the actor was hyper focused within a narrowly defined environment, (c) there was an absence of self-consciousness or ego, (d) the requirements and feedback were clear and consistent, and (d) the reward is the activity itself.

Csikszentmihalyi (1975) asserted the importance of an activity that is evenly matched with a person's ability as a key component of flow. When an activity far exceeded the skills of the actor the result was anxiety. Similarly, when the skills of the actor were somewhat lesser than that required for an activity the result was worry. Conversely, when the skills of the actor are greater than what was required for an activity the result is boredom. Thus, to remain in a state of flow the balance of activity difficulty and person ability becomes a necessary precursor in removing self-conscious behavior.

Flow also requires, what Csikszentmihalyi (1975) refers to as, a "centering of attention" (p.40) on a constrained stimulus field. Csikszentmihalyi described this type of focus as one in which the actor in unaware of anything outside of his environment when within the activity. For this type of laser focus to occur and persist task and ability need to be evenly matched, self-conscious behavior and ego must not exist during the time in flow and the ability of the actor is such that what is required within the activity and the manner in which to respond to feedback in necessary has reached a level of automaticity within the actor (Csikszentmihalyi, 1975).

The final condition of flow is that the reward or goal is not an entity separate from the activity (Csikszentmihalyi, 1975). Instead, the activity itself is the reward or goal. The result of flow is that the person finds the process as intrinsically motivating. Csikszentmihalyi described it best when he said, "the purpose of the flow is to keep on flowing, not looking for a peak or utopia, but staying in the flow" (p.47).

Intrinsic motivation. Deci (1975) defined intrinsic motivation as, "behaviors which a person engages in to feel competent and self-determining" (p.61) in relation to their environment . Thus, intrinsically motivated behaviors are those for which rewards are inherent to the person. According to Deci there are two classes of intrinsically motivated behaviors, seeking and conquering.

People are said to seek out situations that challenge them (Deci, 1975). Similar to components of flow theory, Deci stated that the situation sought after would be one within a person's ability to deal with successfully. Furthermore, he suggested situations that were insufficiently challenging would lead to boredom. In contrast to seeking behavior, conquering behavior focused on challenges that one encountered or created, rather than sought out, and commonly involved reduction in dissonance or uncertainty. Like White (1959) and effectance theory, Deci (1975) attributed the source of energy for intrinsic motivation to involve the needs of the central nervous system.

Learned helplessness theory. Learned helplessness is a state of passively enduring a threat after repeated exposure to events that are perceived to be unavoidable (Maier & Seligman, 1975). This state is primed by the expectation that one's actions and the outcome are independent of one another (Seligman, 1975). Seligman posited that it is the perception that one has no control which fuels a state of helplessness.

Maier and Seligman (1975) described learned helplessness as a three step process, where the subject receives information relative the relationship between their response or action and the outcome, perception of the relationship is then formed, which ultimately affected behavior. In this theory Seligman (1975) stated that when responses and outcomes are perceived as independent, motivation to attempt to control the output was

reduced, learning was retarded, and greater emotional disruption would occur. In what follows, Bandura (1977) differentiated between expectancies resulting from the relationship between response and outcome as compared to expectancies and self-efficacy.

Self-efficacy theory. Bandura (1977) differentiated efficacy expectancies as the conviction that a person could execute the behavior, whereas response-outcome expectancies was one's estimate that a given behavior would lead to certain outcomes. Self-efficacy, then, referred to one's beliefs in his or her ability in such a way as was necessary to yield explicit performance goals. Bandura hypothesized four primary sources of information from which personal efficacy was derived. These sources included, performance accomplishments, vicarious experience, verbal persuasion, and physiological states. Bandura also suggested that each varied in the degree to which they affected personal efficacy and persisted.

Experience of mastery arising from effective performance is said to have the greatest and longest lasting effect on personal efficacy (Bandura, 1977). Repeated successes raised mastery expectations increasing self-efficacy, while failures had the opposite effect. In addition to the frequency of successes or failures, Bandura proposed timing also plays a pivotal role in the strength of personal efficacy. He indicated initial failures followed by repeated successes strengthened self-efficacy, similarly an occasional failure among many successes would not reduce personal efficacy. Although performance accomplishments are said to have the greatest effect on self-efficacy, the remaining sources of information also differentially impact self-efficacy.

The feedback obtained through vicarious experiences provided opportunities for learning new behaviors (Bandura, 1977). Applying such information to one's own circumstances served as a guide in taking corrective action on subsequent behaviors. However, Bandura suggested, on its own, vicarious experiential information was less dependable and more susceptible to change than that obtained through personal experience.

Much like information derived through vicarious experiences, efficacy expectations brought about by verbal persuasion lacks personal experience at its core (Bandura, 1977). It followed then that verbal persuasion would likely result in weaker efficacy expectations. For persons with a history of failure, any gain in mastery expectations could be easily stifled by confounding experiences.

Additionally, Bandura (1977) proposed physiological states such as stress and anxiety might elicit emotional arousal. Subsequently, emotional arousal could impact one's competency and thus their performance (Bandura, 1977). Therefore, in perceptually threatening or anxiety inducing situations self-efficacy is more likely to be negatively affected. In closing, Bandura's hypothesis, in terms of motivation, asserted that expectations of personal efficacy were major determinants in whether behavior would be initiated, the amount of effort that would be expended, and the sustained duration of effort in the face of adversity.

Self-schemas. Self-schematic was defined as the, "cognitive generalizations about the self, derived from past experience, that organize and guide the processing of the self-related information contained in an individual's social experience" (Markus, 1977, p.64). Within self-schema theory, Markus suggested motivation as a function of cognitive self-

representation such that self-schemata mediated the correspondence between self-categorization and overt behavior. Additionally, she proposed a latency component for people with self-schemas relative to a particular dimension of behavior. Whereby those who readily endorsed a behavior based on self-schemas did so with more automaticity than those who did not have similar self-schemas.

Choice theory. Choice theory asserted that individuals only have the power to control themselves and have limited power to control others (Glasser, 1998). The foundational core on which it is based includes 10 axioms (Glasser, 1998):

- The only behavior you can control is your own
- All we can give/get to or from other people is information
- All long-lasting psychological problems are relationship problems
- The problem relationship is always part of our present lives
- Painful past experiences have shaped us, but revisiting them contributes little to what we need to do now
- We are driven by five genetic needs: survival, love and belonging, power, freedom, and fun
- These needs are only satisfied if they are part of our quality worlds
- Human behavior is made up of four inseparable components: acting, thinking, feeling, & physiology
- Behavior is designated by verbs, usually infinitives and gerunds, and named by the component that is most recognizable
- All total behavior is chosen, but we have direct control over only the acting and thinking components

In the section that follows I have elaborated on Glasser's (1998) sixth axiom. Within the sixth axiom I have identified which components were applicable to my current study. Finally, I have tied it to aspects of motivation theory and provided examples from the literature where both have been examined in the context of education.

Applicable Components of Choice and Motivation

My current study draws on both choice and motivation theory. Within choice theory power, freedom, and fun are aspects of the sixth axiom, *we are driven by five genetic needs*, that are applicable to my current study. Additionally, components of several of the mini-theories including achievement motivation, expectancy value, psychological reactance, cognitive evaluation, and intrinsic motivation apply. Connections across theories as well as unique contributions follow.

Glasser (1998) suggested that humans are motivated to act based on five needs of which he believed we were genetically programmed to satisfy. Three of those needs, freedom, power, and fun are applicable to the current study. In terms of power, Glasser referred to our need to have or do things our way, to tell others what to do, how to do it, and to see them do it, but not in their own way, rather in the way we believe it should be done because our way is the best way. Freedom, then, comes into play as a counterbalance to power. Glasser posited that freedom speaks to our need to conduct ourselves the way we choose, irrespective of the wants of others and under our own control. In terms of fun, he suggested it was our "genetic reward for learning" (p.41).

In psychological reactance Brehm (1966) discussed a person's freedom to engage in a behavior that was subsequently withheld as a motivating condition. This ties into the first proposition of cognitive evaluation theory, whereby a change in perceived locus of

causality was proposed by Deci (1975) as impacting intrinsic motivation. In a similar vein, Glasser (1998) described freedom as a genetic need in terms of the threat of losing the freedom to choose as motivated behavior.

By providing assessments where the person in control of the type of assessment taken by the student is no longer an authoritative figure (e.g., testing company, administrator, teacher, etc.), the perception of power can shift. In this scenario, the student is in power. Creating this shift for students can be a motivating factor in student performance.

Glasser (1998) suggested that freedom was a counterbalance to power. Affording students the freedom to select the form of a test suggests that they have some control over their testing experience. Producing tests that provide students with a choice in context can be a counterbalance that is sufficient to engage students while taking an assessment.

Assessments connected to student interest outside of the classroom can be an avenue for more enjoyable testing experiences as inferred by prior research on interest and learning (Anand & Ross, 1987; Bernacki & Walkington, 2014; Brozo et al., 2014; Ivey & Broaddus, 2001; Walkington, 2013; Walkington & Leigh, 2015). Glasser's (1998) *Choice Theory* seems to support this in his view that *fun* is our genetic reward for learning. Introducing an element of fun, by providing assessment context matched to a students interests, can serve as an engaging factor for student performance in testing situations.

Similarly, connections are evident between valence and individual incentive values. The valence aspect of expectancy value theory refers to extrinsic and intrinsic values of a person as a motivating factor of performance (Vroom, 1964). Likewise,

Atkinson (1964) suggested that an individual's incentive to succeed or to avoid failure motivated achievement.

The next chapter introduces the methodology for this study. It describes the instrumentation, sample, and procedures used as well as the analyses employed to address the research questions introduced in Chapter I.

CHAPTER II

METHODS

The methods used in this study have been depicted in four sections. The sections include discussion of the instrument, scales, student sample, and procedures. An overview of sections one through four are provided in the flowing narrative.

I begin with a description of the instrument used to collect the data. Within section one I explain the assessment instrument, including its form, length, target grade range, and content alignment. Additionally, I provide the methods used to construct the test, including the identification of the item sample used, actual item selection, and subsequent item cloning. In section two I discuss the nature of each of the scales utilized. Discussion of the methods used to identify the student sample, including the sampling design and power analysis follow in section three. I conclude in section four where I have outlined the procedures employed in this study. The procedures include selection for school participation, assessment administration, and data analyses used in response to the research questions.

Instrument

Context personalization assessment. The Context Personalization Assessment prototype is a set of fixed-form, 24-item reading assessments. In my current study, the prototyped tests assessed content from the CCSS in ELA for grades 6, 7, and 8. The specific standards assessed included (a) Literature, (b) Informational Text, and (c) Vocabulary Use and Acquisition (NGA & CCSSO, 2010). A subset of skills across standards included: key ideas and details, craft and structure, integration of knowledge and ideas, and range of reading and level of text complexity.

Literature (NGA & CCSSO, 2010)**.**

- Students should be able to cite, determine, describe, and analyze key ideas and details of the text
- Students should be able to determine, analyze, explain, and compare and contrast the craft and structure of the text
- Students should be able to compare and contrast, and analyze the integration of knowledge and ideas in the text
- Students should be able to read and comprehend the following text types: stories, drama, and poetry.

Informational text (NGA & CCSSO, 2010)**.**

- Students should be able to cite, determine, and analyze key ideas and details of the text
- Students should be able to determine and analyze the craft and structure of the text
- Students should be able to trace and evaluate, compare and contrast, analyze, delineate and evaluate, and integrate knowledge and ideas in the text
- Students should be able to read and comprehend the following text types: literary nonfiction.

Vocabulary use and acquisition (NGA & CCSSO, 2010)**.**

- Students should be able to "determine the meaning of unknown and multiple-meaning words" (p.53)
- Students should be able to demonstrate their "understanding of figurative language, word relationships, and nuances in word meanings" (p. 53)

- Students should be able to "acquire and use accurately" (p. 53) grade level vocabulary

Assessment construction. A total of four assessments were constructed, one for each of the three most popular reading contexts as identified by students in grades 6, 7, and 8 who participated in the student interest survey referenced in the prior section of this paper, and a fourth assessment was constructed using items with their original context. Items were ordered from least to most difficult, balancing across goals throughout the assessment. Item order was the same for all versions of the assessment. See *Appendix A* for an example item.

Item sample identification. The pool from which the items were drawn supported computer adaptive tests that spanned grades 2-11, where selection was balanced by goal and item difficulty relative to student ability; grade level was not within the criteria for selection. A cluster sampling technique was used (random stratified sampling at two levels) to identify a sample from which to select reading items for use within the context-personalization assessments. Stratification was based on item difficulty and content at the goal level. Item difficulty level was reported as a RIT[1] scale value and represented a logit transformation of the item after it had been calibrated. The criteria for range of item difficulty was based on fall norms of student reading achievement for sixth through eighth grades published by the Northwest Evaluation Association [NWEA] (Thum & Hauser, 2015). The range utilized was based on the medial quintile of the fall norms across the three grades resulting in a RIT range from 209

[1] An equal interval measure that is one tenth of a logit added to 200. A unit that is derived from test data by applying the Poisson probability theorem. Rasch units, is a name coined by curriculum and evaluation researchers to avoid confusion with other measures (Ingebo, 1997, p. 143)

to 219 (see *Appendix B*). The criteria for content was based on the Common Core State Standards for English Language Arts as they corresponded to sixth through eighth grades (NGA & CCSSO, 2010). Goals included informational text, literature, and vocabulary acquisition and use. The initial extraction of sample items was retrieved from NWEA's item bank resulting in a sample of 6422 items.

Item selection. The resulting item sample was imported into Excel to which a column was added (Random#1) and populated using a random number generating function (=RAND). Items were then sorted by goal, RIT, and Random#1 and selected contingent upon the lowest Random#1 value for each difficulty level (209-219) in each of the three goal areas resulting in items at each of the 11 difficulty levels in the identified range across 2 of the goals, *informational text* and *vocabulary use and acquisition,* but only at 8 of the item difficulties for the goal *literature*. Therefore, 8 items from each goal area were selected in order to achieve tests constructed with 24 items balanced both by content and difficulty.

Item cloning. The resulting 24 items were provided to reading content specialists to create item clones using topics identified through the student interest survey. Content specialists identified nine items in total that would not translate to different contexts. Five from the goal area *informational text*, three from the goal area *literature*, and a single item from the goal area *vocabulary acquisition and use*. Substitute items were selected based on next lowest random#1 value for the same goal area and at the same RIT level. In order to maximize the likelihood that items with the new context would be comparable to items with the original context content specialists were instructed to, (a) write to the same grade level as indicated in the original, (b) write to the same depth of knowledge

(DOK[2]), (c) adhere to the Flesch-Kincaid's grade level readability ± .2 of the original item's readability, (c) use the same number of answer options as the original, (d) persist the correct answer at the same position as the original, and (e) as much as possible persist the same structure and wording as the original. The item transformations resulted in 46% (i.e., 11/24 items) as clones matching all 6 of the criteria with the remaining 44% of the items matching on the first 5 of the 6 criteria. Item clones went through two item reviews by subject matter experts prior to approval for use within the new assessment prototypes. Tables 6-29 depict the four versions (original, animals, fantasy, & sports) of each of the 24 items, the depth of knowledge assessed by each, the target grade to which the item was written based on the Common Core State Standards (NGA & CCSSO, 2010), corresponding word count, readability as measured by Flesch-Kincaid, the content goal name assessed, and the corresponding standard to which each was written.

Table 6

Item 1 Attributes - DOK is 2 and Target Grade is 9 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	15	NA	Vocabulary Acquisition and Use	CCSS.ELA-Literacy.L.9-10.4a
Animals	14	NA		
Fantasy	18	NA		
Sports	15	NA		

[2] Depth of knowledge refers to a framework used to identify the level of cognitive complexity associated with an assessment item/task of which there are four progressively complex levels (i.e., recall, skills/concepts, strategic thinking, & extended thinking) (Webb, 1999).

Table 7

Item 2 Attributes - DOK is 2 and Target Grade is 9 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	67	12.1	Informational Text	CCSS.ELA-Literacy.L.9-10.5
Animals	121	12.0		
Fantasy	148	12.1		
Sports	82	11.9		

Table 8

Item 3 Attributes - DOK is 2 and Target Grade is 7 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	103	2.7	Literature	CCSS.ELA-Literacy.RL.7.4
Animals	111	2.7		
Fantasy	104	2.6		
Sports	98	2.5		

Table 9

Item 4 Attributes - DOK is 2 and Target Grade is 9 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	11	NA	Vocabulary Acquisition and Use	CCSS.ELA-Literacy.L.9-10.4a
Animals	19	NA		
Fantasy	12	NA		
Sports	17	NA		

Table 10

Item 5 Attributes - DOK is 2 and Target Grade is 4 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	54	7.6	Informational Text	CCSS.ELA-Literacy.RI.4.2
Animals	73	7.7		
Fantasy	75	7.7		
Sports	51	7.7		

Table 11

Item 6 Attributes - DOK is 3 and Target Grade is 7 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	256	3.0	Literature	CCSS.ELA-Literacy.RL.7.6
Animals	282	3.2		
Fantasy	344	3.2		
Sports	185	3.1		

Table 12

Item 7 Attributes - DOK is 1 and Target Grade is 11 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	6	NA	Vocabulary Acquisition and Use	CCSS.ELA-Literacy.L.8.5b
Animals	6	NA		
Fantasy	9	NA		
Sports	6	NA		

Table 13

Item 8 Attributes - DOK is 2 and Target Grade is 6 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	156	5.7	Informational Text	CCSS.ELA-Literacy.RI.9-10.1
Animals	122	5.9		
Fantasy	97	5.9		
Sports	179	5.5		

Table 14

Item 9 Attributes - DOK is 3 and Target Grade is 9 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	394	9.2	Literature	CCSS.ELA-Literacy.RL.9-10.5
Animals	509	9.1		
Fantasy	402	9.1		
Sports	319	9.0		

Table 15

Item 10 Attributes - DOK is 1 and Target Grade is 6 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	47	NA	Vocabulary Acquisition and Use	CCSS.ELA-Literacy.L.5.6
Animals	49	NA		
Fantasy	47	NA		
Sports	49	NA		

Table 16

Item 11 Attributes - DOK is 1 and Target Grade is 6 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	244	7.4	Informational Text	CCSS.ELA-Literacy.RI.9-10.1
Animals	279	7.5		
Fantasy	263	7.5		
Sports	262	7.6		

Table 17

Item 12 Attributes - DOK is 2 and Target Grade is 5 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	303	4.3	Literature	CCSS.ELA-Literacy.RL.5.3
Animals	251	4.3		
Fantasy	194	44.0		
Sports	213	4.4		

Table 18

Item 13 Attributes - DOK is 2 and Target Grade is 4 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	78	5.7	Vocabulary Acquisition and Use	CCSS.ELA-Literacy.L.4.6
Animals	136	5.8		
Fantasy	71	5.6		
Sports	96	5.5		

Table 19

Item 14 Attributes - DOK is 2 and Target Grade is 6 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	8	NA	Informational Text	CCSS.ELA-Literacy.RI.6.6
Animals	9	NA		
Fantasy	9	NA		
Sports	9	NA		

Table 20

Item 15 Attributes - DOK is 1 and Target Grade is 9 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	467	3.2	Literature	CCSS.ELA-Literacy.RL.9-10.5
Animals	351	3.0		
Fantasy	332	3.0		
Sports	224	3.4		

Table 21

Item 16 Attributes - DOK is 2 and Target Grade is 3 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	11	NA	Vocabulary Acquisition and Use	CCSS.ELA-Literacy.L.4.5c
Animals	13	NA		
Fantasy	11	NA		
Sports	12	NA		

Table 22

Item 17 Attributes - DOK is 2 and Target Grade is 3 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	151	12.1	Informational Text	CCSS.ELA-Literacy.RI.7.8
Animals	200	12.3		
Fantasy	323	12.0		
Sports	362	12.1		

Table 23

Item 18 Attributes - DOK is 2 and Target Grade is 6 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	543	6.6	Literature	CCSS.ELA-Literacy.RL.6.3
Animals	802	6.8		
Fantasy	528	6.8		
Sports	621	6.5		

Table 24

Item 19 Attributes - DOK is 2 and Target Grade is 7 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	8	NA	Vocabulary Acquisition and Use	CCSS.ELA-Literacy.R.7.5b
Animals	9	NA		
Fantasy	8	NA		
Sports	8	NA		

Table 25

Item 20 Attributes - DOK is 1 and Target Grade is 6 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	166	3.2	Informational Text	CCSS.ELA-Literacy.RI.6.1
Animals	133	3.2		
Fantasy	160	3.2		
Sports	163	3.0		

Table 26

Item 21 Attributes - DOK is 2 and Target Grade is 9 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	543	4.2	Literature	CCSS.ELA-Literacy.RL.8.3
Animals	213	4.2		
Fantasy	166	4.4		
Sports	232	4.2		

Table 27

Item 22 Attributes - DOK is 2 and Target Grade is 8 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	66	5.7	Vocabulary Acquisition and Use	CCSS.ELA-Literacy.L.8.4c
Animals	63	5.8		
Fantasy	69	5.7		
Sports	73	5.9		

Table 28

Item 23 Attributes - DOK is 2 and Target Grade is 7 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	79	7.3	Informational Text	CCSS.ELA-Literacy.RI.7.4
Animals	56	7.2		
Fantasy	80	7.3		
Sports	70	7.2		

Table 29

Item 24 Attributes - DOK is 2 and Target Grade is 7 for all Versions of This Item

Version	Word count	FK	Goal Name	Standard
Original	141	7.9	Literature	CCSS.ELA-Literacy.RL.7.2
Animals	211	8.0		
Fantasy	246	8.1		
Sports	164	8.1		

The Flesch-Kincaid grade level readability formula is computed as:

$$FK = (0.39 \times ASL) + (11.8 \times ASW) - 15.59$$

where FK = the Flesch-Kincaid grade level, ASL = the average sentence length (i.e. average number of words per sentence), and ASW = the average number of syllables per word (Kincaid, Fishburne, Rogers, & Chissom, 1975). Readability scores should be read as grade level followed by months of instruction. Therefore, interpretation of a readability score of 6.3 would be that a sixth grade student at approximately the third month of instruction would be able to read the text.

Scales

The assessment scores for students represent an equal interval scale and were computed using an IRT method of estimation where scores are reported in logits. Scores of effort are reported by three measures, (a) solution behavior, (b) response time effort (RTE), and (c) response time fidelity (RTF). Solution behavior is represented by a dichotomous index where 0 indicates a non-effortful response and 1 indicates an effortful response. Response time effort and response time fidelity are represented by continuous indices from 0 to 1. Values closer to 1 indicate greater effort on the *assessment* while values further from 1 indicate less effort on the *assessment* for RTE. Similarly, values indicate students effort towards *an item* for RTF.

Sample

Sampling design. Participants in the Context Personalization assessment data set represented a convenience sample comprised of 577 middle school students in sixth through eighth grades. Students attended public, charter or private schools from Illinois, North Carolina, Oregon, and Washington. The demographic composition of the original sample was 48.7% female, 45.8% male, 3.2% preferred not to say, and 2.4% non-binary, third gender. Ethnicity was somewhat representative of a national sample (national percentages[3] provided in parentheses) consisting of 61.6% (48.9%) White, 10.1% (3.4%) 2 or more races, 7.8% (15.4%) African American, 7.3% (5.3%) Asian\Pacific Islander, 6.8% (25.9%) Hispanic, 4.9% (NA) Other, and 1.6% (1%) Native American\Alaskan Native. However, one or more responses for 60 students were missing from the data set.

[3] SOURCE: U.S. Department of Education, National Center for Education Statistics, Common Core of Data (CCD), "State Nonfiscal Survey of Public Elementary and Secondary Education," 2000–01 and 2015–16; and National Elementary and Secondary Enrollment Projection Model, 1972 through 2027. See Digest of Education Statistics 2017, table 203.50.

These test events were omitted from the analyses. The final sample size was 517. Demographics, counts, and percentages by state and gender are depicted in Table 30. Similarly, demographics are provided for students by assessment group. In Table 31 the number of students by assessment group, grade and gender have been reported.

Table 30

Demographic Data n(%)

Variable	Illinois	North Carolina	Oregon	Washington	Total
Gender					
Female	35(50.7)	145(47.5)	68(51.1)	34(50)	255(49.3)
Male	32(45.1)	138(45.2)	60(45.1)	32(47.1)	262(45.4)
Non-binary third gender	†	†	†	†	14(2.4)
Unidentified	†	13(4.3)	†	0	18(3.1)
Race / Ethnicity					
African American	†	37(12.1)	†	†	45(7.8)
Asian or Pacific Islander	†	28(9.2)	†	†	42(7.3)
Hispanic	†	19(6.2)	14(10.5)	†	39(6.8)
Native American or Alaskan Native	0	†	0	†	9(1.6)
Other	†	19(6.2)	†	†	28(4.9)
Two or more races	13(18.3)	33(10.8)	†	†	58(10.1)
White	46(64.8)	163(53.4)	102(76.7)	45(66.2)	356(61.7)

Note. † indicates student count less than 10.

Table 31

Preference Groups by Grade and Gender n(%)

Variable		Did not get preference	Got preference	No preference	Grade total
Grade 6		73(44)	74(44.6)	19(11.4)	166(32.1)
	Female	39(53.4)	41(55.4)	†	
	Male	32(43.8)	32(43.2)	12(63.2)	
	Non-binary third gender	0	0	0	
	Unidentified	†	†	†	
Grade 7		92(44.7)	96(46.6)	18(8.7)	206(39.8)
	Female	39(42.4)	50(52.1)	12(66.7)	
	Male	51(55.4)	42(43.8)	†	
	Non-binary third gender	0	†	†	
	Unidentified	†	0	0	
Grade 8		67(46.2)	57(39.3)	21(14.5)	145(28)
	Female	32(47.8)	29(50.9)	†	
	Male	23(34.3)	26(45.6)	11(52.4)	
	Non-binary third gender	†	†	0	
	Unidentified	†	†	†	
Preference total		232(44.9)	227(43.9)	58(11.2)	517

Note. † indicates student count less than 10.

Power analysis. In order to identify a sample size that was sufficient to obtain meaningful outcomes, it was necessary to conduct power analysis. As described here, the

level of power is the probability that a statistical test will yield a significant effect of the phenomenon under study. Statistical power is inversely related to the probability of making a Type II error (β)(i.e., power = 1 – β). Keppel and Wickens (2004) indicated three determinants through which power could be controlled, (a) significance level, (b) size of the treatment effect, and (c) sample size.

I conducted power analysis using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2009) to identify the sample size necessary, for the three-way independent analysis of variance (ANOVA) outlined for Phase VI of the analyses, to detect effect sizes across various levels of power as shown in Figure 2.

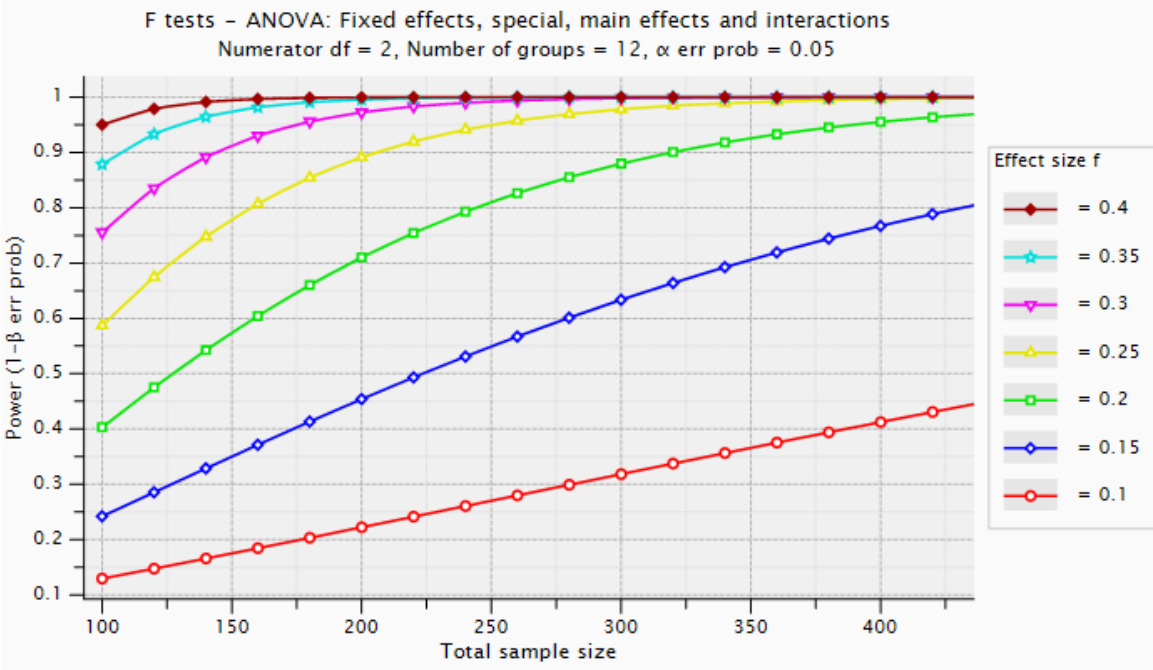


Figure 2. Sample size necessary to identify various effect sizes at various levels of power.

The ANOVA evaluated the independent variance of test condition with two levels, (a) received context of preference and (b) did not receive context of preference, grade level with three levels, (a) grade 6, (b) grade 7, and (c) grade 8, and gender with two levels, (a) male and (b) female, on the dependent variable of response time

engagement (RTE). For this study I chose an alpha of .05 and power of .80, which are typical values used for research in the social sciences (Field, 2018). The smallest effect size I could identify with the given sample ($N = 517$) was $f = .15$.

Procedures

School participant selection. Student selection was as a result of school membership in the *NWEA Partners in Innovation Program* established by NWEA in 2014. The *NWEA Partners in Innovation Program* was funded by the *Product Innovation Center*, a branch of the Research division at NWEA in Portland, Oregon. School partners had entered into a contractual agreement for which they received a stipend in advance of involvement in exchange for participation in a minimum of three research projects. Additionally, schools received supplementary monetary compensation based on individual project participation. Teachers, students and parents did not receive direct compensation by the funding entity for participation.

Assessment administration. Students, for whom parents did not opt out of participation, were provided with scripted information by a member of the project team. Students were told they would be taking a 24-item, fixed-form test online for which they would indicate their preferred context. Additionally, the project team member explained that once a context had been selected, the student would have a 50/50 chance of being administered an assessment with their chosen context. Students who did not get the context they chose were provided with the items in their original context and were used as the control group in the study. Students were told the reason for this was so that researchers could determine whether or not having a choice in context made a difference in how well students performed on the reading test and on how engaged they were

throughout. For students who selected, *context does not matter to me*, they had an equal chance of getting any of the four versions of the test (animals, fantasy, sports, or original context).

A URL and password were provided to students for access to the assessment. The initial login screen requested demographic information such as the name of their school, grade (6, 7, & 8), gender (female, male, non-binary third gender, & prefer not to say), ethnicity (African American, Asian/Pacific Islander, Hispanic, 2 or more races, Other, & White), and preferred context (animals, fantasy, sports, & context does not matter to me). Subsequent demographic and choice information, students were presented with an assent form in which it was stated that their participation was voluntary, and they could elect to opt out with without retribution. Students who opted out were presented with the *thank you for your participation* screen and exited the assessment. Students who provided assent began the assessment and ended either on their own or until completion. Students who ended/completed the assessment prior to the remainder of their classmates were provided with an alternative activity offered by their teacher.

Data Analyses

My study included 2 main research questions with multiple parts as discussed in the prior section of this paper. Each research question has been addressed sequentially in phases where *RQ1a* was addressed in Phase I, *RQ1b* in Phase II, through to *RQ2b* in Phase VI.

In Phases I through III a multi-step process that employed item response theory. IRT was used as the modeling process that described the relationship between the latent

traits (i.e., the constructs I wanted to measure) of student reading ability, the items identified in the measure, and the students' responses to each of the items in the measure.

In Phase I, I addressed RQ1a. RQ1a identified an item response model that best fit the subject matter data set, from among some common operational models including Rasch, 1PL, and 2PL, treating the item responses as a single dimension, by utilizing the R package Test Analysis Modules (TAM version 3.5.1; https://cran.r-project.org/web/packages/TAM/TAM.pdf) to model the data as a unidimensional model. TAM was used to fit a marginal maximum likelihood estimation of a set of unidimensional models using first the 1PL (Rasch) model, then the 2PL and 3PL models.

After an acceptable model was identified in Phase I, I modeled the data using a comparable multidimensional model addressing RQ1b in Phase II. For RQ1b, I evaluated the extent to which the three-dimensional model, based on the assignment of items to the three content goals (literature, informational text, and vocabulary use and acquisition), demonstrated better fit for the instrument. As in Phase I, the R package Test Analysis Modules (TAM) was utilized.

Once I identified the model that exhibited the best fit, I addressed RQ1c in Phase III. Guided by RQ1c, I assessed the extent to which the anchor items exhibited differential item functioning (DIF) for students in the Choice Condition as compared to Comparison Condition 1 (choice followed by random assignment to a context) and Comparison Condition 2 (students indicated no preference and were randomly assigned to a context). I conducted differential item functioning (DIF) analysis, using Winsteps (version 3.91.2).

Following the identification of both, the model that best fit the data and items that exhibited DIF, including removal of the biased items; through Phase IV, I addressed RQ1d. Based on RQ1d, I analyzed the extent to which distributions of proficiency estimates for the subject matter area between groups of students (*Choice* condition as compared to the *Comparison Condition 1*) varied. I used analyses of visual displays such as a Wright Map where both respondents and items can be displayed on the same scale aiding in the visual interpretation of student ability, using EAP estimates, relative to item difficulty for all three groups.

In Phase V, I addressed RQ2a. For RQ2a, I calculated the extent to which patterns of average item engagement varied across the three groups (Choice and the two comparison conditions) as a measured by Response Time Fidelity (RTF). I used analyses of visual displays such as a multiple line chart to identify varied patterns of average item engagement, as the test progressed from beginning to end, between groups tested. Additionally, I used frequency distributions to identify the extent to which the frequency of items at various levels of engagement varied between groups.

Finally, in Phase VI, I addressed RQ2b. In response to RQ2b, I determined whether or not RTE shown main or interaction effects by group, gender or grade, for this data set, using a factorial ANOVA. I also included a robustness test utilizing nonparametric tests to account for violations in one or more of the assumptions underlying factorial ANOVA.

CHAPTER III

RESULTS

As described in Chapter II, the results of my study were comprised of six interrelated phases employed to explore differences in performance of ability and engagement when an assessment offered students a choice in the context for which the content is presented. The *R* package Test Analysis Modules (TAM) (Robitzch, Kiefer, & Wu, 2018) was used to fit a marginal maximum likelihood estimation of a set of unidimensional item response models. TAM is an open-source estimation software in the *R* library that produces some output similar to ACER ConQuest. Using TAM, I modeled the data utilizing unidimensional models, first the 1PL (Rasch) model, then the 2PL model. I then modeled the data using a multidimensional Rasch model.

Subsequently I conducted DIF analysis using Winsteps version 3.91.2.0 (Linacre, 2015) to identify sources of possible bias within the 11 anchor items. Once biased items were removed and the best model fit was identified, I computed response time fluency (RTF) and visually inspected the results. I concluded analyses utilizing 3-way ANOVA to identify possible differences in engagement as measured by response time effort (RTE). The results of each phase have been discussed within this chapter.

Phase I

In Phase I of the analysis I carried out model estimation twice due to DIF found within Phase III. Two items were identified as severely biased and removed from the data. Analysis was recomputed and reported results are based on the second set of analyses (i.e., after items that exhibited DIF were removed). This means that the DIF question in a later phase employs a separate analysis for a full treatment of the subject,

but for the model fit in earlier phases it was important to address extreme DIF issues as shown here.

Model 1, unidimensional 1PL (Rasch) model. For Model 1, I began by calibrating the 11 anchor items in isolation. There are two approaches to anchoring items, simultaneous (concurrent) calibration and the fixed item parameter approach (de Ayala, 2009). In the simultaneous approach both samples' data are concatenated and calibrated in a single analysis. All of the items are on the same metric as defined by all combined individuals across samples. This holds true by definition when IRT estimates are used for person locations because person locations would be on the same metric regardless of the form of the test taken. The other approach is the fixed parameter where item parameter estimates from the first form are input in calibration of the second form and held fixed (anchored). If IRT estimates used to estimate person location, and forms meet the requirements of equating, results for persons are on the same metric and the individuals' estimates are linked.

I used simultaneous calibration to obtain the difficulty levels of the initial set of 11 anchor items. When using anchor items across tests the following criteria were met for this study, (a) together, anchor items should be measuring the same construct, content specifications, and contextual effects as the other, non-common items of the instrument, (b) they must have the same range of items location (difficulty) as the total test, (c) they accounted for no less than 20% of the instrument's length, and (d) in the context of IRT, model assumptions held tenable as discussed below.

The anchor items for the Context Personalization assessment were written to the same specifications as the remaining items in each form of the test and previously

calibrated difficulty levels were also in the same range (.9 – 1.9 logits). Additionally, the number of anchor items exceeded the minimum requirement of no less than 20% of the items in the instruments length (i.e., 11/24 or 45.8%). However, two items exhibited differential item functioning (i.e. items 1 and 3), to be discussed further in Phase III, and were removed from further analysis. The remaining nine items met all criteria.

I then modeled the item set and calibrated all remaining items based on the nine item anchor set. Figure 3 depicts the items for each form of the test, the form with the preferred context and the original context (i.e., not preference). Item numbers in red represent the anchor items and those in black are the context specific items.

Form	Item administration sequence																					
Preference	2p	4	5	6p	7	8p	9p	10	11	12p	13p	14	15p	16p	17p	18p	19	20p	21p	22	23	24p
Not preference	2np	4	5	6np	7	8np	9np	10	11	12np	13np	14	15np	16np	17np	18np	19	20np	21np	22	23	24np

Figure 3. Test item sequence by form after removing biased anchor items.

Using the 26-item subset (13 preference and 13 not preference) anchored to the initial 9 items from the Context Personalization assessment, 28 parameters were estimated, including 26 item difficulty parameters, no item slope parameters (all fixed to 1 initially to assess fit to the set of models described), one regression and one variance parameter. Constraint on persons was specified for this estimation.

Reliability. The IRT EAP reliability, which is one estimate of overall instrument reliability under the IRT model score, was .75. One criterion for minimally acceptable reliability in an assessment of individuals is .70 (Nunnally & Bernstein, 1994). Coefficient alpha from classical estimation was not estimated as the data represented a sparse matrix with approximately 37% of the data missing.

Item fit statistics. Item fit was excellent (infit range 0.87 – 1.22), with all of the estimated item difficulty and step parameters within a 3/4 - 4/3 mean square fit (Wu, Adams, & Wilson, 1998) and for parameters in which the weighted fit T was within ± 2.

Person estimation results. Figure 4 is the *Wright Map*, a graphical representation of the proficiency distribution on the latent trait. It shows the respondent reading scores. The left panel shows a representation of the latent reading proficiency distribution and the right panel indicates the difficulty of the items. Items on the Wright map are plotted at the point on the display where a student falling adjacent has a 50% chance of endorsing the item at that level.

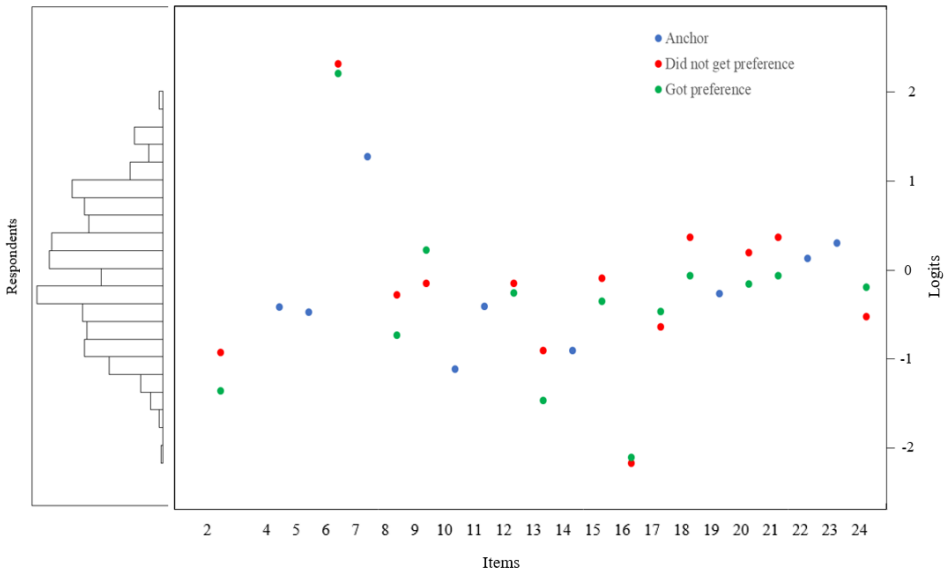


Figure 4. Wright map for Model 1: A unidmensional Rasch model of student performance and item difficulty for students who did and did not get their preferred choice of context.

A respondent's location in the proficiency distribution may be compared to the distribution of items in the instrument where lower values on the scale indicate students with a lower reading proficiency score and higher scale values indicate students with a higher reading proficiency score. The mean of the respondent proficiency on the latent trait, with a scale of approximately 1.81 to -2.09, is 0 (*SD* 0.42) with a constraint on

persons. The item thresholds on the Wright map graphically indicate reasonable distribution of items and scores relative to the distribution of persons for this data set.

Standard errors. The standard errors for respondents are shown in Figure 5. They exhibit a range from 0.41 to 0.51. Where students at either end of the scale (i.e., students with the highest and lowest reading achievement) tend to have higher standard errors, conversely students in the middle of the range tend to have lower standard errors, as is typical on many non-adaptive assessments but not extreme here. The standard errors average approximately .42 logits.

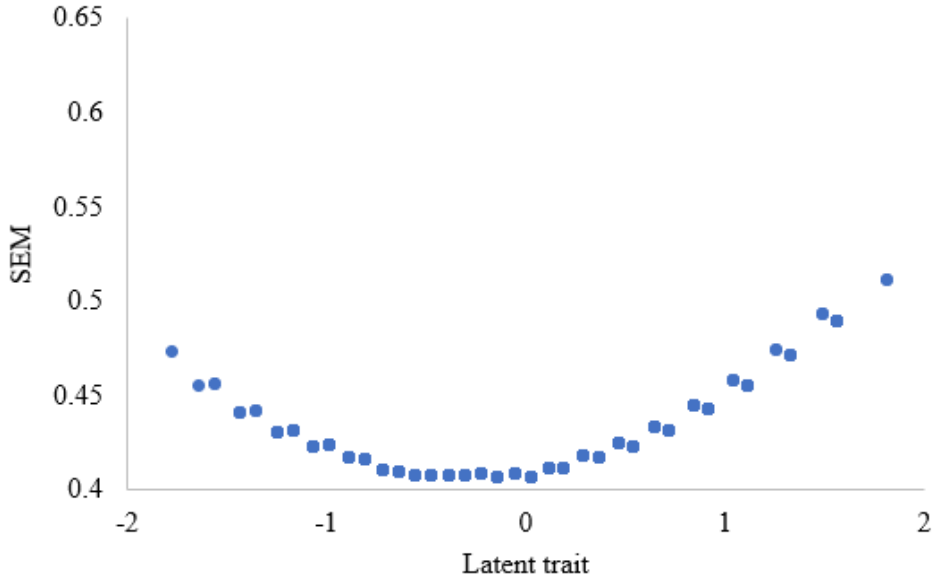


Figure 5. Standard error of measurement (SEM) for the assessment when modeled as a unidimensional Rasch model.

Model 2, unidimensional 2PL model. Next, a comparison of models took place. Model 2 (Unidimensional 2PL) was compared with the Model 1 (Rasch model), again using the same 26-item subset (13 preference and 13 not preference) anchored to the initial 9 items from the Context Personalization assessment. For this model, 62 parameters were estimated, including 26 item difficulty parameters, 35 slope parameters, and one regression parameter .

Reliability. The IRT EAP reliability, which is an estimate of overall instrument reliability under the IRT modeled score, was .77, which was approximately similar to the Rasch model. The general accepted criterion for good reliability is .70 (Nunnally & Bernstein, 1994). As with Model 1, coefficient alpha from classical estimation was not estimated as the data represented a sparse matrix with approximately 37% of the data missing.

Item fit statistics. Item fit again was excellent (infit range 0.97 – 1.07), with all of the estimated item difficulty parameters within the 3/4 - 4/3 mean square fit (Wu, Adams, & Wilson, 1998) and for parameters in which the weighted fit T was within ± 2.

Person estimation results. Figure 6 is the *Wright Map*, or graphical representation of the proficiency distribution on the latent trait. It shows the respondent reading scores. The left panel shows a representation of the latent reading proficiency distribution and the right panel indicates the difficulty of the items. Items on the Wright map are plotted at the point on the display where a student falling adjacent has a 50% chance of endorsing the item at that level.

A respondent's location in the proficiency distribution may be compared to the distribution of items in the instrument where lower values on the scale indicate students with a lower reading proficiency score and higher scale values indicate students with a higher reading proficiency score. The mean of the respondent proficiency on the latent trait, with a scale of approximately 2.15 to -2.45, was 0 (*SD* 0.47).

The item thresholds on the Wright map graphically indicated a somewhat skewed distribution of items and scores relative to the distribution of persons for this data set where item 6 for both forms of the test was nearly 12 logits more difficult than the

average student ability. Under the 2PL model the second parameter taken into account when modeling the data is the items ability to discriminate among individuals. The Wright map depicted in Figure 6 provides an indication that Item 6 failed to sufficiently discriminate between individuals across the continuum in comparison to the remaining items within the assessment.

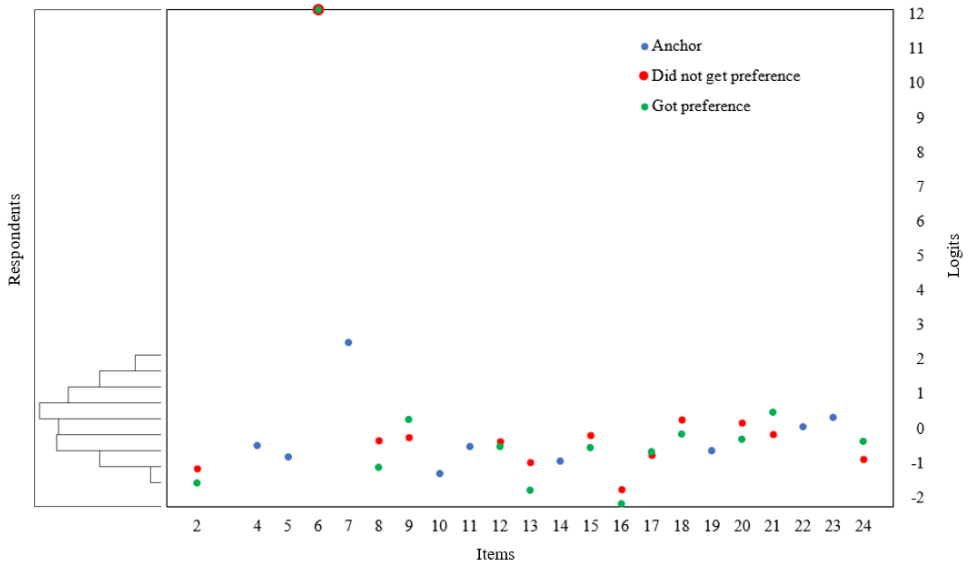


Figure 6. Wright map for Model 2: A unidmensional two parameter logistic (2PL) model of student performance and item difficulty for students who did and did not get their preferred choice of context.

Standard errors. The standard errors for respondents are shown in Figure 7. They exhibit a range from .41 to .65. Where students at either end of the scale (i.e., students with the highest and lowest reading achievement) tend to have higher standard errors, while students in the middle of the range tend to have lower standard errors. The standard errors average approximately .47 logits.

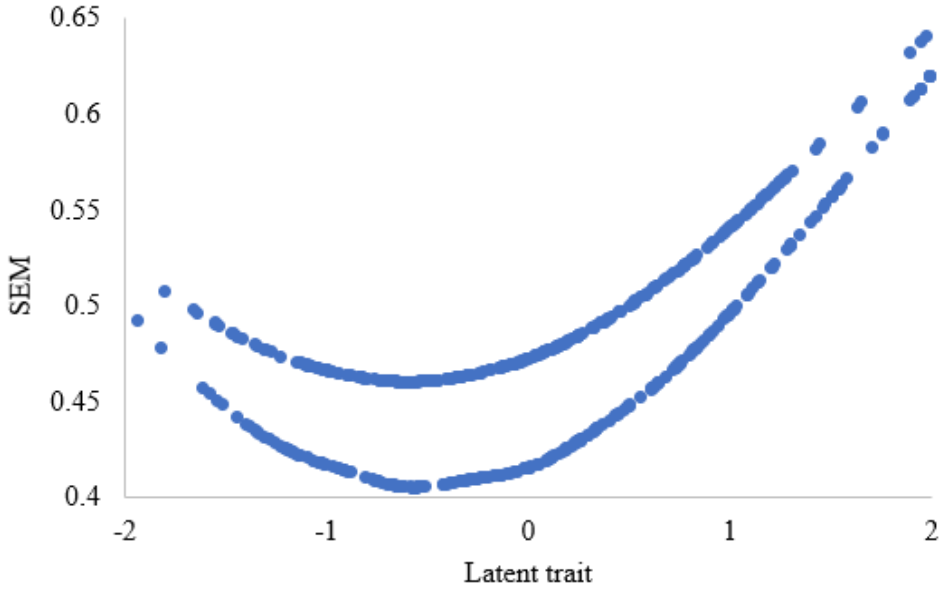


Figure 7. Standard error of measurement (SEM) for the assessment when modeled as a unidimensional two parameter logistic (2PL) model.

Table 32 shows a summary of four pieces of evidence that yielded mixed results. AIC is lower for Model 2; however, BIC is lower for Model 1. As Model 1 is a submodel of Model 2, the difference between the deviance of these two models is distributed as a chi-square with degrees of freedom equal to the difference in parameters estimated between the two models. The estimated deviance difference between the models of

Table 32

Comparison of Unidimensional Rasch and 2PL Models

Model 1 (Rasch)	**Model 2 (unidimensional 2PL)**
[a]Deviance: 13581.24	Deviance: 13434.58
[b]No. estimated parameters = 28	No. estimated parameters = 62
AIC = 13637 (constraint persons)	AIC = 13559 (constraint persons)
BIC = 13756	BIC = 13822

Note. Critical value for Chi-square distribution with df 34 and alpha= .05 is 48.60.
[a]Difference in the deviance: 13581.24 - 13434.58 = 146.66.
[b]Difference in the parameters: 62 - 28 = 34.

146.66 is significant (p = .05). Additionally, the reliability of Model 2 (.77) was only slightly better than that for Model 1 (.75). Inspection of each the Wright Maps and the SEM plots along with BIC results suggest Model 1 as the more parsimonious model is reasonable to use for the purposes of this choice-comparison study. However, a 3-dimensional model warranted examination due to theory considerations so prior to making a final determination, the next phase of my study was conducted.

Phase II

In Phase II of the analysis I examined a three-dimensional 1PL model to compare with the unidimensional 1PL model. For this model 41 parameters were estimated, including 35 item parameters and 6 (co)Variance parameters. Comparisons between Model 1 and Model 3 are depicted in Table 33.

Table 33

Comparison of Unidimensional Rasch and Three-dimensional Rasch Models

Model	*N*	Estimated parameters	Deviance	EAP	AIC	BIC
Unidimensional Rasch	517	28	13581.2	0.749	13637	13756
Three-dimensional Rasch	517	41	13552.0		13634	13808
Dim. 1 - vocabulary acquisition & use				0.725		
Dim. 2 - informational text				0.735		
Dim 3 - Literature				0.674		

As expected, due to the smaller item sets per dimension, reliability was lower for all three dimensions in comparison to model 1 (i.e., information text = .73, literature = .74, and vocabulary acquisition and use = .67). Additionally, AIC was essentially the

same for model 3 (13634) in comparison to model 1 (13637), while BIC was substantially higher for model 3 (13808) in comparison to model 1(13756). Because the reliability for each dimension was lower than for Model 1 and Model 3 was less parsimonious than Model 1 and not supported by BIC, Model 3 was not employed, and Model 1 served for the remaining phases of this study.

Phase III

In Phase III of the analysis I examined the anchor items for bias between students who were assessed using the form of the test that matched their preferred context in comparison to students who were not assessed with the form of the test that matched their preferred context (i.e., original form of the items).

DIF analysis. Winsteps (version 3.91.2) was invoked to carry out the DIF analysis for anchor items used in the Context Personalization test prototype included all completed test events that were administered during the fall of 2018. Each assessment record included the student's context preference and assigned assessment group membership (i.e., students administered preferred context of their choice, students not administered preferred context of their choice, and students for whom choice did not matter). The reference group, *Choice Condition*, is represented by students who were assessed with the preferred context of their choice and the two focal groups, who shall be referred to as *Comparison Condition 1* and *Comparison Condition 2*, are represented by students who did not get their preferred choice and students who indicated choice did not matter to them respectively.

To help in summarizing results, the Educational Testing Service (ETS) delta method of categorizing DIF (Holland & Thayer, 1985) was incorporated. The delta

method allows items exhibiting negligible DIF (difference < .43 logits) to be differentiated from those exhibiting moderate DIF (difference ≥ .43 and < .64 logits) from those exhibiting severe DIF (difference ≥ .63 logits).

As depicted in Figure 8, three of the 11 anchor items exhibited statistically significant DIF between the reference group and at least one of the two focal groups.

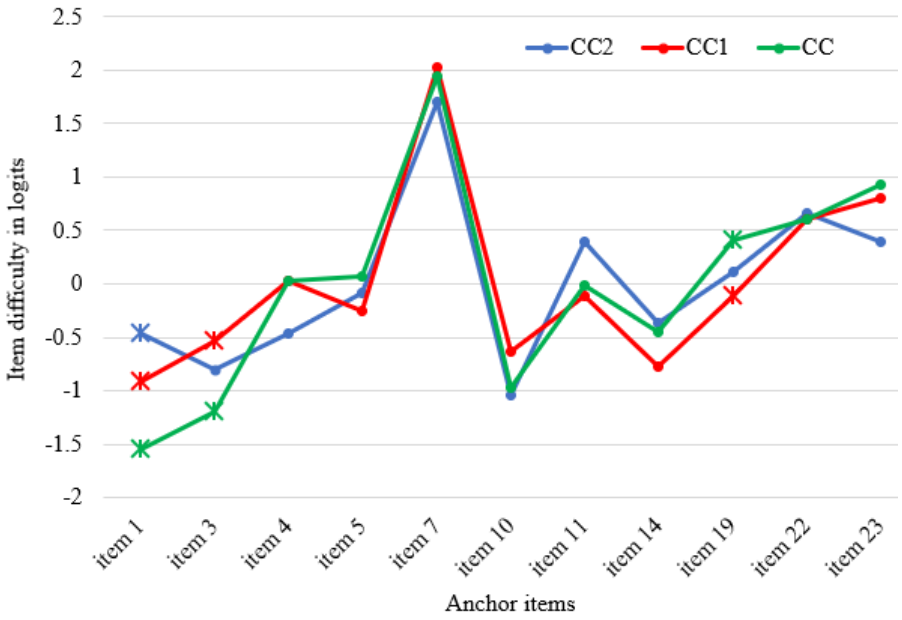


Figure 8. Differential item functioning (DIF) between groups.

The green line represents the *Choice Condition* (reference group), with *Comparison Condition 1* and *Comparison Condition 2* in red and blue respectively and representing the focal groups. Items that indicated statistically significant differences from the reference group are marked with asterisks. Item 1 exhibited severe DIF (i.e., DIF >= .64 logits) between the reference group and both focal groups. It was .64 logits easier for the reference group as compared to *Comparison Condition 1* (p = .025) and 1.09 logits easier when compared to *Comparison Condition 2* (p = .013). Item 3 also exhibited severe DIF, however, only between the reference group and Comparison Condition 1. It was .66 logits easier for the reference group (p = .008). Item 19 exhibited only moderate DIF (>=

.43 logits and < .64 logits). Item 19 favored Comparison Condition 1 indicating the item was .52 logits more difficult for the reference group (p = .025). In the *Discussion* section I will explore plausible explanations for the items that exhibited severe DIF.

The remaining items did not exhibit significant DIF and ranged in magnitude from negligible at .00 to moderate at .53. The items exhibiting statistically significant and severe DIF (i.e., items 1 & 3) were removed from the data set and subsequent analyses. Prior analyses were recomputed with the nine item anchor set and reported as indicated previously in Phases I and II.

Phase IV

In Phase IV, I conducted visual analyses to examine the extent to which distributions of proficiency estimates for the subject matter area between groups of students varied. As illustrated in Figure 9, the distributions of student ability showed considerable similarities across groups tested; although some differences were seen, they did not represent a common trend. A larger data and pretest, or other approach to show equivalency of groups, would be needed to explore this question beyond visual analysis.

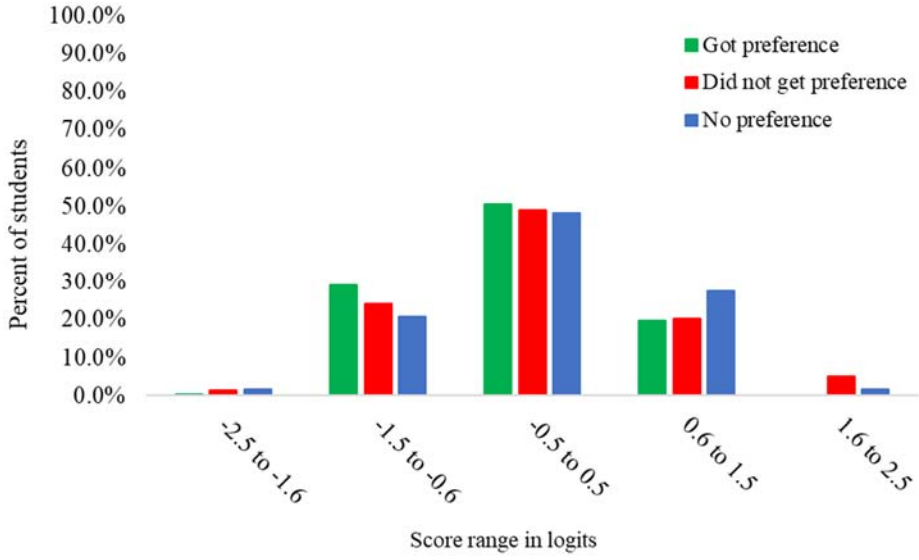


Figure 9. Proficiency estimates by preference group.

Phase V

In Phase V of the analyses I calculated the extent to which patterns of average item engagement varied across the three groups (got preference, did not get preference, and no preference) as a measured by Response Time Fidelity (RTF). As exhibited in Figure 10, students who were assessed with their context of interest showed greater RTF in 68.2% (15/22) of the items as compared to students who were not assessed in their

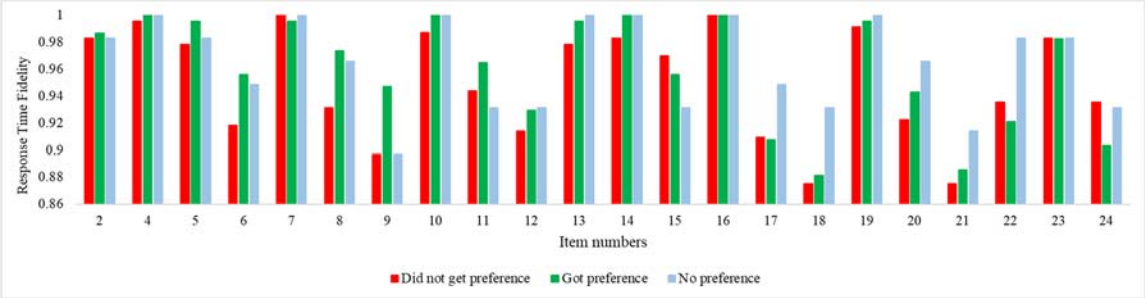


Figure 10. RTF for each item by preference group.

context of interest. Response time fidelity did not differ between preference groups in 9.1% (2/22) of the items. Conversely, students who were assessed with their context of interest only showed greater RTF in 31.8% (7/22) and no difference in 22% (5/22) of the items as compared to students who indicated no preference. Further exploration of the composition of tests (i.e., original context and animal, fantasy, or sports contexts) administered to students in the no preference group are provided in the *Discussion* section.

Response time fidelity exhibited a pattern by content goal area (vocabulary acquisition and use, informational text, and literature) across all groups in general. The pattern begins at item 4, as items 1 and 3 were removed due to DIF, where vocabulary acquisition and use exhibits the highest RTF followed by informational text, then literature. Overall, RTF was high with a range between .875 to 1 across all groups of

students. This means that even for the items with the lowest RTF, 87.5% of students met the minimum criteria for solution based behavior.

Phase VI

Finally, in Phase VI, I examined the extent to which response time effort (RTE), a measure of overall effort on individual test events, varied based on choice of test context and gender, as students progressed through middle school. Choice of testing context was the first independent variable with two levels (assessment matched student preference and assessment did not match student preference). Gender was the second independent variable also with two levels (male & female). The final independent variable was grade that consisted of three levels (6^{th}, 7^{th}, & 8^{th}).

I began analysis by testing the assumptions underlying a 3-way ANOVA. The assumptions included, (a) interval data of the dependent variable, (b) data are approximately normally distributed, (c) homogeneity of variance, and (d) no multicollinearity (Howell, 2013). The data did not conform to the assumptions of normality as depicted in Figure 11, nor homogeneity of variance as concluded using Levene's test $F(11, 424) = 1.99$, $p = .03$.

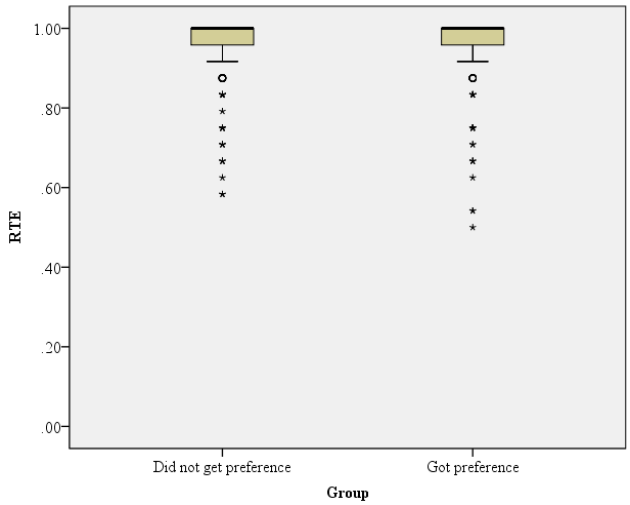


Figure 11. Distribution of RTE by preference group.

Due to the characteristically severely negatively skewed nature of the data (i.e., the expectation is that the majority of students engage most of the time), transformation of the RTE variable failed to yield a normal distribution. However, ANOVA is fairly robust to both of these violations. Therefore, I proceeded with a factorial ANOVA and then applied non-parametric tests as a robustness check. Table 34 provides a report of the descriptive statistics.

Table 34

Descriptive Statistics for Response Time Engagement by Context, Gender, and Grade

		Did not get preferred context			Got preference of context		
Variables		*M*	*SD*	*N*	*M*	*SD*	*N*
Grade 6							
	Female	.97	.08	39	.97	.07	41
	Male	.98	.07	32	.94	.10	32
Grade 7							
	Female	.97	.07	39	.97	.10	50
	Male	.93	.11	51	.96	.07	42
Grade 8							
	Female	.96	.08	32	.96	.10	29
	Male	.96	.09	23	.96	.09	26
Total		.96	.09	216	.96	.09	220

The analysis of variance results are reported in Table 35. The second order effects were not statistically significant, $F(2, 424) = 1.59$, $p = .204$, partial eta squared = .007. Additionally, first order interaction effects were not significant, nor were main effects.

Subsequently, I conducted nonparametric tests as a robustness check. However, there were no nonparametric measures comparable to a three-way ANOVA. This limited the use of nonparametric measures to those comparable to *t*-tests and one-way ANOVAs. Therefore, it was necessary to conduct multiple tests. First, I conducted separate Mann Whitney tests (comparable to an independent sample *t*-test), one for the main effects of

Table 35
Three-Way Between-Subjects Analysis of Variance Summary Table for the Effects of Context, Gender, and Grade on Response Time Engagement

Source	*df*	*SS*	*MS*	*F*	*p*	Partial eta squared
Group	1	1.416E-6	1.416E-6	0.00	.989	.000
Gender	1	0.018	0.018	2.39	.123	.006
Grade	2	0.002	0.001	0.16	.856	.001
Group * Gender	1	0.002	0.002	0.23	.629	.001
Group * Grade	2	0.025	0.012	1.61	.202	.008
Gender * Grade	2	0.008	0.004	0.55	.580	.003
Group * Gender * Grade	2	0.024	0.012	1.59	.204	.007
Error	424	3.248	0.008			
Total	436	405.361				

preference group (did not get/got context preference) and the other for the main effect of gender (female/male). Then I conducted a Kruskal Wallis (comparable to a one-way between subjects ANOVA) test for grade level (6, 7, & 8).

Similar to the three-way ANOVA, using nonparametric measures, I found no significant main effects of either group nor grade. However, I did find significant differences based on gender (p = .018). I employed Mann Whitney for post hoc tests and identified significant differences by gender for 7th grade students who were in the testing group who did not get their preference for context (p = .008), see Figure 12.

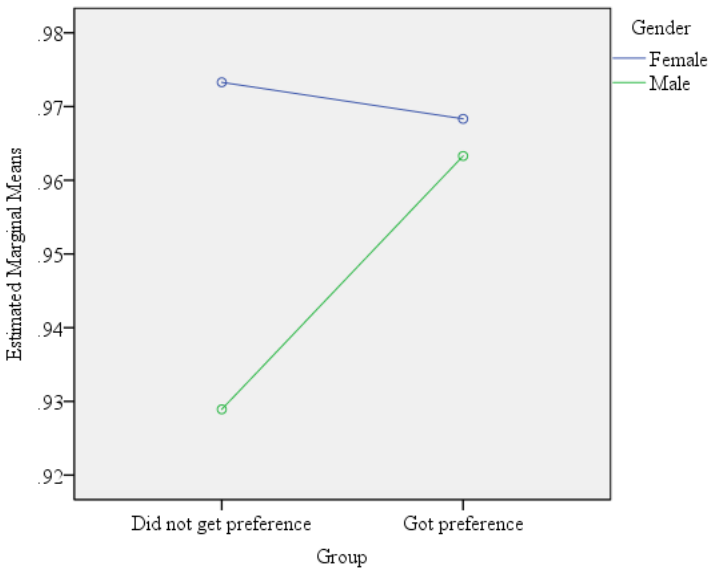


Figure 12. Mean response time effort (RTE) for male and female 7th grade students by preference group.

However, these results may be confounded due to family-wise error. Running the comparable parametric tests (i.e., *t*-tests and one-way between subjects ANOVA) I yielded the same results as the nonparametric tests, confirming that conflicting results were due to family-wise errors. Therefore, the results of the factorial ANOVA were confirmed, and the nonparametric results rejected, for this phase of the investigation.

CHAPTER IV

DISCUSSION

This final chapter was divided into six sections. In the first section I have provided a summary of the results based on the research questions and hypotheses. Within the second section I have discussed the contributions this work has made to the body of knowledge. In the third and fourth sections I have discussed the limitations of the study and threats to validity respectively. In closing, I have provided recommendations and implications for future research as well as concluding remarks.

Summary of Research Question Findings

The purpose of this study was to examine the impact of affording students choice over the context within a reading assessment by introducing them to an interest-based choice opportunity. In so doing, the intent was to examine possible improvements in the measurement of student ability and increased motivation when students engaged with an assessment that addressed their self-reported interests. Ultimately the hope was to provide a model of assessment that better measured students ability and sufficiently engaged them in the test so as to provide a representative account of their ability.

The research questions in this study were comprised of two focal areas. The first was the estimation of subject matter assessment examined in four phases (i.e., Phases I through IV). These four phases examined the structure and bias of the interest-based reading assessment across groups of middle school students (i.e., grades 6-8). The second focal area was the estimation of engagement examined in two phases (i.e., Phases V and VI). In Phase V, I examined response time fidelity (RTF) (Wise, 2006) as a measure of

engagement at the item level and in Phase VI, I examined response time engagement (RTE) (Wise & Kong, 2005) as a measure of engagement at the test level.

In Phases I and II of this research, I identified the unidimensional Rasch model as the item response model that exhibited best characteristics for this study of the *Context Personalized* reading assessment. This was consistent with the model for which the original items were based (NWEA, 2011). Additionally, the mean SEM for this 23-item fixed-form test (.42 logits) indicated that it was well targeted to the sample used. The original 45-item computer adaptive tests on which the prototyped test was based reported a mean SEM that was more precise (.35 logits), as would be expected in a computer adaptive test (CAT) of longer length, however precision was degraded by less than one tenth of a logit. This is an indication that a version of the prototyped test of similar length would likely be comparable in precision as well.

In Phase III of this study differential item functioning (DIF) was indicated for two of the items. Due to the severity and statistical significance of the DIF it was necessary to remove the two items, recompute calibrations and rerun analyses in Phases I and II, as reported in the *Results* section. As stated in the *Results* section, Item 1 exhibited severe DIF favoring the reference group (*Choice Condition*) over both focal groups (*Comparison Condition 1 & Comparison Condition 2*) by .64 logits (p = .025) and 1.09 logits (p = .013) respectively. Item 1 was a vocabulary question for which each version of the item (i.e., original, animals, fantasy, & sports) was written to the 9th grade level and assessed the use of context clues to determine the meaning of the word. Due to the word count for item 1 (i.e., <50 words) additional readability measures cannot be reasonably applied.

Item 3 also indicated severe DIF favoring the reference group (*Choice Condition*) by .66 logits ($p = .008$) relative to *Comparison Group 1* (students who were not assessed with their preferred context). Item three was a literature question for which each version of the item (i.e., original, animals, fantasy, & sports) was written to the 7th grade level and assessed the interpretation of symbolism in literary text.

As previously indicated in the *Methods* section, all versions of the items were written to the same grade level based on grade specific standards and adhering to grade level vocabulary, same depth of knowledge, and within ± .2 of the original item's readability as computed by Flesch-Kincaid's readability formula. However readability does not take into consideration the comprehensibility of an item. Readability, as measured by Flesch-Kincaid, considers average sentence length and average syllables per word and is a characteristic of the text itself. However, comprehensibility also considers the context in which the words are used and the relationship between the reader the reader's knowledge of the content read.

Through the use of Coh-Metrix (McNamara & Graesser, n.d.), a text analysis tool that considers comprehensibility in addition to readability, evidence of differences that could account for bias have been identified. These characteristics include narrativity (N), syntactic simplicity (SS), word concreteness (WC), referential cohesion (RC), and deep cohesion (DC) as identified Tables 36 through 39 provide the additional comprehensibility and readability information for each version of the item (i.e., original, animal, fantasy, & sports). Additional comprehensibility and readability information for remaining items are located in *Appendix C*.

Table 36

Additional Readability Information for the Original Context of Item 3

Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has low word concreteness, which means there are many abstract words that are hard to visualize. Abstract texts may be more difficult to understand. It is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	75%	87%	21%	12%	2%

Note. FK = 2.7 and word count = 103.

Table 37

Additional Readability Information for the Animal Context of Item 3

Readability narrative	N	SS	WC	RC	DC
This text is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. It is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	69%	82%	33%	9%	2%

Note. FK = 2.8 and word count = 111.

Table 38

Additional Readability Information for the Fantasy Context of Item 3

Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has low word concreteness, which means there are many abstract words that are hard to visualize. Abstract texts may be more difficult to understand. It is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	78%	88%	12%	9%	2%

Note. FK = 2.8 and word count = 104.

Table 39

Additional Readability Information for the Sports Context of Item 3

Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	81%	82%	37%	12%	2%

Note. FK = 2.8 and word count = 98.

The animal version of the item, previously depicted in Table 37, varied by both narrativity and word concreteness ranking as well as word count. Similarly, the sports version of the item, previously depicted in Table 39 also varied in word concreteness as

compared to the original version of the item. The lower narrativity ranking for the animal version of the item could increase its difficulty. However, only 22% (i.e., 49/227) students in the reference group saw this version. Additionally, the item favored rather than biased the reference group. Therefore, it is less likely that narrativity would account for the bias. Conversely, the higher word concreteness ranking in both the animal and sports versions of the item could make the item easier for the reference group. As a higher word concreteness ranking indicates that there were fewer abstract words. The combined percentage of students who saw either the animal or sports version of the item (61%) represents a plausible explanation of bias favoring the reference group for item 3.

Phase IV of the analyses showed similar achievement performance across the three groups assessed as indicated previously in Figure 9. Similarly, differences in the precision across groups were negligible as depicted in Figure 14. Precision in scores was best measured for students who scored at and just below the mean (0 to -0.7 logits).

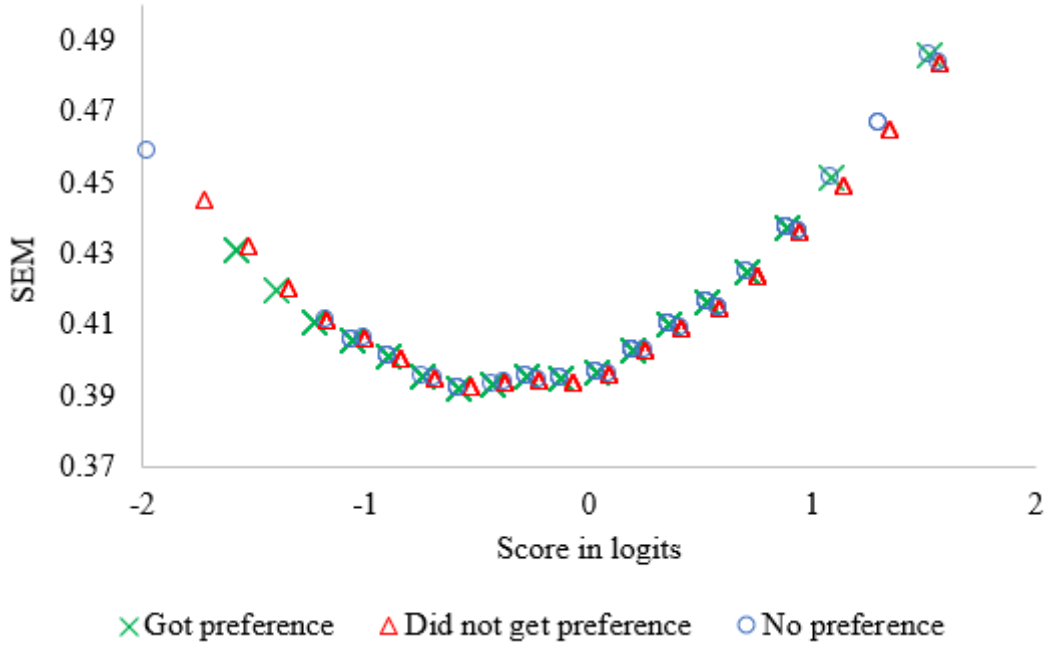


Figure 13. Precision of Context Personalization reading scores by assessed group.

Through examination of response time fidelity (RTF), in Phase V of the analyses, students across all groups (got preferred context, did not get preferred context, and no preference in context) showed relatively high engagement across all items in general (i.e., between .875 and 1) as identified through solution behavior. An important distinction to call to the fore is that RTF is a measure of the percentage of students who met or exceeded the minimum time threshold (i.e., 10% of the average time spent on a particular item by a group of students) to be identified as providing solution behavior in response to an item as opposed to the average length of time spent on an item per group.

A pattern emerged relative to item engagement by goal area, whereby items with the lightest reading load (i.e., vocabulary use and acquisition) where those items for which students most often exhibited solution behavior. In contrast, items with the heaviest reading load (i.e. literature) where those for which students exhibited solution behavior less often. There were two observed exceptions to this pattern in the data. Students showed greater engagement for items 14 and 23, both assessed aspects of informational text, as compared to the prior items (i.e., 13 and 22) that each assessed vocabulary acquisition and use.

Upon closer inspection of each of these behavior anomalies, the supplemental passage for item 14 contained fewer total words (i.e., 8) in comparison to the passage for item 13 which contained 78 words. Therefore, the reading load was greatly reduced for item 14 as compared to item 13 which indicated student solution behavior followed the same pattern previously observed. The anomaly in solution behavior between items 22 and 23 cannot be explained based on passage length as both were of similar length. They

varied on additional readability attributes as well. Subsequent research beyond the scope of this paper would be necessary to provide plausible explanations.

In the final phase of this study a 3-way ANOVA failed to yield significant main nor interaction effects of choice of test context in the engagement of students, as measured by RTE, when assessed in this reading assessment using this data set. The results from subsequent nonparametric tests were rejected because they could not account for family wise errors, but they brought to light a possible extension of the current research within a more narrowed focus on reducing gaps between boys and girls when engaging in reading assessments.

Recent research conducted by Wise and Jensen (2019) reported that 14.3% of reading assessments administered in the fall to middle school students exhibited rapid guessing behavior (i.e., non- solution-based behavior for 10% or more of the items in an assessment). My study yielded similar results. Again, though significant differences across groups were not identified, my study's results were consistent with similar research in this finding (Wise & Jensen, 2019). Students in the no preference group exhibited the lowest percentage of students with rapid guessing behavior (i.e., 10%) as compared to both the students who did (13%) and did not get the context preference of their choice (16%), although this study design was not designed to interpret statistical significance of this small difference. Worth noting was that two thirds of the students in the no preference group were assessed with a test from among three interest based contexts (the same as the *preferred context group*) while the remaining one third of them received the original context (the same context as the *did not get preferred context group*).

Contributions to the Body of Knowledge

Throughout the literature student interest has been indicated as a motivating factor shown to impact student learning, engagement and literacy (Anand & Ross, 1987; Bernacki & Walkington 2014; Brozo et al, 2014). Additionally, my review suggested a need to examine the relationship between student interest and the nature of the assessment models used in measuring achievement as well as test engagement. My research examined one aspect in a single reading data set of assessment models using areas of interest to personalize context in general to groups of students.

Prior research had examined interest-based learning in mathematics with mixed results (Anand & Ross, 1987; Bernacki & Walkington, 2014; Brozo et al., 2014). Anand and Ross (1987) found significant differences[4] in performance ($p < .01$) and attitude ($p < .05$) and described findings regarding transference of lessons to subsequent learning as *marginally* significant ($p < .10$), on math lessons favoring students in the personalized condition. However, they (Anand & Ross, 1987) did not find significant differences based on gender nor learning time under the same conditions.

Bernacki and Walkington (2014) yielded similarly mixed results when they examined the impact of personalization using out of school interests within algebra lessons. They found that personalization differentially impacted individual interest, as reported by a self-reported measure of student interest, in algebra based on students level of interest in mathematics overall, where students with lower levels of interest in mathematics reported higher triggered interest in mathematics as compared to students with previously reported high interest in mathematics. However, Bernacki and

[4] The study authors cited did not provide effect sizes for significant findings. As such *p*-values were reported in their stead.

Walkington found no significant difference in maintained situational interest relative to perceived value or enjoyment.

My research sought to identify differences in achievement as well as test engagement, using a computer-tracked measure of engagement, between groups of students afforded the opportunity to participate in an interest-based reading assessment. Dissimilar to mixed results from the prior research, my research found no significant differences in neither achievement nor engagement although in some phases only visual analysis was employed so further investigations could explore this topic more completely. This finding seems to suggest several hypotheses, although note limitations in the upcoming sections: (i) design of the study without pretest and exploration of equivalent groups following randomization may be insufficient to explore the topics, (ii) randomization via the process employed may not have been sufficiently complete and might need designed stratification for instance, (iii) interest-based engagement may not operate in such a way given the particular data set, instrument and approach used here to significantly alter inferences about respondents, and/or (iv) a time-based measure of engagement may not be an appropriate proxy for interest-based engagement.

Brozo et al., (2014) reported significant and substantial gender differences, favoring girls, in overall print reading literacy as measured by an optional test of digital literacy within PISA. In the same report they reported significantly higher indices of reading for enjoyment, also favoring girls. As previously indicated in Chapter I, the subject area assessed within my study was reading. Although significant differences by gender were not identified in my research, the potential of family-wise error led to

rejection of non-parametric results so this information provided may be used to guide subsequent research in the sphere of personalized assessment.

Limitations

There were limitations that bear mention within this study. They included limitations relative to the sample, measure of engagement, components of technology used, and testing environment. Each are discussed in turn.

Sample limitations. The sample used in this study represented a sample of convenience followed by some degree of random assignment rather than a random sample from among the population of students in the United States. Students in this sample were purposively drawn from a group of schools who were compensated for participation through a contractual agreement with a research organization. Although students were not directly compensated for their participation it was not possible to control for any unforeseen mitigating effects of such an agreement, for the sample.

Random assignment of students to testing context was based on test login information provided by the student. Students were given the option of selecting the context for the reading assessment from among four options (*animals, fantasy, sports, and context doesn't matter to me*). Prior to testing students were informed they would be asked about the context they preferred for the reading assessment and that they would have a 50/50 chance of getting their preferred context, unless they selected *context doesn't matter to me*. In that case students were informed they would be randomly assigned to one of the four testing contexts. It was explained that context referred to the reading within each item which would be based on a topic of interest.

Additional attributes such as school type, socioeconomic status (SES) and prior achievement were not controlled for in this study. Although the sample composition was somewhat heterogenous both geographically and ethnically, it was more homogenous based on school type and SES, and prior achievement levels were unknown. Four of the five schools represented students of greater affluence as indicated by low percentages of students who qualified for free or reduced lunch programs or where funding for such a program was unnecessary (i.e., private schools) as depicted in Table 40. Thus, 77% of students in this sample were comprised of students with greater affluence as compared to the national average of 21% (National Center for Education Statistics, 2016). Therefore, it is plausible that exclusion of such attributes may have played a role in the results obtained.

Table *40*
School Demographics by Type and Free and Reduced Lunch Program (FRL)

School and type		*N*	Percent of sample	Percent of FRL
Private		224	43%	
	School A	66	13%	0%
	School B	158	30%	0%
Charter		102	20%	
	School C	102	20%	12%
Public		191	37%	
	School D	121	23%	50%
	School E	70	14%	6%

Measure of engagement limitations. The use of time to measure interest and motivation was attractive in that time could be measured precisely and without bias

through the use of technology. However, using response time fluency and response time engagement as identified through solution based behavior may have been insufficient as a proxy for student interest and motivation on its own. Although self-reported measures may be more susceptible to bias, a time-based measure cannot provide information explaining the time students spend with an item nor across a test. The inclusion of both computer-tracked and self-reported measures may provide a more complete picture of actual interest and motivation, over multiple measures.

Technology limitations. Differences in technology used across schools may have limited the extent to which computer-tracked timing components were measured. The use of trackpads versus a mouse for response selection, scrolling item text, initiating screen magnification, and item advancement could have differentially impacted student time spent on an item and/or across a test event based on students familiarity with the technology used and the key-stroke short cuts available to them in comparison to menu driven or browser-based tools. Additionally, the use of Chromebooks, versus laptops, versus desktop computers could also impact the rate at which students were able to interact with items and the test overall. These differences could have confounded, or introduced construct irrelevant variance, into the individual results for RTE and RTF.

Testing environment limitations. Finally, the setting in which students were administered the assessments varied both across and within schools. Four of the five schools assessed students either in the classroom or within a computer lab. However, one of the schools utilized both settings. Additionally, seating configuration varied across schools with some students seated immediately adjacent one another in rows, within grouped desks facing one another, and in a row and column configuration. The row and

column configuration elicited the least amount of interactions between students, followed by students seated immediately adjacent one another, and with the most interactions observed by students within the grouped desks that faced one another. Additional interactions between students could have impacted both time on task and test scores.

Threats to Validity

In this study I have acknowledged the limitations as identified in the previous section. Additionally, each of the aforementioned limitations are considered as possible threats to the validity of the study. In this section I have discussed specific threats to both internal and external validity.

Internal validity. Threats to internal validity include selection, history and possibly instrumentation. Because sample selection was non-randomized and random assignment not stratified, selection could have been a threat to validity. The absence of prior information regarding student ability curtailed the assumption of equivalent groups, as well as the absence of a pretest measure of some other criterion reference.

Additionally, some of the students who used Chromebooks for the assessment were unaware of the keystroke sequence necessary to quickly zoom in and out on the item text and initially resorted to using the browser-based tool to obtain the same functionality. Use of the browser-based functionality could have artificially inflated item response times and or lead to frustration resulting in the opposite effect on time. Possible threats to validity due to instrumentation could have occurred in two ways. The first was relative to proctor knowledge of the keystroke sequence to instantiate the zoom feature in Chromebooks. The other could occur if proctors did not consistently provide students with the scripted instructions for participation.

External validity. A singular possible threat to external validity was identified for this study. Reactive effects of experimental arrangements could have posed a threat to the external validity. The Hawthorne effect could have come into play because students were aware that they were participating in an experimental study.

Recommendations and Implications

The scope of this study was limited to fitting interest-based assessment models of reading and exploring assessment engagement of middle school students based on the use of interest-based tests. Future research may branch out to include additional measures of engagement, exploring differential item functioning across all items within an interest-based assessment, examining student strategies relative to their choice of context, assessing different subject areas (e.g., math or science), explore possible effects at varied grade levels (e.g., elementary or high school students) or study specific student populations (e.g., students for whom English is their second language, students with disabilities, student populations that vary by SES).

My study examined standard IRT models of achievement (unidimensional1PL/Rasch, 2PL, and multidimensional 1PL) to identify which, if any, best fit the data from this sample. Although the unidimensional Rasch model was identified as the model with the most reasonable overall fit given the purposes of the study, and was consistent with the model under which the original form of the items were calibrated, the absence of students' prior achievement levels in reading constrained analyses to that of model fit, differential item functioning of the anchor items across groups, and visual inspection of performance on the measure. Furthermore, item characteristic curves were not compared between predicted and observed data. Finally,

visual inspection of achievement by group is a small part of this study and indicated no overall trends seen, but such visual inspection has limited utility without a more robust statistical framework such as being able to assume equivalent groups following random assignment and using proficiency estimates such as EAP to compare for significance should trends be seen. Future research should, at minimum, include prior achievement level so as to provide the opportunity to more clearly establish the relationship of interest-based assessment and achievement estimation, as well as whether the relationship was substantially dependent on location on the latent trait or other aspect of the student distribution.

As previously suggested, a time-based measure of engagement may be an overly coarse representation of student engagement when used in isolation. This might be more prevalent for items with few words because the normative threshold (i.e., 10% of the average student response time) may be so low that all students would likely meet the criterion for solution based behavior. Solution based behavior in conjunction with other measures of engagement might be necessary to fully capture student engagement. Subsequent studies might consider additional measures such as tracking student movement or interaction within assessment tasks based on screen clicks. Tracking of eye movement, as in pupillometry, where expansion and constriction of the pupils may be monitored (Ahern & Beatty, 1979) as students engage with assessment tasks might also contribute information on engagement.

Expanding the research to include examination of differential item functioning across all items within the test, as opposed to constraining it to the anchor items alone, represents another opportunity for future research. Although the use of DIF in my study

was to ensure anchor items were equivalent in difficulty for all groups of students (i.e., got preference, did not get preference, and no preference), retaining them might have contributed evidence of import to student performance. Subsequent studies might consider moving anchor items that exhibited statistically significant and severe DIF from the anchor set and calibrating them with the remaining items.

Findings from my study relative to test engagement failed to show significant differences between groups of middle school students who were assessed using interest-based test items in comparison to those students who did not received interest-based items within a reading assessment. Although significant differences were not detected between students in this sample, 89% of students indicated a preference in context. Subsequent studies might consider examining the strategies that students report using when selecting the context of interest when given a choice. Choice may have had little to do with interest and more to do with which context students believed might be easier. For students who enjoyed reading it might have been that one context was just as good as the next. Still for others it might have been that irrespective of context they felt confident in their ability regardless of context, or for very low achievers, it might have been that they felt they would do poorly. Another avenue for exploring interest based choice might include administering clustered sets of items based on students choice interspersed by clusters of items that do not correspond to the students choice.

Much of the prior research that examined student interest and/or personalization focused on learning mathematics (Anand & Ross, 1987; Bernacki & Walkington, 2014; Bernacki & Walkington, 2018; Cordova & Lepper, 1996; Walkington, 2013; Walkington & Leigh, 2015; Walkington, Petrosino, & Sherman, 2013). Far fewer focused on reading

(Brozo et al., 2014; Ivey & Broaddus, 2001). My study sought to extend the literature by examining student interest and/or personalization focused on engagement within a reading assessment. Future research examining engagement in mathematics assessment may be able to close the loop between interest-based learning and interest-based assessment in the literature (e.g., do interest-based tests yield more favorable achievement levels initially, but fail to sustain interest; do interest based tests differentially impact low vs high achieving students). Furthermore, examining additional subject areas with interest-based assessments or matching students' interest-based learning with interest-based assessment may provide insight into the ways in which interest differentially impacts education.

Previous studies examined the impact of interest using a sample comprised of high school students (Bernacki & Walkington, 2014) while other studies used elementary school students in grades 4 and 5 (Cordova & Lepper, 1996). My study was comprised of middle school students (grades 6 – 8). Although grade level was factored in to the analysis for this study, the introduction to a wider grade range may yield different results. Future research may examine transitional grade ranges between elementary and middle school and then again from middle school to high school.

Another avenue of research to be explored could be inclusion of school and student attributes. School level attributes could include socioeconomic status or locale (i.e., urban, suburban, and rural). Additional student level attributes could include students with disabilities and students for whom English is their second language. Extending the research in this manner could provide insights into whether or not specific

student populations may better demonstrate their abilities or respond in a more engaging manner to interest-based tests.

Conclusion

In closing I have demonstrated the use of a unidimensional Rasch model to examine both interest-based and non- interest-based items in a measure across groups. I found that based on this assessment and sample, item level engagement varied little by group (*got preferred context*, *did not get preferred context*, and *no preference in context*). Furthermore, I found no significant differences in overall test level engagement by test (got preferred context, did not get preferred context), grade (6^{th}, 7^{th}, & 8^{th}), nor gender (male & female). Nor did I find significant interactions between test, grade, and gender.

A modification in my study design might provide for more generalizable results. Controlling for prior achievement level and drawing a more representative sample that takes into account socioeconomic status in addition to gender and ethnicity should be considered. Additional considerations such as other measures of engagement (e.g., self-reported measure, click tracking, eye movement tracking), expanding to a mixed methods design (e.g., student interviews), and ensuring data collection across similar devices to reduce any potential model effect may be beneficial as well. Such extensions may provide complementary information and insights into student interest as a component to engagement overall.

Choice of context matched to student interest, represented by a small change in text alone, may be insufficient to change the engagement patterns of 21^{st} century students within a short standardized assessment. Based on my study, there was no significant

impact of interest based choice on either reading assessment performance nor engagement for middle school students.

APPENDIX A

SAMPLE ITEM

Item 13: DOK: 2; Target grade 4

Goal Assessed: Vocabulary Acquisition and Use

CCSS.ELA-Literacy.L.4.6 - Acquire and use accurately grade-appropriate general academic and domain-specific words and phrases, including those that signal precise actions, emotions, or states of being (e.g., quizzed, whined, stammered) and that are basic to a particular topic (e.g., wildlife, conservation, and endangered when discussing animal preservation). (NGA & CCSSO, 2010 p.77)

Original item (FK: 5.7)

(a)

Animals (FK: 5.8)

(b)

Fantasy (FK: 5.6)

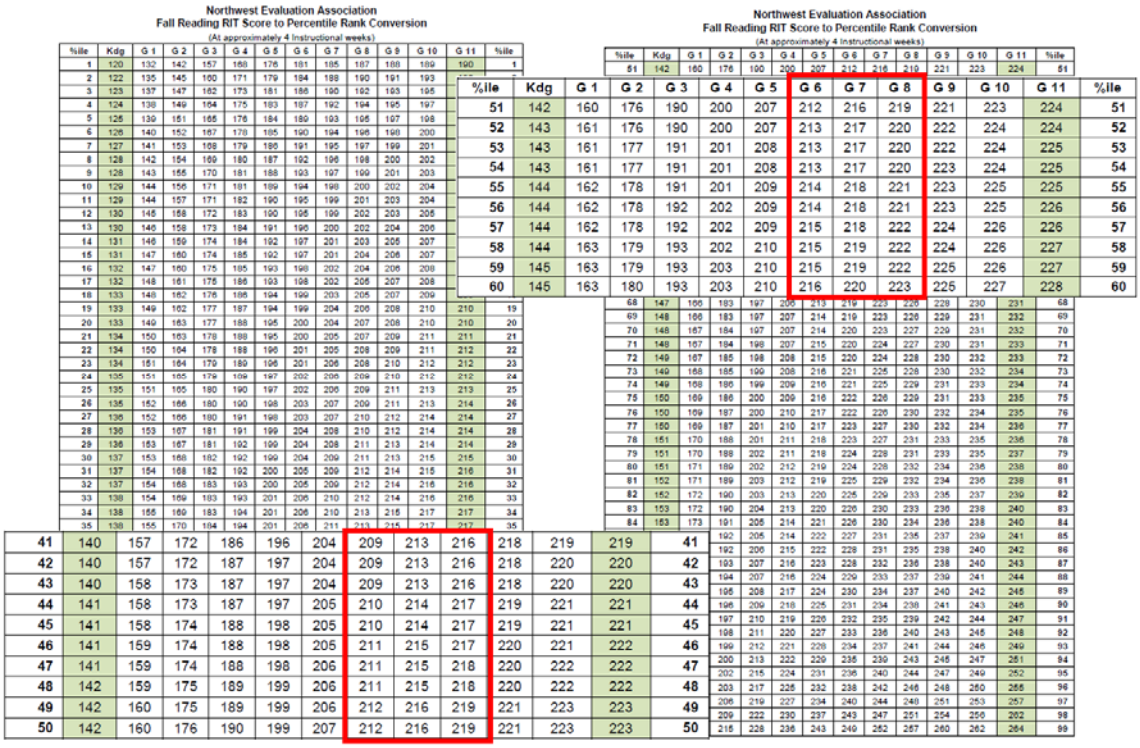
(c)

Sports (FK: 5.5)

(d)

APPENDIX B

MEDIAL QUINTILE OF FALL RIT SCALE NORMS FOR GRADES 6, 7, & 8



Northwest Evaluation Association
Fall Reading RIT Score to Percentile Rank Conversion
(At approximately 4 instructional weeks)

%ile	Kdg	G 1	G 2	G 3	G 4	G 5	G 6	G 7	G 8	G 9	G 10	G 11	%ile
51	142	160	176	190	200	207	212	216	219	221	223	224	51
52	143	161	176	190	200	207	213	217	220	222	224	224	52
53	143	161	177	191	201	208	213	217	220	222	224	225	53
54	143	161	177	191	201	208	213	217	220	223	224	225	54
55	144	162	178	191	201	209	214	218	221	223	225	225	55
56	144	162	178	192	202	209	214	218	221	223	225	226	56
57	144	162	178	192	202	209	215	218	222	224	226	226	57
58	144	163	179	193	202	210	215	219	222	224	226	227	58
59	145	163	179	193	203	210	215	219	222	225	226	227	59
60	145	163	180	193	203	210	216	220	223	225	227	228	60

Northwest Evaluation Association
Fall Reading RIT Score to Percentile Rank Conversion
(At approximately 4 instructional weeks)

%ile	Kdg	G 1	G 2	G 3	G 4	G 5	G 6	G 7	G 8	G 9	G 10	G 11	%ile
41	140	157	172	186	196	204	209	213	216	218	219	219	41
42	140	157	172	187	197	204	209	213	216	218	220	220	42
43	140	158	173	187	197	204	209	213	216	218	220	220	43
44	141	158	173	187	197	205	210	214	217	219	221	221	44
45	141	158	174	188	198	205	210	214	217	219	221	221	45
46	141	159	174	188	198	205	211	215	217	220	221	222	46
47	141	159	174	188	198	206	211	215	218	220	222	222	47
48	142	159	175	189	199	206	211	215	218	220	222	222	48
49	142	160	175	189	199	206	212	216	219	221	223	223	49
50	142	160	176	190	199	207	212	216	219	221	223	223	50

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth*. Portland, OR: NWEA.

APPENDIX C

TEST ITEM READABILITY CHARACTERISTICS

Items with fewer than 50 words are of insufficient length for in-depth readability analysis. Readability characteristics include narrativity (N), syntactic simplicity (SS), word concreteness (WC), referential cohesion (RC), and deep cohesion (DC) as identified through the use of a text analysis tool (McNamara & Graesser, n.d.).

Item 2: Original context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	30%	8%	99%	98%	93%
Item 2: Animal context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	48%	3%	95%	87%	51%
Item 2: Fantasy context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	35%	2%	97%	64%	74%

Item 2: Sports context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	67%	2%	99%	88%	54%

Item 5: Original context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	4%	77%	82%	10%	44%
Item 5: Animal context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	5%	72%	60%	7%	18%
Item 5: Fantasy context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It is high in syntactic simplicity	2%	78%	72%	2%	34%

Item 5: Fantasy context; Goal: Informational text					
which means that it has simple sentence structures. Simple syntax is easier to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.					
Item 5: Sports context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	9%	94%	83%	6%	9%

Item 6: Original context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. It is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	90%	95%	33%	18%	80%

Item 6: Animal context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend.	90%	68%	85%	68%	63%
Item 6: Fantasy context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	93%	87%	66%	29%	45%
Item 6: Sports context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	94%	41%	88%	64%	94%

Item 8: Original context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	47%	93%	99%	72%	96%
Item 8: Animal context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text has high word concreteness, which means there are many words that are easier to visualize	50%	58%	96%	6%	98%

Item 8: Animal context; Goal: Informational text					
This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. This text is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.					
Item 8: Fantasy context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	32%	84%	98%	64%	99%
Item 8: Sports context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	65%	62%	76%	74%	99%

Item 9: Original context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the	80%	12%	93%	38%	12%

Item 9: Original context; Goal: Literature					
reader has insufficient prior knowledge these gaps can be challenging.					
Item 9: Animal context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process.	94%	12%	66%	68%	4%
Item 9: Fantasy context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge.	74%	30%	82%	24%	53%
Item 9: Sports context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process.	89%	27%	69%	41%	49%

Item 11: Original context; Goal: Informational test					
Readability narrative	N	SS	WC	RC	DC
This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. It is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support,	51%	68%	59%	9%	82%

Item 11: Original context; Goal: Informational test					
comprehension may be facilitated, especially when the topic is unfamiliar.					
Item 11: Animal context; Goal: Informational test					
Readability narrative	N	SS	WC	RC	DC
This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge.	50%	61%	61%	4%	69%
Item 11: Fantasy context; Goal: Informational test					
Readability narrative	N	SS	WC	RC	DC
This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge.	47%	68%	61%	6%	61%
Item 11: Sports context; Goal: Informational test					
Readability narrative	N	SS	WC	RC	DC
This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge.	48%	62%	57%	7%	62%

Item 12: Original context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend.	70%	53%	94%	37%	52%
Item 12: Animal context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend.	92%	47%	74%	33%	41%

Item 12: Fantasy context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	85%	51%	88%	46%	97%
Item 12: Sports context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	86%	77%	93%	38%	95%

Item 13: Original context; Goal: Vocabulary Acquisition and Use					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has low word concreteness, which means there are many abstract words that are hard to visualize. Abstract texts may be more difficult to understand. It is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	89%	85%	16%	65%	99%

Item 13: Animal context; Goal: Vocabulary Acquisition and Use					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	98%	21%	96%	82%	92%
Item 13: Fantasy context; Goal: Vocabulary Acquisition and Use					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has low word concreteness, which means there are many abstract words that are hard to visualize. Abstract texts may be more difficult to understand. This text is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	82%	55%	2%	37%	89%
Item 13: Sports context; Goal: Vocabulary Acquisition and Use					
Readability narrative	N	SS	WC	RC	DC
This text is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. It is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	53%	72%	97%	25%	83%

Item 15: Original context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	64%	76%	93%	8%	4%
Item 15: Animal context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. It is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	81%	73%	67%	5%	94%
Item 15: Fantasy context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	78%	84%	32%	41%	85%

Item 15: Sports context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. This text is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	81%	88%	75%	5%	75%

Item 17: Original context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. This text is low in deep cohesion. This means there are few connective words that help to clarify relationships between events, ideas, and information. Because of this, the text may be more difficult to comprehend, especially for unfamiliar topics.	7%	20%	51%	51%	17%
Item 17: Animals context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps	15%	38%	75%	11%	51%

Item 17: Animals context; Goal: Informational text					
require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge.					
Item 17: Fantasy context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It has low word concreteness, which means there are many abstract words that are hard to visualize. Abstract texts may be more difficult to understand.	28%	31%	29%	47%	42%
Item 17: Sports context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is low in deep cohesion. This means there are few connective words that help to clarify relationships between events, ideas, and information. Because of this, the text may be more difficult to comprehend, especially for unfamiliar topics.	15%	28%	72%	59%	15%

Item 18: Original context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has low word concreteness, which means there are many abstract words that are hard to visualize. Abstract texts may be more difficult to understand. This text is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	85%	46%	18%	15%	8%

Item 18: Animal context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend.	91%	43%	71%	45%	47%
Item 18: Fantasy context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is low in deep cohesion. This means there are few connective words that help to clarify relationships between events, ideas, and information. Because of this, the text may be more difficult to comprehend, especially for unfamiliar topics.	70%	30%	77%	56%	18%
Item 18: Sports context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	91%	40%	59%	49%	91%

Item 20: Original context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It has high referential cohesion, suggesting that explicit words and ideas overlap between sentences. This overlap supports readers	29%	86%	97%	83%	3%

Item 20: Original context; Goal: Informational text					
by referring to ideas introduced earlier in the text, helping the reader make the connections the author intended. This text is low in deep cohesion. This means there are few connective words that help to clarify relationships between events, ideas, and information. Because of this, the text may be more difficult to comprehend, especially for unfamiliar topics.					
Item 20: Animal context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	35%	76%	99%	71%	52%
Item 20: Fantasy context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text has high referential cohesion, suggesting that explicit words and ideas overlap between sentences. This overlap supports readers by referring to ideas introduced earlier in the text, helping the reader make the connections the author intended. It is low in deep cohesion. This means there are few connective words that help to clarify relationships between events, ideas, and information. Because of this, the text may be more difficult to comprehend, especially for unfamiliar topics.	31%	88%	98%	74%	4%
Item 20: Sports context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is low in both referential and deep cohesion, suggesting that the reader may have to infer the	44%	80%	73%	17%	46%

Item 20: Sports context; Goal: Informational text					
relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.					

Item 21: Original context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend.	95%	53%	79%	55%	45%
Item 21: Animal context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	89%	77%	88%	59%	99%
Item 21: Fantasy context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand.	86%	54%	69%	30%	38%
Item 21: Sports context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	96%	66%	86%	56%	88%

Item 22: Original context; Goal: Vocabulary acquisition and use					
Readability narrative	N	SS	WC	RC	DC
This text is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge.	62%	15%	99%	16%	68%
Item 22: Animal context; Goal: Vocabulary acquisition and use					
Readability narrative	N	SS	WC	RC	DC
This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	67%	33%	81%	24%	31%
Item 22: Fantasy context; Goal: Vocabulary acquisition and use					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. It is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	78%	31%	96%	24%	31%
Item 22: Sports context; Goal: Vocabulary acquisition and use					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is low in syntactic simplicity which.	89%	25%	74%	43%	87%

Item 22: Sports context; Goal: Vocabulary acquisition and use					
means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.					

Item 23: Original context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text has high referential cohesion, suggesting that explicit words and ideas overlap between sentences. This overlap supports readers by referring to ideas introduced earlier in the text, helping the reader make the connections the author intended.	23%	32%	91%	70%	31%
Item 23: Animal context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge.	40%	54%	85%	4%	51%
Item 23: Fantasy context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. It is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.	48%	22%	67%	24%	26%

Item 23: Sports context; Goal: Informational text					
Readability narrative	N	SS	WC	RC	DC
This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. This text is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	22%	85%	87%	10%	70%

Item 24: Original context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. This text is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	61%	66%	88%	2%	99%
Item 24: Animal context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is low in both referential and deep cohesion, suggesting that the	75%	45%	93%	14%	12%

Item 24: Animal context; Goal: Literature					
reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.					
Item 24: Fantasy context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. This text has high word concreteness, which means there are many words that are easier to visualize and comprehend. It is high in both referential and deep cohesion, which may scaffold the reader, particularly if the content is challenging.	74%	22%	88%	60%	74%
Item 24: Sports context; Goal: Literature					
Readability narrative	N	SS	WC	RC	DC
This text is high in narrativity which indicates that it is more story-like and may have more familiar words. More story-like texts are typically easier to understand. It has low referential cohesion, indicating little overlap in words and ideas between sentences. Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge. This text is high in deep cohesion. There are relatively more connecting words to help clarify the relationships between events, ideas, and information. Because of this added support, comprehension may be facilitated, especially when the topic is unfamiliar.	73%	52%	59%	23%	98%

REFERENCES CITED

Achieve. (2013). *Closing the expectations gap*. Retrieved from http://www.achieve.org/ClosingtheExpectationsGap2013

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science, 205(4412),* 1289-1292.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological tests*. Washington, D.C.: AERA.

Anand, P. G., & Ross, S. M. (1987). Using Computer-Assisted Instruction to Personalize Arithmetic Materials for Elementary School Children. *Journal of Educational Psychology*, *79*(1), 72–78.

Atkinson, J. W. (1964). *An introduction to motivation*. Princeton, NJ: Van Nostrand.

Attribution. (2019). Retrieved from Lexicon.com, Oxford University Press website: https://www.lexico.com/en/definition/attribution

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.

Bernacki, M. L., & Walkington, C. (2018). The role of situational interest in personalized learning. *Journal of Educational Psychology*, *110*(6), 864–881.

Bernacki, M., & Walkington, C. (2014). The Impact of a Personalization Intervention for Mathematics on Learning and Non-Cognitive Factors. *EDM*.

Brehm, J. W. (1966). *A theory of psychological reactance*. New Yory, NY: Academic Press.

Brehm, J. W., & Brehm, S. S. (1981). *Psychological reactance - a theory of freedom and control*. New York, NY: Academic Press.

Brozo, W. G., Sulkunen, S., Shiel, G., Garbe, C., Pandian, A., & Valtin, R. (2014). Reading, gender, and engagement. *Journal of Adolescent & Adult Literacy*, *57*(7), 584–593.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715–730.

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety*. San Francisco, CA: Jossey-Bass, Inc.

de Ayala, R. J. (2009). *The theory and practice of item response theory* (D. A. Kenny & T. D. Little, Eds.). New York, NY: The Guildford Press.

Deci, E. L. (1975). *Intrinsic motivation* (E. Aronson, Ed.). New York, NY: Plenum Press.

Descartes, R. (1911). *Meditations on first philosophy (E. S. Haldane, Trans.)*. London, England: Cambridge University Press (Original work published 1641).

Descartes, R. (1975). *The passions of the soul (E. S. Haldane & G. R. T. Ross, Trans.)*. London, England: Cambrdge University Press (Original work published 1649).

Dillard, J. P., & Shen, L. (2005). On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, *72*(2), 144–168.

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

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Field, A. (2018). *Discovering Statistics Using IMB SPSS Statistics* (5th Ed.). SAGE Publications.

Fredricks, J., McColskey, W., Meli, J., Mordica, J., Montrosse, B., & Mooney, K. (2011). Measuring student engagement in upper elementary through high school: A description of 21 instruments. In *(Issues & Answers Report, REL 2011–No. 098)* (Vol. 098). Retrieved from http://ies.ed.gov/ncee/edlabs

Glasser, W. (1998). *Choice theory: A new psychology of personal freedom*. New York, NY: HarperCollins Publishers.

Greenwood, C. R., Horton, B. T., & Utley, C. A. (2002). Academic engagement: Current perspectives on research and practice. *School Psychology*, *31*, 328–349.

Greenwood, Charles R, Carta, J. J., Kamps, D. M., Terry, B., & Delquadri, J. (1994). Development and validation of standard classroom observation systems for school practitioners: Ecobehavioral assessment systems software (EBASS). *Exceptional Children*, *61*(2), 197–210.

Guthrie, J. T., Hoa, A. L. W., Wigfield, A., Tonks, S. M., Humenick, N. M., & Littles, E. (2007). Reading motivation and reading comprehension growth in the later elementary years. *Contemporary Educational Psychology*, *32*(3), 282–313.

Hintze, J., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, *33*(2), 258–270.

Holland, P. W., & Thayer, D. T. (1985). An alternative definition of the ETS delta scale of item difficulty. In *ETS Research Report No. 85-43*. Princeton, NJ: Educational Testing Service.

Hong, S. M., & Faedda, S. (1996). Refinement of the Hong psychological reactance scale. *Educational & Psychological Measurement*, *56*, 173–182.

Howell, D. C. (2013). *Statistical methods for psychology* (Eighth; J. Hague, Ed.). Belmont, CA.: Wadsworth, Cengage Learning.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory* (R. M. Elliott, Ed.). New York, NY: D. Appleton-Century Company, Inc.

Ingebo, G. S. (1997). *Probability in the measure of achievement: Rasch measurement*. Chicago, IL: MESA Press.

Ivey, G., & Broaddus, K. (2001). "Just plain reading": A survey of what makes students want to read in middle school classrooms. *Reading Research Quarterly*, *36*(4), 350–377.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researchers handbook* (4th Ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf

Kukla, A. (1972). Attributional detreminants of achievement-related behavior. *Journal of Personality and Social Psychology*, *21*, 166–174.

LaPorte, R. E., & Nath, R. (1976). Role of performance goals in prose learning. *Journal of Educational Psychology*, *68*(3), 260–264.

Linacre, J. M. (2015). *Winsteps: Rasch Measurement Computer Program*. Retrieved from www.winsteps.com

Locke, E. A. (1968). Toward a theory of task motivation and incentives. *Organization Behavior and Human Performance*, *3*(2), 157–189.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717.

Logan, S., Medford, E., & Hughes, N. (2011). The importance of intrinsic motivation for high and low ability readers' reading comprehension performance. *Learning and Individual Differences*, *21*(1), 124–128.

Maier, S. F., & Seligman, M. E. P. (1975). Learner helplessness: Theory and evidence. *Journal of Experimental Psychology*, *105*(1), 3–46.

Markus, H. (1977). Self-schemata and processing informaion about the self. *Journal of Personality and Social Psychology*, *35*(2), 63–78.

McDougall, W. (1908). *An introduction to social psychology*. London, England: Methuen & Co.

McKenna, M. C., Conradi, K., Lawrence, C., Janj, B. G., & Meyer, J. P. (2012). Reading attitudes of middle school students : Results of a U . S . survey. *Reading Research Quarterly*, *47*(3), 283–306.

McLeod, S. A. (2018). Cognitive dissonance. Retrieved from https://www.simplypsychology.org/cognitive-dissonance.html

McNamara, D., & Graesser, A. (n.d.). Coh-Metrix. Retrieved from http://tea.cohmetrix.com

Mühlberger, C., Klackl, J., Sittenthaler, S., & Jonas, E. (2019). The approach-motivational nature of reactance: Evidence from asymmetrical frontal cortical activation. *Motivation Science*, 1–26.

Mullder, P. (2018). Vroom's expectancy theory. Retrieved from https://www.toolshero.com/psychology/theories-of-motivation/vrooms-expectancy-theory/

National Center for Education Statistics. (2016). Public school students eligible for free or reduced-price lunch. Retrieved from https://nces.ed.gov/fastfacts/display.asp?id=898

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards, English Language Arts Standards*. Washington, D.C.: National Governors Association Center for Best Practices, Council of Chief State School Officers.

Northwest Evaluation Association. (2011). *Technical manual for Measures of Academic Progress (MAP) and Measures of Academic Progress for Primary Grades (MPG)* (pp. 1–204). pp. 1–204. Portland, OR.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (Third; J. Vaicunas & J. R. Belser, Eds.). New York, NY: McGraw-Hill, Inc.

Reeve, J. (2015). *Understanding motivation and emotion* (C. Johnson, Ed.). Hoboken, NJ: Wiley.

Robitzch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules*. Retrieved from https://cran.r-project.org/package=TAM

Schnipke, D. L. (1996). How contaminated by guessing are item-parameter estimates and what can be done about it? *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*. New York, NY: (ERIC Document Reproduction Service No. ED400276).

Schnipke, D. L., & Scrams, D. J. (1997). *Modeling item response times with a two-state mixture model: A new approach to measuring speededness*. *34*(3), 213–232.

Seligman, M. E. P. (1975). *Helplessness. On depression, development, and death.* San Francisco, CA.: W. H. Freeman.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*(1), 34–49.

Shapiro, E. S. (2004). *Academic skills problems: Direct assessment and intervention* (3rd ed.). New York, NY: The Guilforde Press.

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., & Greenberg, J. (2015). Understanding psychological reactance: New developments and findings. *Zeitschrift Fur Psychologie / Journal of Psychology*, *223*(4), 205–214.

Steindl, C., Klackl, J., & Jonas, E. (2016). *The neural correlates of psychological reactance: An fMRI study*. University of Salzburg, Austria.

Sundre, D. L. (1999). Does examinee motivation moderate the relationship between test consequences and test performance? *Annual Meeting of the American Educational Research Association*.

Sundre, D. L., & Moore, D. L. (2002). The student opinion scale: A measure of examinee motvation. *Assessment Update*, *14*(1), 8–13.

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth*. Portland, OR: NWEA.

Vroom's expectancy theory. (n.d.). Retrieved from University of Cambridge's Institute for Manufacturing website: https://www.ifm.eng.cam.ac.uk/research/dstools/vrooms-expectancy-theory/

Vroom, V. H. (1964). *Work and motivation*. Oxford, England: Wiley.

Walkington, C. A. (2013). Using Adaptive Learning Technologies to Personalize Instruction to Student Interests: The Impact of Relevant Contexts on Performance and Learning Outcomes. *Journal of Educational Psychology*, *105*(4), 932–945.

Walkington, C., & Leigh, L. A. (2015). *Personalization to Student Interests in Reasoning Mind : Depth , Grain Size , and Ownership*. Retrieved from http://www.cwalkington.com/RM_Study_Walkington_Mingle.pdf

Walkington, C., Petrosino, A., & Sherman, M. (2013). Supporting algebraic reasoning through personalized story scenarios: How situational understanding mediates performance. *Mathematical Thinking and Learning*, *15*(2), 89–120.

Webb, N. (1999). *Alignment of science and mathematics standards and assessments in four states*. Madison, WI: Council of Chief State School Officers and National Institute for Science Education Monograph No. 18.

Weiner, B. (1972). Attribution theory, achievement motivation, and the educational process. *Review of Educational Research*, *42*(2), 203–215.

Weiner, B., Heckhausen, H., Meyer, W. U., & Cook, R. E. (1972). Causal ascriptions and achievement motivation: A conceptual analysis of effort and reanalysis of locus of control. *Journsal of Personality and Social Psychology*, *21*, 239–248.

Weiner, B., & Kukla, A. (1970). An attibutional analysis of achievement motivation. *Journal of Personality and Social Psychology*, *15*, 1–20.

Weiner, B., & Potepan, P. A. (1970). Personality characteristics and affective reactions towards exams of succeeding and failing college students. *Journal of Educational Psychology*, *61*, 144–151.

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, *66*(5), 297–333.

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.

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment*, *10*(1), 1–17.

Wise, S. L., & Jensen, N. (2019). Student test engagement: How it's measured and why you should care. *Paper Presented at Fusion*. St. Louis, MO.

Wise, S. L., Kingsbury, G., Thomason, J., & Kong, X. (2004). An investigation of motivation filtering in a statewide achievement testing program. *Annual Meeting of the National Council of Measurement in Education*. San Diego, CA.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010). An investigation of the relationship between time of testing and test-taking effort. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*. Denver, CO.