AN INVESTIGATION OF TEST-TAKING EFFORT IN A COMPUTER-ADAPTIVE TEST OF READING

By

James Eugene Los

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

School Psychology—Doctor of Philosophy

2019

ABSTRACT

AN INVESTIGATION OF TEST-TAKING EFFORT IN
A COMPUTER-ADAPTIVE TEST OF READING

By

James Eugene Los

Educators use academic testing to measure the knowledge and skills of their students, but research has shown that some students exhibit minimal test-taking effort (TTE) on low-stakes tests (Wise & Kong, 2005; Wise 2015). Given the importance of test validity, there is a need for further research on the contexts in which low TTE occurs and possible correlates of low TTE. The goals of the current study were to measure the prevalence of low TTE in elementary and middle school contexts, identify groups for whom low TTE is particularly apparent, and examine whether motivational variables are associated with low TTE. Study I involved the analysis of item-level test data for students in grades four and eight (N = 572,847) to identify the proportion of students who submitted responses so rapidly that they are not considered valid. In Study II, students in grades seven and eight (N = 675) completed an online survey that measured their expectancy and value beliefs (Wigfield & Eccles, 2000) related to taking the STAR Reading test (Renaissance Learning, 2014). Results of logistic regression analyses indicated that grade level, gender, race/ethnicity, attainment value beliefs, and cost beliefs were significantly associated with the odds of low TTE. These findings suggest that potential ways to improve student TTE may include informing students about how a test will be used to enhance their learning. Related suggestions for future research that might meaningfully extend the present findings are provided.

Leah Jean,
I like you and I love you.

ACKNOWLEDGMENTS

Thank you to all of my friends, family, and colleagues who have supported me. There are many other people who deserve my deepest gratitude. If every one of them were written down, I suppose that even the whole world would not have room for the pages that would be written.

Thank you to my family: Scotty, Wendi, Kevin, Lisa, Charissa, Tricia, Jake, Jon, Ady, Tyler, Mitch, Cara, Paul, Jeff, Cole, Jackson, Alex, grandpas, grandmas, and all the rest of you.

Thank you to my crew: Aaron, Jack, Lucas, Parker, Kailie, Kali. I'm glad you're mine.

Thank you to my friends: Lee Gordo, Ben, Derek, Kelli, Jenna, Rheadon, Dana, Ali, Jeshua, Allie, Zach, Matt, Tyler, AJ, Kyle, Bryant, Peter, Donny, Bradley, Chris, Drew, Derek, Jacob, and Corey. In loving memory of my friend Dr. Adam Winstrom.

Thank you to my cohort of friends and esteemed colleagues: Addam, Ali, Allie, Becky, Courtney, Dani, Danielle, Jamie, Katie, Kiley, and Rick.

Thank you to Rob, Science Mike, Vishnu, Hillary, William, Father Richard, and Rachel.

Thank you to my supervisors and mentors: Kurt, Sherri, Jason, Luke, Lillian, and Trisha.

Thank you to my professors at MSU and Calvin: Dr. Aupperlee, Dr. Carlson, Dr. Fine, Dr. Oka, Dr. Rispoli, Dr. Windram, Dr. Yonker, Dr. Tellinghuisen, Dr. Riek, Dr. DeHaan, Dr. Stehouwer, et al. Thank you to Holly Boehle, Brandi-Lyn Mendham, and Calvin DeKuiper.

Finally, thank you to my adviser and dissertation committee members: Dr. Sara Witmer, Dr. Cary Roseth, Dr. Adrea Truckenmiller, and Dr. Martin Volker.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Purpose

The purpose of the current study was to investigate student-level correlates of test-taking effort (TTE) on a low-stakes, computer-adaptive test (CAT) in reading. Educators regularly use testing to gather information about the knowledge and skills of students, but inferences made based on test scores are only appropriate if these scores represent reliable and valid indicators of the students' "true" proficiency (Salvia, Ysseldyke, & Bolt, 2013). In contemporary models of assessment, test users rely on the assumption that the examinees have given appropriate effort when completing a test (Eklöf, 2010; Wise, 2015). However, research on educational testing has suggested some test-takers show little effort during testing, as demonstrated by responding with rapid-guessing behavior (RGB; Schnipke & Scrams, 1997) with accuracy rates comparable to chance (Setzer, Wise, van den Heuvel, & Ling, 2013; Swerdzewski, Harmes, & Finney, 2011).

This problem is especially evident in low-stakes testing contexts, in which test scores may be significant to educators but carry no personal consequences for the students (Eklöf, 2010; Wise, 2014; Wise & DeMars, 2005). In fact, researchers have documented that as many as 35% of students exhibit low TTE if scores do not affect their grades (Rios, Liu, & Bridgeman, 2014), although the *reasons why* some students tend to show non-effortful responding are unknown. Widespread disengagement during low-stakes educational testing could be a considerable threat to the validity of testing systems because scores from low-effort respondents yield no meaningful information about the actual proficiency of the students (Cronbach, 1960; Haladyna & Downing, 2004; Wise & DeMars, 2005). Moreover, if a large proportion of students exhibit low TTE, this can adversely affect the psychometric properties of the aggregate test data (Wise & Kong, 2005).

Even though ensuring that students exert adequate TTE is essential for appropriate testing practices, there have been surprisingly few empirical studies of TTE in K–12 academic contexts (Wise, 2014). Given the importance of test validity, there is a clear need for additional research examining both the extent to which low TTE is a problem for educational tests and the possible correlates of low TTE. With a greater understanding of the factors related to disengagement from testing, it may be possible to develop targeted strategies for promoting more effortful responding on low-stakes tests. Therefore, to help inform policies and practices that address the problem of low TTE, the primary goals of the current investigation were to contribute to the extant research on the following: a) the contexts in which low TTE occurs, b) the groups of students for whom low TTE is particularly apparent, and c) the motivational variables associated with low TTE.

Background

The appropriate use of educational testing practices has been one of the central concerns in the national discourse on education over the past two decades, and recent federal legislation in the Every Student Succeeds Act of 2015 (ESSA, 2015) further emphasized the importance of using educational testing for measuring student proficiency and informing classroom instruction. According to the Institute of Educational Sciences (IES), educational testing data should be used as "part of an ongoing cycle of instructional improvement" (Hamilton et al., 2009, p. 8). Indeed, reports suggest that K–12 students have experienced a substantial increase in time spent taking state-mandated accountability tests, district-wide standardized tests, and teacher-made tests over the past few years (Hart et al., 2015). Further, the proliferation of technology in the classroom has contributed to an increase in the development and use of computer-based testing (CBT), which can have practical advantages over paper-and-pencil testing (PPT) in terms of efficiency and measurement precision (Barnard, 2015; Shapiro, Dennis, & Fu, 2015; Weiss; 2011).

In response to growing concerns from educators and policy-makers about the expanding use of educational testing programs, the U.S. Department of Education (ED) called for efforts to promote more appropriate testing practices in K–12 public schools. The Obama administration's Testing Action Plan (ED, 2015) stated that all educational tests must be high quality, supportive of fairness, worthwhile for students and teachers, and tied to improved learning. This demand for quality testing practices reflects the consensus among educators that test scores are only useful if they convey *valid* information about the knowledge and skills of students (Salvia et al., 2013).

According to the *Standards for Educational and Psychological Testing*, validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 11). That is, test developers and users need to consider the evidence that scores are reasonably free from the influence of *construct irrelevance* (i.e., processes extraneous to the intent of the test).

In general, test validation includes the evaluation of five sources of data: 1) evidence based on test content, 2) evidence based on response processes, 3) evidence based on internal structure, 4) evidence based on relations to other variables, and 5) evidence based on the consequences of testing (AERA et al., 2014; Salvia et al., 2013). Even though it is standard practice for technical manuals of assessments to discuss particular types of validity evidence (e.g., convergent or discriminant evidence, predictive or concurrent test-criterion relationships), little attention has been given to test validity based on response process. This issue is concerning because unless test developers and users have data to suggest the test takers engaged in cognitive processes consistent with the intended cognitive model of the test, it remains possible that some extraneous factors could have differentially influenced the performance of specific test takers.

Researchers can gather data based on examinee response processes by asking them to explain how they reached their answers, maintaining and analyzing records of their work, or tracking and recording eye movement or response times (Salvia et al., 2013; AERA et al., 2014). Indeed, a central assumption in educational testing is that students are using some meaningful cognitive processes directed toward determining the answer. For instance, if a test is purported to measure reading comprehension, a fundamental assumption is that students *actually engaged in reading* the content of the passage, item prompt, and response options. However, several experts have suggested that non-effortful responding (i.e., guessing) may be one construct-irrelevant response process that could pose a significant threat to test validity. That is, scholars have asserted that the validity of any inferences made based on test scores is directly dependent on the amount of TTE students exerted while taking the test (Wise & DeMars, 2006; Wise & Kong, 2005). Indeed, growing research evidence has supported this notion, as several scholars have documented the adverse effects of low TTE on test performance (Cronin et al., 2005; Sundre & Kitsantas, 2004; Wise, 2006; Wise, Bhola, & Yang, 2006; Wise & DeMars, 2006).

For this reason, educational researchers, test developers and test users need to recognize that responding correctly to test items requires examinees to possess and demonstrate not only the *target academic skills* of the assessment but also the level of *test-taking motivation* necessary for appropriately engaging in the test and providing a valid response (Eklöf, 2010). Although there has been empirical research on the extent to which deficits in specific *academic skills* adversely influence student performance on tests not designed to measure those deficit academic skills, few researchers have examined whether student differences in *motivational factors* could similarly affect how students demonstrate their academic skills during educational testing.

Moreover, it is possible that having low motivation for test-taking could prevent some students from exhibiting their actual knowledge and skills on low-stakes tests in a variety of subjects (when the tests demand high levels of mental effort), in the same way that deficits in specific academic skills can prevent some students from showing their true ability when taking tests in other subject areas (e.g., students with low reading skills might not be able to adequately demonstrate their underlying math problem-solving skills if the test has high reading demands).

Despite growing recognition in several countries that appropriate TTE is fundamental to valid test interpretation and use (e.g., Barry, Horst, Finney, Brown, & Kopp, 2010; Eklöf, 2010), there have been few studies on the extent to which low TTE contributes construct-irrelevant variance to test scores and thereby limits the measurement of the academic skills of some subgroups or individual students. Ultimately, this issue relates to the core principle of fairness in assessment, and the current lack of research on TTE in K–12 academic contexts points to an important area for further empirical investigation. Guided by an application of the contemporary expectancy–value theory of motivation and engagement (Wigfield & Eccles, 2000), the current investigation aimed to extend the previous research on student TTE through two empirical studies focused on the student-level correlates of TTE in K–12 educational testing contexts.

Importance

Currently, questions remain about why some students exhibit appropriate TTE whereas others disengage and respond with minimal effort (Wise, 2014). More specifically, educators, researchers, and policy-makers could benefit from additional information on three major issues related to TTE: 1) the prevalence of K–12 students who exhibit low TTE, 2) the characteristics of students who exhibit low TTE, and 3) the alterable psychological correlates of exhibiting low TTE (which could thereby help inform efforts designed to prevent low TTE from occurring).

First, it is crucial for test developers and educators to identify the extent to which low TTE might be a problem for school-age students. Each year, millions of elementary and middle school students across the nation take low-stakes CATs in reading (e.g., STAR Reading) as part of school-wide benchmark testing programs (Renaissance Learning, 2016), and educators use scores from these tests to inform curricula, instruction, and intervention (Ysseldyke et al., 2006). However, as previously stated, the resulting scores of test examinees who engaged in cognitive processes extraneous to the target construct (such as RGB) might not be considered valid. That is, interpreting and using testing data containing scores from disengaged test-takers may be inappropriate due to the deleterious effects of low TTE on test scores (Sundre & Wise, 2003; Wise & DeMars, 2009). Thus, if any particular students or groups of students exhibit low TTE, their resulting scores might under-represent their actual skills in the tested domain. Given the potential consequences of using invalid testing data, it is essential for additional research to clarify the proportion of students whose scores could potentially be invalid due to low TTE.

Despite growing evidence that low TTE is a substantial problem for college students when taking university-mandated tests (which often have practical utility for the institutions but no personal consequences or incentives for students), few researchers to date have studied TTE in *elementary* or *middle school* samples. In the previous research on student TTE in elementary and middle school samples, the proportions of students identified as exhibiting low TTE have varied considerably, ranging from 1.4% (Wise, Kingsbury, Thomason, & Kong, 2004) to 11.6% (Wise, Ma, Kingsbury, & Hauser, 2010). These findings suggest that some elementary and middle school students indeed demonstrate high rates of RGB when taking educational CBTs in low-stakes contexts, although researchers have not reported consistent estimates of the current prevalence of low TTE on the low-stakes academic tests commonly used in K–12 schools.

For instance, Wise and colleagues (2010) documented the proportions of students in grades 3–8 who were identified with low TTE on a CAT in reading and disaggregated the results by grade level and the time of day the test was taken. The proportions of low-effort respondents ranged from 0.5% (for third-grade students taking a test at 7:00 a.m.) to 5.9% (for eighth-grade students taking a test at 2:00 p.m.). These figures suggest a notable proportion of students exhibit low TTE on educational tests under certain conditions, although replication studies are needed to establish further that these rates exist in other contexts. If high proportions of elementary or middle school students exhibit RGB on low-stakes tests, that would suggest low TTE might be distorting the results of the aggregate test data in these contexts (Wise, 2015). Still, even if the overall rate of RGB is small, educators would need to acknowledge that the validity of *individual* test scores would be compromised severely for any student identified as exhibiting exceptionally low TTE. Better understanding this issue in the context of K–12 schools may help inform future testing research focused on examinee response processes. For this reason, one of the primary goals of the current study was to investigate the current proportion of elementary or middle school students who exhibit low TTE when taking a commonly used, low-stakes CAT in reading.

Second, if additional research is conducted to identify whether any subgroups of students are particularly likely to show low TTE, this information could be useful for informing efforts to improve student TTE. Previous research on academic motivation has indicated that students from different demographic groups may vary in their general motivation (Wentzel & Brophy, 2014). In the area of reading, there are documented differences in motivation at school by age, gender, race, ethnicity, and disability status (Archambault, Eccles, & Vida, 2010; Baird, Scott, Dearing, & Hamill, 2009; Baker & Wigfield, 1999; Battle, 1979; Durik, Vida, & Eccles, 2006; Eccles, 1984; Grolnick & Ryan, 1990; Schunk, Meece, & Pintrich, 2014; Wentzel & Miele, 2016).

Whether similar group differences might emerge in the specific domain of test-taking is currently unknown. This issue warrants further investigation, as there has been growing interest in research that addresses the degree to which tests might differentially support the test-taking motivation of specific subgroups of examinees (AERA, APA, & NCME, 2014). For example, if some subgroups of students perceive the content in an assessment to be especially uninteresting, culturally irrelevant, unfamiliar, or confusing, it could differentially limit the TTE of students from that group. In sum, it would be helpful to identify student characteristics associated with low TTE to help inform how educators might ameliorate any group disparities in TTE that exist.

Lastly, it is essential for researchers to investigate potential *reasons why* some students or groups of students disengage from testing. Scholars have argued that better understanding the dynamics of test-taking is critical for developing effective strategies for eliciting appropriate TTE from students on low-stakes educational tests (Wise & DeMars, 2005). The identification of *malleable* correlates of TTE could be particularly useful for informing how educators design and implement prevention and intervention strategies intended to improve student TTE. Indeed, there is research evidence to suggest educators can use instructional practices or targeted interventions to enhance motivation in academic domains (Guthrie et al., 2004; Wigfield & Wentzel, 2007). In fact, a recent intervention study by Liu, Rios, and Borden (2015) showed that test practitioners could proactively improve the TTE of college students by administering a brief motivational prompt to the experimental group before taking the test, and this increase in TTE was associated with outperforming a control group by 0.63 standard deviations. Accordingly, if additional research is conducted to identify alterable motivational variables associated with low TTE, it is possible that future efforts to improve student TTE (and prevent RGB) could be strengthened by explicitly targeting the unique motivational needs of students from different groups.

Rationale for Current Study

Therefore, it is critical for researchers and practitioners to learn more about the issue of low TTE in academic contexts to ensure test validity, promote fairness, and develop strategies that support student TTE on non-consequential tests. With students taking increasing numbers of tests at school, it is essential for educators to understand how serious the problem of low TTE might currently be in K–12 schools. Relatedly, it would be helpful for researchers to investigate why some students demonstrate appropriate levels of engagement whereas others have been found to disengage and guess randomly. According to Eccles and colleagues' expectancy–value theory of motivation (EEVT; see Wigfield & Eccles, 2000), students' achievement-related behaviors (such as their effort and persistence on a task) can be explained by the students' subjective perceptions concerning 1) how successfully they expect to perform the task (expectancy beliefs) and 2) how much they value engaging in the task (value beliefs).

Generally, the EEVT would imply that students who believe they are unable to answer a test item correctly (low expectancy) or believe they have no meaningful reason to try (low value) would be hypothesized to exhibit lower TTE (which could be manifested by high rates of RGB). Recently, scholars have cited the expectancy–value theory to explain why some test-takers have exhibited high levels of RGB in previous studies, asserting that low-stakes academic tests may have exceptionally low perceived value for particular students, such that these students exert minimal TTE (Setzer et al., 2013). Wise and DeMars (2005) postulated, "For these students, the task of doing well on the test will have little attainment, intrinsic, or utility value. Moreover, these students will be aware of the costs associated with the assessment test (i.e., being denied the opportunity to engage in more valued activities). Thus, the Eccles–Wigfield model would predict low effort on low-stakes assessment tests from students with weak value beliefs" (p. 3).

Still, there have been few studies to date in which researchers have directly tested the extent to which students' expectancy or value beliefs about testing relate to the likelihood they exhibit low TTE (i.e., demonstrate pervasive RGB) on low-stakes educational tests. Guided by an application of the EEVT, the current investigation was designed to extend previous research on the dynamics of TTE to inform future research and practice. The purpose of this study was to identify the prevalence of low TTE in students in grades 4–8 and examine the extent to which *student demographic characteristics*, *test-taking expectancy beliefs*, and *test-taking value beliefs* are associated with the likelihood that students exhibit *low TTE* on a low-stakes CAT in reading. To that end, the specific research questions for the current investigation were as follows:

Research Questions

1. What proportion of students in grades 4–8 exhibit low TTE on a CAT in reading, as determined by Response Time Effort (RTE; Wise & Kong, 2005)?
2. To what extent do student demographic variables relate to the likelihood students exhibit low TTE on a CAT in reading?
3. Do students differ in test-taking expectancy and value beliefs by student demographic variables?
4. To what extent do student test-taking expectancy and value beliefs relate to the likelihood students exhibit low TTE on a CAT in reading?

CHAPTER II

LITERATURE REVIEW

This literature review summarizes previous research relevant to the current study. First, the fundamental constructs of motivation, engagement, and effort are described. Next, the contemporary EEVT is presented as it relates to the current investigation, and a conceptual model for empirical research on TTE is described. After that, critical issues in research on TTE are discussed, the use of response time methods for measuring TTE on CBTs is explained, and previous empirical studies of TTE are reviewed. An overview of the proposed student-level correlates of TTE follows, focusing on the previous empirical research on student demographic characteristics and motivational variables associated with TTE. This literature review concludes with the rationale for the current application of the expectancy–value theory to an empirical study of TTE, and the specific research questions and hypotheses for this study are presented.

Theoretical Background and Conceptual Framework

Motivation, engagement, and effort. The current investigation was contextualized within the more general research literature on academic motivation and student engagement. Broadly construed, motivation and engagement refer to multidimensional patterns of thoughts, feelings, and behaviors that facilitate, explain, or indicate individuals' goal-directed actions. As such, several aspects of academic motivation and student engagement are integral to all aspects of the learning process, and so these two concepts have received considerable attention over recent decades. For more comprehensive reviews of the major theoretical perspectives on academic motivation and student engagement, readers are directed to volumes by Wentzel and Miele (2016) and Christenson, Reschly, and Wiley (2012), respectively.

Currently, there are competing views about how researchers should define motivation and engagement, differentiate and measure their components, and conceptualize the relationship between these two concepts, but the general consensus among scholars is that motivation and engagement are considered two distinct, yet related "metaconstructs" or organizing frameworks (Christenson et al., 2012). Moreover, there are no universally accepted definitions for motivation or engagement, but experts in these areas of educational psychology research have proposed the following. Wentzel and Miele (2016) defined motivation broadly as "a set of interrelated desires, goals, needs, values, and emotions that explain the initiation, direction, intensity, persistence, and quality of behavior" (p. 1). Christenson and colleagues (2012) defined student engagement as "the student's active participation in academic and co-curricular or school-related activities, and commitment to educational goals and learning," adding, "It is a multidimensional construct that consists of behavioral (including academic), cognitive, and affective subtypes" (p. 816).

Thus, for the sake of conceptual clarity in this review, motivation refers to psychological processes (i.e., thoughts and feelings) which *facilitate* active participation in an academic task, whereas engagement refers to observable *indicators* of active participation in the task. In other words, "Motivation refers to the underlying sources of energy, purpose, and durability, whereas engagement refers to their visible manifestation" (Skinner & Pitzer, 2012, p. 22). To summarize, academic motivation represents the latent psychological processes (i.e., energy and purpose) that initiate and sustain students' goal-directed actions, and student engagement refers to students' goal-directed actions themselves (characterized by effort, intensity, and persistence). As such, students with higher motivation for a task are likely to demonstrate higher engagement in the task (as indicated by the amount of effort the students exert toward completing the task).

Furthermore, because *effort* (the focus of the current study) is closely related to other key constructs in theories of motivation and engagement—with definitions of effort overlapping with numerous constructs in both fields—differentiating motivation and engagement is pertinent to research on TTE. Broadly construed, there are competing perspectives on whether effort should be considered an *indicator* (i.e., "markers or descriptive parts inside a construct") or *facilitator* (i.e., "explanatory causal factors, outside the target construct, that have the potential to influence the target") of engagement (Skinner & Pitzer, 2012, p. 25). Numerous experts have argued effort is understood best as one behavioral *indicator* of engagement, whereas motivational variables are the *facilitators* of effort (Appleton, Christenson, & Furlong, 2008; Newmann, Wehlage, & Lamborn, 1992; Skinner & Pitzer, 2012; Skinner, Kindermann, Connell, & Wellborn, 2009).

In the domain of educational test-taking, one of the widely accepted definitions of TTE is "a student's engagement and expenditure toward the goal of attaining the highest possible score on the test" (Wise & DeMars, 2005, p. 2). Accepting this definition, the primary construct of interest in the current study, TTE, can be conceptualized as one specific facet of *test-taking engagement* (which refers more broadly to psychological processes and observable solution-focused behaviors directed toward responding correctly to test items). Other scholars described TTE as "the extent to which an examinee gives his or her best effort to the test, with the goal being to accurately represent what one knows and can do in the content area covered by the test" (Barry et al., 2010). By contrast, *test-taking motivation* represents the underlying psychological processes whereby solution-focused responses to items (i.e., TTE) are instigated and sustained. Test-taking motivation has been defined as "the willingness to engage in working on test items and to invest effort and persistence in this undertaking" (Baumert & Demmrich, 2001, p. 441).

To summarize, a fundamental assumption in the current study is that students' test-taking motivation directly influences the amount of TTE the students exert when engaging in the test. Although several theories of motivation and engagement could be useful for informing empirical research on student TTE, one theoretical framework guided the development of the current study. Specifically, the contemporary Eccles et al. expectancy–value theory (EEVT), derived from the work of Eccles, Wigfield, and colleagues (see Wigfield & Eccles, 2000; Wigfield, Tonks, & Klauda, 2009; 2016), provided the theoretical framework for the current investigation.

Expectancy–value theory of achievement behaviors. The EEVT proposes the existence of relationships among individual's expectancies for success, personal values, task choices, beliefs about achievement, self-concepts of ability, goals, self-schemata, affective memories, perceptions of others' attitudes and expectations for them, and perceptions of past achievement outcomes (Eccles, 2005; Wigfield & Eccles, 2000; Wigfield et al., 2009). The primary constructs in the EEVT, expectancies and values, are considered internal, cognitive beliefs that influence the individual's observable, measurable behaviors (Schunk et al., 2014). Furthermore, a primary assumption of the EEVT is that individuals' beliefs about the following questions can explain their achievement behaviors: 1) Can I do this task? and 2) Do I want to do this task and why? (Wentzel & Brophy, 2014). According to the EEVT, expectancy and value beliefs would be hypothesized to directly influence effort on achievement-related tasks (Eccles & Wang, 2012; Wigfield & Eccles, 1992, 2000). Thus, the EEVT suggests individuals would exert more effort toward initiating and completing achievement-related tasks if they believe they can succeed and that succeeding will result in a desirable outcome. Indeed, expectancy and value beliefs predict student effort, persistence, and achievement outcomes in academic and recreational activities (Bong, Cho, Ahn, & Kim, 2012; Meece, Wigfield, & Eccles, 1990; Wigfield et al., 1997).

Eccles, Wigfield, Harold, and Blumenfeld (1993) described expectancies for success as subjective evaluations about whether one can perform a task successfully, whereas task value beliefs refer to subjective evaluations about whether one has a personally meaningful reason to engage in the task. Expectancy beliefs have commonly been differentiated from more general academic ability beliefs, with expectancy beliefs referring to task-specific beliefs as opposed to representing more global perceptions of self-competence (Schunk & Pajares, 2009).

Subjective value beliefs can also be general or task-specific (Higgins, 2007), and there are at least four types of task value beliefs (Conley, 2012; Wigfield & Eccles, 2000). First, *attainment value* (or *importance*) is the perceived importance of a task based on how it allows one to express an important aspect of one's self-identity. Second, *intrinsic value* (or *interest*) is enjoyment or interest in a task. Third, *utility value* refers to perceptions of a task's usefulness based on how it aligns with or advances one's future aspirations or goals. Finally, *relative cost* is a dimension of task value belief representing perceptions of the alternative opportunities that are forfeited when engaging in the activity (Eccles-Parsons et al., 1983; Wigfield & Eccles, 2000).

In the context of academic test-taking, the EEVT can serve as a useful theoretical basis for explaining how student perceptions of test-taking might relate to their subsequent TTE on the test. Generally, the EEVT model would imply that the effort students exert toward responding to a test item would be most proximally determined by their subjective beliefs about the following questions: 1) Can I respond to this test item successfully? and 2) Do I have a meaningful reason to try to respond to this test item successfully? (Eklöf, 2010). In the current study, the EEVT model was applied to the domain of test-taking to help inform an empirical investigation of particular relationships of interest between individual student characteristics, their expectancy and value beliefs related to test-taking, and the TTE they demonstrate while taking the test.

Demands-capacity model of test-taking effort. Over the past two decades, researchers have attempted to identify variables that might be associated with test examinees' test-taking motivation and their subsequent TTE. This research has primarily focused on three categories of variables that have been hypothesized to influence examinee TTE: 1) characteristics of the test (e.g., format, content, or item features), 2) characteristics of the individual completing the test (e.g., demographic, psychosocial, or motivational variables), and 3) the context in which testing occurs (e.g., purpose of testing, test setting, or consequences associated with test performance). Based on this research, assessment researchers have suggested that a "test event" (i.e., one completion of a test by one student) can be conceptualized as a series of interactions between examinee-level variables and item-level variables, each of which occurs within a particular assessment context (Wise & Cotten, 2009; Wise & Smith, 2011; Wise, 2015). Wise and Smith (2011) proposed a conceptual model of TTE in which the TTE exhibited by an examinee on a test item is "influenced by the dynamic interplay" among these factors (p. 147).

In Wise and Smith's (2011) demands–capacity model (see Figure 1), the TTE a student exerts toward responding correctly to test items is regarded as a function of two primary model constructs: 1) resource demands (RD), and 2) effort capacity (EC). More specifically, RD is considered "an item characteristic representing the effort that must be expended by an examinee to correctly answer the item," whereas EC is said to represent "the amount of effort the examinee is willing to devote to answering test items" (p. 147). According to Wise and Smith, RD is considered a fixed, item-level variable based on characteristics of the item (e.g., item length), whereas EC is considered a dynamic, examinee-level characteristic based on motivational differences (e.g., confidence to answer items) that can change during a single test session.

The demands–capacity model implies that when an examinee's internal EC exceeds the RD of an item (at the time of the examinee-item encounter), the examinee is expected to give an effortful response (i.e., solution-focused behavior; SB); conversely, when the item's RD exceeds the student's momentary EC, the examinee is expected to exhibit low TTE by 1) omitting the item or 2) guessing quickly (i.e., rapid-guessing behavior; RGB). With that said, it must be noted that a third possibility would be for students to respond *effortlessly* but *not rapidly*, which would thereby not constitute RGB. Therefore, this model can account for only two of three possible response processes (i.e., the model explains SB and RGB but not non-rapid guessing). Further, as Wise and Smith (2011) acknowledged, "There are no methods currently available to quantify EC and RD on a common scale, which precludes a literal comparison of their values" (p. 150). As such, it is currently unclear how the concept described as EC in the Wise–Smith model might relate to other constructs in the research literature on motivation and student engagement.

Still, the primary strength of the demands-capacity model is that it provides researchers a relatively parsimonious framework for explaining the variables that research and theory suggest may be significant correlates of student TTE. In this model, TTE is assumed to be multiply determined by numerous test characteristics, individual differences, and test context factors. The empirical research support for the components of the model is described later in this chapter.

Another useful aspect of the demands–capacity model is that it conceptualizes TTE as a dynamic, changing construct that can vary between students or within a single student across the duration of a testing session. As such, the demands–capacity model accounts for the examinee's initial motivation upon beginning the test, as well as changes in the student's motivation during the test. This notion is helpful because TTE is understood best as fluid, with test examinees often showing variation in effort on different items, subtests, or tests within assessment batteries.

Finally, the relationships between the test-level variables, test context variables, and student-level variables proposed in the Wise–Smith model could be tested through empirical research on student TTE. Indeed, several scholars have begun testing some of the hypothesized relationships among variables in the demands–capacity model, and the model serves as a useful framework for identifying which of the potential correlates of TTE warrant additional research.

Conceptual framework of current study. Given the potential value of employing the demands–capacity model for informing research on student TTE in a variety of testing contexts, the current study aims to build on the previous work of Wise and colleagues, using the EEVT as the theoretical basis for an empirical study of the correlates of low TTE on a low-stakes CAT. The demands–capacity model assumes that TTE is multiply influenced by the dynamic relationships among 1) the characteristics of the test, 2) the characteristics of the individual test examinee, and 3) the context in which testing occurs, but an investigation of all these factors would far exceed the scope and purpose of the current study. For this reason, the current investigation focused only on the *student-level* correlates of TTE.

The conceptual framework for the current study (see Figure 2) represents an application of the EEVT in the context of academic test-taking. The EEVT implies that achievement-related choices, effort, and persistence (which in the context of test-taking would be a student's TTE) are influenced most directly by an examinee's expectations for success on the test and perceived value of the test. The relationships of interest in the current research study are considered factors comprising the student's EC in the demands–capacity model of TTE. Wise and Smith (2011) had proposed that examinees possess several individual "internal factors" (i.e., expectations about test demands; desire to please teachers or parents; citizenship; competitiveness; ego satisfaction) that would be hypothesized to be determinants of their EC prior to engaging in the test (p. 149).

In summary, the contemporary EEVT is a useful theoretical framework for explaining student engagement in academic tasks in terms of their expectancies for success on the task and the extent to which they value engaging in the task (Wigfield et al., 2009). By contextualizing the current investigation of student-level correlates of TTE within the EEVT, this study draws from a robust theoretical framework of motivation and engagement. The EEVT would suggest students' perceptions of test-taking (i.e., expectancy beliefs and value beliefs) should directly influence their subsequent test-taking engagement, which can be indicated by their TTE. This model was used to inform an empirical study of the relationships between student-level variables (i.e., grade, gender, race/ethnicity, test-taking expectancy beliefs, and test-taking value beliefs) and student TTE. In doing so, the current study extends our understanding of the demographic characteristics and internal motivational variables associated with low TTE, which may help to eventually inform the development of practices and policies aimed at addressing low TTE.

Test-Taking Effort

Issues resulting from low test-taking effort. Despite the substantial variability in how scholars have defined TTE in previous empirical studies, the research on test-taking motivation and effort has consistently suggested that lower TTE is associated with poorer test performance (Haladyna & Downing, 2004; Sundre & Kitsantas, 2004; Wise & DeMars, 2005). Wise and DeMars (2005) conducted a meta-analysis of empirical research on the relationships between TTE and test scores, comparing the magnitude of the differences in test performance between groups of high-effort and low-effort test-takers. The authors found that results from 24 of the 25 reviewed studies indicated significant, positive effects of TTE on test performance. The effect sizes (ES) ranged from –0.04 to 1.49 (mean ES $g = 0.59$), suggesting that high-effort test-takers scored more than one-half of a standard deviation higher than low-effort examinees on average.

Such findings suggest TTE may contribute substantial construct-irrelevant variance to test scores. Furthermore, high rates of non-effortful responding can have spurious effects on the reliability and validity of aggregate test datasets (Wise, 2015). Several studies have shown that test scores from examinees with low TTE can significantly influence the reliability coefficients for the test (Sundre & Wise, 2003; Wise & DeMars, 2009; Wise & Kong, 2005) as well as the correlations between test scores and external criteria (Wise, 2009).

In a seminal article on TTE, Wise and Kong (2005) hypothesized that removing the test scores of any examinees identified as exhibiting low TTE would result in a set of remaining scores that more accurately represents the actual knowledge and skills of students in the sample. Their argument assumed that these scores had been invalidated based on the evidence from the examinee's response processes, and therefore the non-effortful results should not be considered valid indicators of those students' true ability. Wise and Kong postulated that removing scores from low-effort respondents would result in a) higher test scores, b) equal or greater test score reliability, and c) increased external validity (i.e., higher correlations with scores from measures expected to relate to test performance). The results of Wise and Kong's (2005) study supported their hypotheses, and their findings have been replicated in numerous studies using comparable methods (e.g., Kong, Wise, Harmes, & Yang, 2006; Swerdzewski et al., 2011). This process of removing scores from students flagged for low TTE is commonly called *motivational score filtering*, and several researchers have evaluated its utility as a post hoc statistical method for "data cleaning" (Sundre & Wise, 2003; Wise, 2015). Motivational filtering might be a reasonable solution for addressing problems related to low TTE, and some scholars have advocated for the use of this approach as a way to improve the validity of educational measurement (Wise, 2009).

On the other hand, it is possible excluding the scores of low-effort responders from aggregated data reports could *diminish* the validity of resulting test interpretations if the subset of examinees retained is not representative of the students for whom the data are used. Currently, it is unclear whether score filtering tends to differentially remove test scores from any subgroups who are more likely to exhibit low TTE. Because of these possible unintended consequences of TTE-based filtering approaches are still unknown, research should address whether any groups are especially likely to exhibit low TTE. If students vary in TTE by demographic group, it might suggest score filtering has the potential to exclude scores from some subgroups. Systematically removing the scores of students from certain groups could lead to educational decision-making less informed by their underlying academic needs. For this reason, it is essential for researchers to address whether various subgroups of students may differ in the odds they exhibit low TTE. Furthermore, if educators could take preventative steps to improve TTE, it is possible statistical adjustments like score filtering would not be necessary. Rather than removing the test scores of students with low TTE, it would be more appropriate to find ways to increase student test-taking motivation in order to promote more effortful responses to test items. In sum, identifying groups of students who are most likely to disengage and malleable predictors of disengagement could inform the development of targeted strategies for promoting TTE in low-stakes contexts.

In addition to better understanding the potential consequences of low TTE for any interpretations made about *aggregate* test data, another critical issue concerns the effects of low TTE on the validity of interpretations about *individual* scores. If a student has low test-taking motivation, it might be manifested as low TTE by the student refusing to comply with the testing procedures, omitting answers, leaving entire sections of the test blank, cheating, guessing, or responding in a non-effortful pattern (Haladyna & Downing, 2004; Wise, 2015).

These types of unexpected response processes are problematic for test interpretation and use, as they make it difficult to ascertain whether an incorrect response indicated a) the examinee *did not* have the target skill being measured, b) the examinee *did* have the target skill but made an error, c) the examinee had *partial* knowledge but made an incorrect educated guess, or d) the examinee *might have* had the target skill but did not exert enough effort to give a meaningful answer. Test users respond to this issue by evaluating whether evidence suggests the student's score has been invalidated due to construct-irrelevant variance associated with low TTE. One recommendation for identifying potentially invalid scores is a process Wise (2015) termed the *individual score validation* (ISV) approach, characterized by the following five-step procedure:

> First, the test user identifies the construct-irrelevant factors that represent the greatest validity threat. Second, suitable indicators of each construct-irrelevant factor are chosen or developed…The third step is to establish criteria for classifying test scores as invalid, by defining procedural rules for invalidating scores. The fourth step is to apply the indicators and criteria to the set of test events in question and identify those test events whose scores are invalidated. Once invalid test scores have been identified, a final step is to decide what course(s) of action to take. (p. 246)

Wise's (2015) ISV approach gives researchers and practitioners a systematic method for evaluating whether low TTE might have invalidated a particular student's test score. To use this method, test users must select an appropriate indicator of TTE to "flag" students with low TTE. Because there is no single criterion for measuring "adequate TTE," one must define *low TTE* and provide evidence that the validity flag can reliably classify effortful and non-effortful responders. Therefore, an important issue pertinent to research on student TTE concerns how TTE should be measured. In the following section, commonly used methods for measuring TTE are discussed.

Issues in measuring test-taking effort. In a volume on motivation in school, Schunk and colleagues (2014) stated that effort is considered a latent variable measured through a) direct observations of behavior or b) self-report via questionnaires, interviews, or think-aloud. Most researchers have measured TTE using brief self-report measures, which are typically Likert-type questionnaires that yield general, global estimates of students' perceived test-taking motivation before testing or their post-test perceptions of their TTE (Wise & DeMars, 2006). One commonly used measure of test-taking motivation and TTE is the Student Opinion Scale (SOS; Sundre, 1999), comprised of subscales measuring perceptions of Effort (e.g., "I gave my best effort on this test.") and Importance (e.g., "Doing well on this test was important to me.").

Most of the research on TTE is based on self-report, but scholars have raised several unique concerns about the conclusions that can be drawn based on the TTE students report after testing. First, self-reports only yield a general estimate of self-perceived TTE, as opposed to providing more objective data about student TTE for a particular section of a test or on specific items. Second, although they are useful as global estimates of TTE, self-report measures cannot provide data about changes in students' TTE that might occur during a single testing session. Third, researchers have questioned how truthfully or reliably students respond to self-report measures of TTE (Wise & Kong, 2005). Indeed, previous research from the attribution theory literature suggests that some individuals might tend to attribute their poor performance to a lack of effort to help themselves preserve a positive self-concept (Weiner, 1992). Lastly, it is unlikely that examinees (especially younger students) could accurately report the proportion of items on which they guessed, which suggests self-report is not useful for estimating the total prevalence of low-effort responses. Overall, there is inconclusive evidence about the extent to which self-report of TTE might be adequately reliable, valid, and useful for informing educational decisions.

For these reasons, several researchers have indicated the need for more objective methods for quantifying TTE. One alternative to self-report questionnaires offered by some measurement experts is the use of person-fit statistics, which compare examinees' responses to a theoretical model for the test. According to Meijer (2003), "Person-fit statistics have been proposed that can be used to investigate whether a person answers the items according to the underlying construct the test measures or whether other answering mechanisms apply… Most statistics are formulated in the context of item response theory (IRT) models…and are sensitive to the fit of an individual score pattern to a particular IRT model" (72). As such, the person-fit approach enables test users to evaluate whether responses are improbable given the estimated item and examinee parameters. That is, when item responses deviate substantially from the estimated IRT model, the observed aberrant responses can be interpreted as likely due to guessing rather than effortful responding.

However, some experts have asserted that this assumption may not always be correct. As Lord and Novick (1968) argued, identifying all responses that deviate from expected IRT models as "random" guesses would be inconsistent with our understanding of test-taking behavior. For instance, examinees who have partial knowledge about an item might be able to eliminate clearly incorrect responses and answer with the most reasonable remaining option. Others could have misconceptions about items and thereby respond in a way that does not fit the IRT model due to their misunderstanding. Thus, it may be problematic to interpret every item flagged as violating the person-fit model as a completely random guess; however, most of the indices for flagging non-model-fitting responses treat all aberrant responses as guesses (Weiss, 1983). In summary, both self-report and IRT-based person-fit methods for measuring TTE have notable limitations, and neither method identifies non-effortful item responses with complete precision. In response to this issue, Wise and Kong (2005) proposed an alternative method for measuring student TTE.

Response time effort. Over the past decade, educational measurement experts have suggested item response latency could be used as a more reliable and practical indicator of TTE (Wise, 2014), citing an observation by Schnipke and Scrams (1997) that some test-takers tend to respond very quickly when they approach the end of timed tests. This finding had suggested response times might be used to differentiate between effortful and non-effortful item responses. Wise and Kong (2005) applied this idea to develop an innovative approach for measuring TTE on CBTs by analyzing item response times.

Specifically, Wise and Kong's RTE method is used to classify responses as effortful or non-effortful using Schnipke and Scrams' (1997) dichotomous distinction between SB and RGB by comparing response times to a prespecified threshold (for a review of research on response time thresholds, see Kong, Wise, & Bhola, 2007). Researchers derive RTE for a test event (i.e., one completion of a test by an individual student) by summing the number of items classified as SB and dividing that number by the total number of items. Hence, researchers can calculate the student's RTE score—indicating the proportion of responses classified as SB—a value ranging from 0.0 (no responses classified as SB) to 1.0 (all responses classified as SB).

Wise and Kong (2005) listed several reasons why response time data may be preferable for research on student TTE. First, collecting item response times on CBTs is unobtrusive, as researchers can measure latency without examinees realizing that the data are being recorded. Second, this indicator of TTE is not based on subjective judgments of effort; instead, scores represent examinees' observed response behaviors during testing. Third, response time data can be collected for each item, which allows for the analysis of changes in TTE across different test items. A final practical advantage is that RTE can be used to measure TTE on a CAT without requiring the IRT-based item parameters for each of the items in a large CAT item bank.

In Wise and Kong's (2005) seminal RTE study, they proposed five hypotheses about the validity of RTE as an indicator of student TTE, and a decade of subsequent research on RTE in higher education contexts has provided support for the following hypotheses (Wise, 2015).

1. RTE scores should demonstrate adequate levels of reliability.

The first hypothesis was supported by the results of Wise and Kong's (2005) study, as the observed coefficient alpha for RTE scores was 0.97 (indicating high internal consistency). Other RTE research studies in higher education contexts have consistently corroborated these findings, with the observed internal consistency reliability coefficients for the RTE index ranging from 0.81 (Wise & DeMars, 2010) to 0.99 (Kong et al., 2007; Wise & DeMars, 2006) and exceeding 0.90 in several subsequent research studies (e.g., Kong et al., 2006; Wise et al., 2006).

2. RTE scores should be correlated with other measures of examinee test-taking effort.

The second hypothesis was that RTE should demonstrate concurrent validity. Wise and Kong's (2005) findings supported this hypothesis, as RTE scores were significantly correlated with both self-reported TTE ($r = 0.25$) measured by the Effort subscale of the SOS and person-fit estimates ($r = –0.42$) measured by the Modified Caution Index (Harnisch & Linn, 1981). Other researchers have found similar associations between RTE and self-reported TTE, with the observed validity correlations ranging from 0.38 (Kong et al., 2007) to 0.61 (Rios et al., 2014).

3. RTE scores should not be correlated with measures of academic ability.

Wise and Kong's (2005) third hypothesis pertained to evidence for discriminant validity, as the authors hypothesized RTE scores should not be related to students' academic ability. To test this hypothesis, the authors analyzed relationships between student RTE scores and previous scores on the Scholastic Assessment Test (SAT) and found no significant correlations between RTE and SAT–Verbal ($r = 0.06$) or SAT–Quantitative ($r = –0.02$) scores.

Other scholars have found similar results (Rios et al., 2014; Wise & DeMars, 2010), suggesting academic ability does not appear to be a correlate of low TTE on low-stakes tests. These results could suggest that the students' *low test-taking motivation* (rather than *low academic skills*) might have been a reasonable explanation for their low TTE.

4. Instances of rapid-guessing behavior should yield item scores that are correct at a rate consistent with chance.

Research has also supported Wise and Kong's (2005) fourth hypothesis that RGB should have accuracy rates comparable to chance guessing. In previous studies, scholars have found the accuracy of responses classified as RGB have not significantly exceeded the accuracy expected by random guessing, whereas those classified as SB have significantly exceeded chance. For instance, Wise (2006) found 25.5% of items identified as RGB were correct, but 72.0% of item responses identified as SB were correct; relatedly, Setzer and colleagues (2013) analyzed over one million item responses and found the accuracy rates were 27.9% for RGB and 51.7% for SB.

5. RTE scores should show motivation filtering effects similar to those found with other measures of examinee effort. (p. 174–175)

Lastly, several empirical studies have supported Wise and Kong's (2005) fifth hypothesis that RTE scores should be comparable to other measures of TTE for motivational score filtering (Rios et al., 2014; Swerdzewski et al., 2011; Wise & Cotten, 2009, Wise & Kong, 2005). Results from these studies have indicated filtering the test scores of students identified as exhibiting low TTE (as determined by RTE score), average test scores increase, test score standard deviations decrease, and the magnitudes of correlations with external variables—indicative of convergent validity—increase (Kong et al., 2007; Wise, 2015; Wise & DeMars, 2010; Wise & Kong, 2005). In the next section, empirical research on Wise and Kong's (2005) RTE approach is described.

Empirical Research on Student Test-Taking Effort

Prevalence of low test-taking effort. Empirical studies using Wise and Kong's (2005) RTE index have consistently shown that a small proportion of test examinees exhibit low TTE when taking a CBT. However, the results from these studies have varied considerably in the observed prevalence rates of low TTE. As shown in Table 1, existing research studies in college samples have indicated that the proportions of students identified as exhibiting low TTE (commonly defined as RTE scores below 0.90) have ranged from 0.6% (Wise & DeMars, 2010) to as high as 35.6% (Swerdzewski et al., 2011). Altogether, the observed rates of low TTE in these studies are alarming, given that the results from numerous studies have indicated that more than 10% of students exhibited low TTE (i.e., RGB on at least 10% of items) when taking educational CBTs (Kong et al., 2007; Wise et al., 2006; Wise & DeMars, 2006; Wise, Pastor, & Kong, 2009; Wise & DeMars, 2010; Swerdzewski et al., 2011; Rios et al., 2014).

Even though there have been fewer studies on the RTE scores of school-age students on low-stakes educational tests, there is some research evidence to suggest that RGB does indeed occur in K–12 contexts. As displayed in Table 1, the proportion of examinees flagged as having low TTE were reported in four empirical studies, with these observed proportions ranging from 1.1% (Wise et al., 2004) to 11.9% (Wise, 2015). Of the research in school-age samples, two studies in particular (Wise et al., 2010; Wise & Ma, 2012) warrant more detailed discussions in the current literature review. Both of these large-scale studies by Wise and colleagues used data collected from students who used the Northwest Evaluation Association (NWEA)'s *Measures of Academic Progress* (MAP) assessment, a multiple-choice CAT system with mathematics and reading comprehension tests. The results of these studies provide strong support for the notion that widespread RGB may pose a significant threat to the validity of commonly used CATs.

Wise and colleagues (2010) investigated the relationship between TTE and the time the testing occurred. In a large-scale secondary data analysis, researchers analyzed data from all of the students in grades 3–9 who used the MAP math (n = 355,116) or reading (n = 356,715) CAT. The authors found that average RTE scores for these test events decreased as the time of day was later. For instance, results indicated the mean RTE score for MAP reading tests taken at 7:00 a.m. was 0.994, whereas the mean RTE score for reading tests taken at 2:00 p.m. was 0.974.

In another related study, Wise and Ma (2012) analyzed MAP reading and math data to calculate RTE for students in grades 3–9. Comparing several methods for RTE time thresholds, the authors found that even the most conservative approach (a common three-second threshold) flagged a considerable proportion of students as having low TTE. In math, 4.8% of the 286,150 analyzed test events were flagged as invalid, whereas 10.6% of the 287,843 analyzed reading test events were flagged as invalid. These estimates exceeded those by Wise and colleagues (2010).

Although the researchers did not perform comparative or correlational analyses to test the significance of group differences in RTE, results from these two studies still have important implications for future research in this area. First, researchers conducted population analyses and thus observed the true mean RTE scores. To date, no other scholars had reported the mean RTE scores for K–12 students disaggregated by grade level and gender. Next, given their large sample sizes, these studies provide clear evidence that RGB does indeed occur when K–12 students use CATs at school. Lastly, the tests in these studies are used widely in schools for measuring math and reading skills. These types of assessments can be considered low-stakes from the perspective of the examinees because they carry no (or minimal) consequences for the students, but they are used to inform critical educational decisions. The high rates of RGB in these studies support the need for additional research on the nature of TTE in these types of assessment conditions.

Support for demands–capacity model. As previously stated, Wise and Smith's (2011) demands–capacity model assumes TTE is influenced by three types of factors: *item-level variables*, *examinee-level variables*, and *testing context variables*. Although it is only a preliminary model for explaining why RGB occurs, the Wise–Smith model is a useful framework for describing what scholars have learned so far about the correlates of TTE based on evidence from the extant research literature. In several studies, researchers have investigated *item-level correlates* of TTE by testing the relationships between item characteristics and the TTE students exhibited on the items. In these studies, researchers used Response Time Fidelity (RTF), an index analogous to RTE representing the rate of RGB exhibited across all examinees for a given item (as opposed to the RGB exhibited across all items for a given examinee).

The results of previous studies have consistently shown three primary characteristics are strongly associated with student TTE: item position in the test, item length, and the presence of additional reading materials (Wise, 2006; Wise et al., 2009; Setzer et al., 2013). By contrast, item difficulty has not been found to predict RTF (Wise, 2006; Setzer et al., 2013). The lack of a significant association between item difficulty and the likelihood students exhibit RGB was unexpected because the EEVT seems to imply students would have lower expectancies for success on difficult items compared to easy items (and therefore exhibit less effort). In fact, the research has suggested students who exhibit low TTE tend to respond so rapidly that they do not accurately judge the difficulty of the item; instead, they make quick decisions about the amount of "mental taxation" the item will likely cause them (Wolf, Smith, & Birnbaum, 1995). These findings seem to suggest students do have a limited amount of "mental energy" they can exert during a testing session and that the perceived RD of a test item does relate to TTE on the item.

Further empirical support for the Wise–Smith model is derived from research on the effects of the *testing context* on student TTE. Most early research studies in this area focused on one specific contextual variable: the consequences of test performance. Research has shown that offering a reward of one dollar for performance on a low-stakes test increased the performance of eighth-grade students (O'Neil, Sugrue, & Baker, 1995). Other studies focused on differences in TTE between graded and ungraded tests. Wolf and Smith (1995) found that college students reported higher levels of TTE on tests they were told would count toward a grade, though Smith and Smith (2002) later found that TTE and test performance did not improve in consequential conditions for students with high test anxiety. Collectively, these studies have provided some support for *testing context* factors as correlates of TTE in the demands–capacity model of TTE.

In contrast to the extant research literature on the item-level correlates and context-level correlates of TTE, there are relatively few studies on the *student-level correlates* of TTE. In the demands–capacity model, Wise and Smith (2011) proposed that the following "internal factors" may be determinants of students' EC when they begin a test: level of proficiency, amount of test preparation, expectations regarding test demands, desire to please teachers, parents, and others, citizenship, competitiveness, and ego satisfaction (p. 149). Even though there is evidence that student expectations regarding test demands do indeed relate to the amount of TTE students exert (Wise, 2006; Wolf et al., 1995), few researchers to date have empirically tested the hypothesized relationships between student "internal factors" and TTE in the Wise–Smith model. This notable gap in the extant research on TTE points to a particularly important area for further research. As Wise (2015) concluded, "research should be directed toward better understanding the dynamics of test-taking motivation" (p. 250). The next section of this literature review describes what we currently know about the psychological and motivational factors that may relate to student TTE.

Student test-taking beliefs and test-taking effort. According to the EEVT, expectancy beliefs and value beliefs presumed to be the most proximal predictors of achievement-related behaviors, which would suggest student perceptions of test-taking should be expected to relate to their TTE directly. Indeed, there is research evidence to suggest student beliefs about the value of test performance is a correlate of their TTE. Those who perceive a test as affecting their grades (and thus having higher utility value) report higher levels of test-taking motivation and outperform students who are told a test will not be graded (Wolf & Smith, 1995). Even in the absence of test consequences, students are more likely to omit or respond incorrectly to test items they *perceive* as having greater mental taxation (i.e., higher relative cost), even if the items are not more difficult (Wolf et al., 1995). In one study, students varied in their perceptions of how relevant a test would be to their future employment (i.e., utility value), and these group differences in task value beliefs partially explained the variation in self-reported TTE and test performance (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997). Cole, Bergin, and Whittaker (2008) analyzed an expectancy–value model of test-taking by measuring the TTE of 1,005 college students who took a low-stakes exam (used for institutional evaluation purposes). Mediational analyses indicated students' test-taking value beliefs predicted their self-reported TTE, which in turn predicted their test performance. Together, these findings provide support for the proposition that student beliefs about testing may be individual-level correlates of TTE.

Furthermore, one of the only studies to directly test the student-level correlates of TTE (using the RTE method) was Wise and Cotten's (2009) investigation of a low-stakes exam taken by 802 college students. The purpose of the study was to test whether TTE was associated with "attitudinal and affective determinants" (p. 192), as suggested by the demands–capacity model. The researchers provided the following rationale for examining student beliefs about testing:

> The Wise–Smith model of test-taking effort would consider student conceptions of assessment as important components of the internal motivational factors that contribute to effort capacity. In a low-stakes testing situation, in which the dominant motivational factors of test consequences are absent, the model would predict that test-taking effort should be related to the conceptions of assessment that the student brings into the testing session. (Wise & Cotten, 2009, p. 194)

To empirically test this proposition, researchers gathered survey data from students about their perceptions of test-taking using four subscales from the Student Conceptions of Assessment Scale (SCoA; Brown, Irving, Peterson, & Hirschfeld, 2009): Improvement (measuring attitudes toward the use of the testing for academic improvement), Affect (measuring how positively one feels about testing), Irrelevant (measuring the degree to which one believes testing is irrelevant to learning), and Accountability (measuring the degree to which one believes testing is important for holding students and schools accountable). Using hierarchical multiple regression analyses, Wise and Cotten (2009) analyzed the extent to which scores on the four subscales of test-taking perceptions derived from the SCoA predicted student TTE, controlling for gender and SAT scores. Descriptive statistics for the model indicated that RTE was significantly associated with three of the four SCoA subscales (Improvement, Irrelevant, and Accountability), although the block of four SCoA predictors explained just 6% of the variance in RTE scores. Further, only the SCoA Improvement and Affect scales had statistically significant regression weights, with higher Improvement scores associated with an increase in RTE score and higher Affect scores associated with a decrease in RTE score. Based on these findings, the authors concluded that "student conceptions of assessment were clearly related to test-taking effort" (p. 200).

Still, Wise and Cotten's (2009) study had several limitations that may weaken these conclusions. First, the authors did not provide a clear rationale for the use of the SCoA or clear conceptual definitions of the constructs its four subscales were intended to measure. As such, the practical significance of the observed relationships between RTE scores and the Improvement and Affect predictor variables remains unclear. Although the subscales in the SCoA may share some conceptual overlap with the four task value beliefs in the contemporary EEVT—items in the Improvement and Irrelevant subscales seem to measure facets of utility value, items in the Affect subscale seem to measure intrinsic value, and items in the Accountability subscale seem to measure attainment value—the major shortcoming of this study was that it lacked a strong theoretical foundation for investigating the motivational variables associated with TTE. Further, Wise and Cotten's (2009) finding that positive affect toward testing was negatively associated with student RTE appears to be inconsistent with Cole and colleagues' (2008) finding that personal interest in testing was *positively* associated with student TTE. Altogether, Wise and Cotten's (2009) study was an influential contribution to research on the relationships between student test-taking beliefs and TTE (as proposed in the demands–capacity model), but an unclear justification for the student-level motivational variables in their investigation does limit the conclusions one should make about the psychological processes that contribute to low TTE.

In another study focused on the relationships between student test-taking beliefs and TTE, Zilberberg, Finney, Marsh, and Anderson (2014) used a measure of test perceptions that could be considered more consistent with an expectancy–value model of test-taking than the measure used by Wise and Cotten (2009). Zilberberg and colleagues examined the extent to which first-year college students' perceptions about K–12 accountability assessment systems predicted their test-taking motivation on a university-mandated academic achievement test.

Participants responded to the Students' Attitudes towards Institutional Accountability Testing in K–12 (SAIAT-K–12; Zilberberg, Anderson, Finney, & Marsh, 2013), completed quantitative and scientific reasoning tests, and responded to the SOS. The researchers tested a hypothesized fully-mediated path model and found that "students' attitudes toward institutional accountability tests in K–12 directly affect perceived importance of university accountability tests, which in turn directly affects test-taking effort, which in turn directly affects test performance" (p. 367). In other words, the results from this study suggested the relationship between students' perceptions about K–12 testing and their performance on a university test could be explained by the effects of test perceptions on TTE.

The specific perceptions that were associated with TTE and subsequent performance on the test were Purpose ("students' understanding of the purpose of such tests") and Parents ("students' parents paid attention to test scores"); by contrast, subjective beliefs about Validity ("K–12 institutional accountability tests are adequate measures of ability") and Disillusionment ("student dissatisfaction with these tests") were not significant predictors in this mediation model. Thus, Zilberberg and colleagues' (2014) findings supported Wise and Cotten's (2009) conclusion that certain beliefs about assessment are significantly associated with the TTE students exhibit in low-stakes assessment scenarios. Altogether, these studies suggest that beliefs about the purpose of an assessment (i.e., utility value or attainment value) relate to student TTE.

Summary and remaining questions. This review of empirical research on TTE has presented what we currently know and pointed toward gaps in the literature that may warrant further research. First, most students tend to exhibit (or self-report) adequate levels of TTE when completing low-stakes academic tests, but scholars have clearly documented that a small proportion of test takers disengage from testing and exhibit inappropriately low TTE.

The proportions of college students flagged with low TTE have ranged from less than 1% (e.g., Wise & DeMars, 2010) to greater than 20% of test takers (e.g., DeMars, 2007; Rios et al., 2014; Swerdzewski et al., 2011; Wise et al., 2009). However, despite this growing interest in TTE on low-stakes university-mandated assessments, relatively few studies to date have focused on the prevalence of low TTE in K–12 settings. As such, questions remain about whether elementary and secondary students show similar rates of disengagement on low-stakes tests.

Second, the Wise–Smith (2011) model of TTE has accumulated some empirical support, but the research to date has primarily focused on correlates of TTE that are considered *test-level variables* or *testing context variables* (as opposed to *student-level variables*). The effects of consequences, incentives for performance, and test item-level characteristics on TTE have been documented clearly, but less is known about the demographic characteristics or motivational patterns of students most likely to exhibit disengagement from low-stakes testing.

Finally, only a few empirical studies have addressed the extent to which motivational factors may be associated with non-effortful response processes. Moreover, these studies have not been designed in a way that explicitly draws support from a prominent theory of motivation and engagement, such as the EEVT. Further research on "internal factors" related to TTE in the Wise–Smith (2011) model is needed, as these relationships are not understood fully at this time.

Student-Level Correlates of Test-Taking Effort

Numerous studies based in the EEVT have demonstrated the existence of relationships between demographic characteristics, beliefs, and effort and persistence on achievement-related tasks. Also, the extant research has revealed several key group differences in the expectancy and value beliefs students endorse in various academic domains, and these relationships have indeed explained subgroup differences in school engagement and overall academic performance.

Although few studies have explicitly focused on relationships between student demographic characteristics and test-taking expectancy and value beliefs, there is research evidence to suggest that certain groups of students may exhibit differences in their broad expectancy or value beliefs in a given subject area (such as reading, math, or science). This section describes prior research on subgroup differences in a) TTE, b) expectancy and value beliefs in the domain of reading, and c) expectancy and value beliefs specific to test-taking.

Age/grade level. Several studies have indicated that students differ in their test-taking motivation and TTE by age or grade level, although few scholars have directly tested whether age or grade are significantly associated with the likelihood students exhibit low TTE. Still, the trend in previous empirical research is that TTE appears to decrease slightly over time, with students in higher grades reporting and exhibiting lower TTE.

For example, Wolf and colleagues (1995) studied TTE in a sample of students in grades 10 and 11, and results indicated that grade 11 students reported lower test-taking motivation than grade 10 students and consequently were more likely to omit items that were high in mental taxation. This finding by Wolf and colleagues suggests students may experience decreases in their TTE as they take tests multiple times over consecutive years; indeed, a handful of other studies have also pointed to grade-level differences in average RTE scores or the proportion of students with low TTE. Hauser and Kingsbury (2009) identified low-effort test respondents on a reading CAT in a large sample of students in grade 3 (n = 16,209) and grade 9 (n = 18, 705). The proportion of grade 9 students who were flagged as exhibiting low TTE (7.7%) was higher than the proportion flagged in grade 3 (6.9%), although researchers did not test whether there was a statistically significant difference by grade in the likelihood of exhibiting low TTE.

Relatedly, Wise and DeMars (2010) found sophomore students had a lower mean RTE score ($M = 0.943$) than freshman students ($M = 0.996$). In their study, 11% of second-year students were identified as exhibiting low TTE (defined as RTE below 0.90), whereas 0.6% of first-year students were flagged. Results from this study showed 43 of the 45 examinees (95.6%) removed from the sample through motivational score filtering were sophomores and indicated that the lower TTE of sophomores had significantly distorted the estimates of student growth.

Even more notably, Wise and colleagues (2010) demonstrated strong evidence of a decline in RTE in their analysis of test scores from students in grades 3–9. Analyzing 356,715 test events from the MAP reading CAT, they found the following mean RTE scores: 0.995 in grade 3 ($n = 138,016$), 0.994 in grade 4 ($n = 134,535$), 0.995 in grade 5 ($n = 131,818$), 0.988 in grade 6 ($n = 122,963$), 0.984 in grade 7 ($n = 121,273$), 0.984 in grade 8 ($n = 120,375$), and 0.971 in grade 9 ($n = 79,189$). Their results showed a consistent decline beginning in sixth grade.

In sum, previous RTE studies suggest TTE declines as students age, with older students generally having lower RTE. This trend is consistent with developmental declines in overall academic motivation in previous research (Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002; Wigfield et al., 1997). There is growing evidence that academic motivation decreases as students get older, and research has indicated that students' general expectancy and value beliefs tend to decline as they age (Durik et al., 2006; Wigfield & Eccles, 1994). Some researchers have even documented distinct trajectories of general motivation as students age (Archambault, Janosz, Morizot, & Pagani, 2009; Baker & Wigfield, 1999; Ratelle, Guay, Larose, & Senécal, 2004). Archambault and colleagues (2010) found changes in children's motivation in literacy using cross-sectional and longitudinal studies of students in grades 1–12, and results indicated there are several distinct motivational trajectories students might experience throughout their school years.

Relatedly, Eccles-Parsons and colleagues (1983) examined student beliefs about literacy at five ages throughout elementary, middle, and high school years. Results suggested students display as many as seven distinct trajectories in task value and reading ability beliefs over time. Three of these trajectories indicated decreasing expectancy and value beliefs in reading as students aged. Students in the "Early Decline Trajectory" group (7.8%) showed steep declines in their reported ability beliefs and literacy task value starting in second grade and continuing until ninth grade; students in the "Constant Decline Trajectory" group (28.1%) showed consistently decreasing expectancy and value beliefs from elementary to high school; lastly, students in the "Late Decline Trajectory" group (13.3%) showed decreases in literacy value beliefs that dropped more significantly in high school—although this group showed increases in ability beliefs until fifth grade, at which time expectancies began to fall dramatically. Altogether, nearly half of the students in this longitudinal study ($N = 655$) demonstrated a decrease in expectancy or value beliefs in reading. The results were consistent with previous findings that many students develop more negative expectancy and value beliefs about reading as they get older (Jacobs et al., 2002).

Given the evidence that older students tend to exhibit lower TTE than younger students and the evidence that students tend to show developmental declines in their general expectancy and value beliefs for reading tasks, it is reasonable to consider whether students may experience similar declines in their expectancy or value beliefs specific to test-taking as they get older. Indeed, Paris, Lawton, Turner, and Roth (1991) identified developmental trajectories in student perceptions of academic testing based on survey data collected from students in grades 2-11 about attitudes toward test-taking at school. In one study, researchers found significant age differences in how students perceived standardized test systems (Paris, Turner, & Lawton, 1990).

Paris and colleagues found that younger students agreed tests were useful for measuring their learning, whereas older students disagreed that academic test scores were valid to them. Older students reported testing was less important, and they disagreed that standardized testing provided useful information to their families or that the purposes of tests were explained clearly. Paris and colleagues (1991) described this trend as the development of "disillusionment" with testing, as results pointed to "a native presumption of the positive value of the test among young students, as well as increasing skepticism among older students about the importance of the test" (p. 15). Other trends in their survey results included increases in both hostile attitudes toward testing and anxiety about social comparisons based on test performance. Given that previous findings have shown decreases in overall motivation in reading, these results are reasonable evidence that students may be more likely to display inappropriately low TTE as they age.

Gender. Another consistent trend in the extant research literature is that female students exhibited higher TTE than male students in general. By contrast, male students have been identified with low TTE at disproportionately higher rates than females. For instance, Wise and colleagues (2004) found that 85.2% of students in grades 6–10 flagged as non-effortful respondents were male, even though their sample was balanced by gender. Similarly, Wise and DeMars (2010) studied a group that was 67% female; however, 25 of the 43 students (58%) with low TTE were male. In another study, Wise and Cotten (2009) directly tested whether college students differed by gender in mean RTE and found that male students ($m = 0.92$) demonstrated lower RTE than females ($m = 0.96$), with an ES of 0.18. Male college students showed poorer RTE scores on science and business tests (DeMars, Bashkov, & Socha, 2013), and similar gender differences were found in Sweden (Eklöf, 2007) and New Zealand (Brown & Hirschfeld, 2008).

Other researchers have identified distinct gender differences in expectancies for success and value beliefs at school. According to the EEVT, gender differences in ability beliefs and the valuing of certain subjects may arise from gender stereotypes that are communicated to children as they progress through school (Eccles, Wigfield, & Schiefele, 1998). Indeed, numerous studies suggest boys report higher expectancy beliefs for activities in male-stereotyped domains like math and sports, whereas girls report higher expectancies for activities in female-stereotyped domains, such as reading and social activities (Eccles et al., 1989; Eccles et al., 1993; Jacobs et al., 2002; Robinson & Lubienski, 2011; Watt, 2004). Still, it is unclear when and why these differences may develop and whether age might moderate the magnitude of these differences. Another critical issue requiring additional research pertains to how social, cultural, familial, or school environment variables might contribute to gender differences in expectancy beliefs at school (Bailey, 1993; Meece, Glienke, & Askew, 2009).

Relatedly, researchers have also examined how subjective task value beliefs may differ between male and female students. Once again, the research has yielded mixed findings concerning the possible presence of gender differences in subjective task value beliefs across different ages. For instance, Wigfield and Eccles (1992) reported that boys valued math more than girls, whereas girls valued English more than boys. However, this pattern was not consistent across age groups. By contrast, Wigfield and Eccles (2002) found no gender differences in task value for math or computer activities but found that girls valued reading and music more than boys. Together, these findings suggest the emergence of gender differences in value beliefs is not fully understood, and additional research is needed to determine how these differences might be influenced by other student characteristics or environmental variables in the school context.

Little is known about whether gender differences may exist in test-taking perceptions. Still, given the evidence that female students tend to exhibit higher TTE than male students and the evidence that female students tend to endorse greater expectancy and value beliefs for reading tasks, gender differences might arise in test-taking expectancy or value beliefs. In one study, Cole and colleagues (2008) found significant relationships between gender and student perceptions of test usefulness, perceptions of test importance, and self-reported TTE, with male students showing lower test-taking motivation and TTE across four subject areas. However, it is unclear whether gender differences in test-taking value beliefs are present in younger students.

Race/ethnicity. Compared to the research on age or gender, few scholars have examined whether student race or ethnicity might be related to TTE. As such, it is unclear whether students from different racial or ethnic minority groups may hold different test-taking beliefs or exhibit different levels of TTE. However, there is research evidence to suggest that students from racial and ethnic minority groups may be at-risk for maladaptive patterns of academic motivation in general (Graham & Taylor, 2002), and so it could be helpful to know whether such students are at-risk for disengagement from test-taking. It should be noted that scholars have argued that previous comparisons by race or ethnicity might have been confounded by socioeconomic status (SES) if researchers had not controlled for this variable (Graham, 1994; Pollard, 1993). Therefore, research on differences in motivation or engagement by race or ethnicity should be interpreted with the recognition that the interacting influences among multiple demographic variables on motivation are still not well understood. Although there is little empirical research that addresses relationships between race or ethnicity and TTE in low-stakes testing contexts, a few studies have examined associations between race or ethnicity and test-taking beliefs or TTE.

In one study, Chan and colleagues (1997) investigated the performance of 210 college students on an assessment of cognitive problem-solving skills (with no consequences), as well as student perceptions about the validity of the tests and their test-taking motivation. Black students reported lower motivation for testing, and lower performance on two parallel forms of the test compared to White students, and mediational analyses indicated the relationship between race and test performance was mediated partially by differences by race in perceptions of test validity. Based on these findings, the authors concluded that part of the achievement gap between Black and White students could be explained by group differences in perceptions about whether the test adequately measured their knowledge and skills, which in turn predicted variation in test-taking motivation. In another study, Brown and Hirschfeld (2008) found ethnic minority status related to student perceptions of assessment and test performance in a group of students in New Zealand.

Notably, a recent investigation of RTE by Setzer and colleagues (2013) is one of the only studies to date in which researchers a) reported descriptive statistics for the sample disaggregated by demographic subgroups and b) tested the statistical significance of subgroup differences in RTE. Comparing the RTE scores of non-White (n = 1,646) and White (n = 6,436) students, Setzer and colleagues found a significant difference between the average RTE scores of White (m = .990) and non-White (m = .980) students (with an ES of 0.10).

Taken together, little is known about whether similar effects may be found when analyzing the test responses of K–12 students, and so one intent of the current investigation is to add to the extant research on the demographic characteristics of students who may be most likely to exhibit low levels of TTE in low-stakes testing contexts. With that being said, the present analysis of the potential associations between student race/ethnicity and student TTE is primarily exploratory in nature, given there has been little empirical research on this topic to date.

Disability status. The final student-level demographic variable of interest in the current study is student learning disability (LD) status. This variable in the proposed model of TTE is also exploratory, given that few researchers to date have reported data about students with disabilities (SWD) in empirical studies of test-taking motivation or TTE. Still, previous research on motivation and engagement has suggested students with an LD report lower self-efficacy, make maladaptive attributions about their abilities, and show less effort and persistence on difficult tasks (Battle, 1979; Butkowsky & Willows, 1980; Chapman & Boersma, 1979; Licht & Kistner, 1986). For instance, a report from the National Joint Committee on Learning Disabilities (NJCLD, 2008) suggested that students with an LD have difficulties maintaining sufficient motivation in school. Wiest, Wong, Cervantes, Craik, and Kreil (2001) compared the self-reported intrinsic motivation of secondary students in general education, special education, and alternative education settings, and the results suggested that students in special education settings reported lower perceptions of personal competence than students in regular education settings.

These findings were consistent with previous research on elementary students with an LD. For instance, Grolnick and Ryan (1990) had documented that children with an LD tended to report feeling lower levels of cognitive competence, were more likely to report that academic outcomes were out of their control and had lower levels of motivation (compared to a control group) per teacher report. To summarize, the limited knowledge about this issue points to the potential importance of further research on the motivation and engagement of SWD. Still, there is some evidence that could suggest SWD may have low expectancy and value beliefs compared to their typically developing peers. A better understanding of whether group differences may exist in these alterable motivational variables would be necessary for informing strategies for addressing the unique motivational needs of different students.

Current Study and Research Questions

Remaining gaps in the research literature. This review of research on student TTE in educational contexts revealed three significant gaps in the research literature that warrant further investigation. First, previous studies of TTE (using multiple methods for measuring TTE) have not provided conclusive estimates of the prevalence of low TTE on low-stakes tests. Even less is known about the occurrence of low TTE for school-age students in K–12 settings. Second, scholars have tested some of the hypothesized relationships proposed in the Wise–Smith (2011) model of TTE, but few scholars have studied *student-level* correlates of TTE. As such, it is currently unknown whether students from any particular demographic groups may be more likely to exhibit low TTE in low-stakes testing contexts. Third, another critical gap in the extant research literature concerns student-level motivational correlates of TTE. Although there is some evidence that examinee beliefs about test-taking are related to TTE, it is clear that the "internal factors" component of the Wise–Smith model warrants further investigation. Taken together, a review of the literature on student TTE points to the need for additional research addressing 1) the prevalence of students who exhibit low TTE, 2) subgroup differences in low TTE, and 3) relationships between malleable motivational factors and the odds students exhibit low TTE. The research questions in this study were guided by an application of the EEVT to the domain of test-taking to identify variables that could be targeted to help ameliorate the problem of low TTE on low-stakes academic tests. The EEVT suggests that internal factors (i.e., students' expectancies and values) may be particularly relevant to consider in order to understand TTE better, and previous research has suggested that student demographic characteristics may be associated with group differences in test-taking expectancy and value beliefs, which in turn may predict TTE in school-age students. The research questions and hypotheses for this study are described below.

Research question 1. What proportion of students in grades 4–8 exhibit low TTE on a CAT in reading, as determined by RTE (Wise & Kong, 2005)? It was hypothesized that the proportion of students in grades 4–8 identified as exhibiting low TTE would be similar to the proportions observed in previous studies of school-age students, which have ranged from 0.2% to 11.9%, depending on grade level, subject area, and the time of day (Wise et al., 2010).

Research question 2. To what extent do student demographic variables relate to the likelihood students exhibit low TTE on a CAT in reading? It was hypothesized that student demographic characteristics (i.e., grade, gender, race/ethnicity) would be correlates of the odds that students exhibit low TTE. Specifically, based on previous research in the domains of reading and testing, it was hypothesized that a) students in grade 8 would be more likely to exhibit low TTE than grade 4, b) male students would be more likely to exhibit low TTE than females, and c) non-White students would be more likely to exhibit low TTE than White students.

Research question 3. Do students differ in test-taking expectancy and value beliefs by student demographic variables? It was hypothesized that student demographic characteristics (i.e., grade, gender) would be correlates of test-taking expectancy and value beliefs. Specifically, based on previous research on reading and test-taking, it was hypothesized that a) students in grade 7 would report more positive test-taking expectancy and value beliefs than those in grade 8, and b) female students would report more positive expectancy and value beliefs than males.

Research question 4. To what extent do student test-taking expectancy and value beliefs relate to the likelihood students exhibit low TTE on a CAT in reading? It was hypothesized that test-taking expectancy and value beliefs would be correlates of low TTE. Specifically, based on previous research, it was hypothesized that a) students with lower test-taking expectancy beliefs and b) students with lower test-taking value beliefs would be more likely to exhibit low TTE.

CHAPTER III

METHODS OF STUDY I

Rationale for Two Studies

Two studies of student TTE were carried out to address the research questions (RQs) that guided the current investigation. First, Study I addressed RQ1 and RQ2 through a large-scale, nationally representative, secondary analysis of data from the STAR Reading test. Second, Study II replicated and expanded Study I through an empirical study in one school district, using both survey research and a secondary analysis of data from the STAR Reading test.

Purpose and Design of Study I

The primary purposes of Study I were to a) describe the proportion of students in grades four and eight who exhibited low TTE on a STAR Reading test and b) examine the hypothesized relationships between three stable individual student characteristics and the likelihood of being identified as a student with low TTE. The predictor (independent) variables of interest were grade level, gender, and race/ethnicity. The criterion (dependent) variable of interest, *low TTE*, was a binary categorical variable operationalized as RTE scores falling at or below 0.90.

Sampling Procedure

STAR assessment database. Using data from administrations of STAR Reading assessments completing during the 2014–15 school year, a targeted sample of item-level testing data were acquired from the testing company. The dataset used for the current analyses included student demographic information, test performance, and item-level testing data (including response times) for students in grades 4 and 8 who had completed a STAR Reading test in the winter of 2014–15 and for whom the following demographic information was input into the Renaissance Learning database by their school: grade, gender, and race/ethnicity.

Thus, the testing data represented a subset of all students in grades 4 and 8 in the existing STAR Reading database. It was assumed that most students who used a STAR Reading test in the winter of that school year had taken the test at least once before, so TTE was not expected to be unduly influenced by unfamiliarity with the test. To be included in the sample, the student must have completed a STAR Reading interim test such that item responses for each question were recorded and a standard score was derived from the student's performance. Cases with missing item response data were excluded (per the data screening procedure described below).

The researcher requested de-identified item-level data from the STAR Assessment database per the inclusion criteria (see Appendix B). Personally identifiable information (PII) was removed before receiving the data. Educational agencies that use the STAR Assessment system provided informed consent to the testing company prior to the storage of PII pursuant to the Application and Hosting Privacy Policy (Renaissance Learning, 2017). Parents of students who used a STAR Reading test who wished to revoke their consent for the storage of PII were given the opportunity to contact the child's school or district to have their educational records disclosed, changed, or removed from the STAR Assessment system. As stated in the privacy policy notice, available on the Renaissance Learning website (see Appendix C), "Renaissance Learning does not use your child's PII for any purpose other than to provide services to your child's school. Combined information that has been stripped of PII, and therefore not traceable to any student, is used for research and development so we can continuously improve our products and accelerate learning for all students" (Renaissance Learning, 2014, "Frequently Asked Questions About Student Information in our Software Products"). Therefore, participant consent for inclusion in research like the current Study I was obtained at the time students participated in a STAR Reading test, per the conditions of the user agreement for this assessment system.

Measures

STAR Reading test. The STAR Reading test is a CAT of reading comprehension skills completed by students in grades 1–12 using a computer or tablet device. According to the test developers (Renaissance Learning, 2016), STAR Reading tests are typically administered three times per year to all students for the purpose of universal screening of reading skills, administered monthly to monitor student reading progress and match instruction to the ability of each student, and/or administered weekly for monitoring the progress of students who receive intensive reading intervention. Each test consisted of 34 vocabulary-in-context items, which required students to read a passage and select from three choices the word that completes a sentence about the passage. The average test takes approximately 15 minutes to complete.

Research on the psychometric properties of STAR Reading has indicated that the test has adequate reliability and validity for the purposes of universal screening and progress monitoring, according to the testing standards developed by the National Center on Response to Intervention (U.S. Department of Education: National Center on Response to Intervention, 2010). Internal consistency reliability was measured using a random sample of 1.2 million STAR Reading test administrations, and results indicated high reliability for each grade level (range of 0.93 – 0.95), with a reliability coefficient of 0.97 for the full sample (Renaissance Learning, 2016). Evidence for the concurrent and predictive validity of the STAR Reading test has been demonstrated through empirical research on the correlations between student scores on the STAR Reading test and their current or future scores on other established measures of reading skills (Renaissance Learning, 2016). The results of more than 400 research studies have suggested there is strong evidence for the concurrent and predictive validity of STAR Reading tests, given that the average correlations between STAR Reading and other reading tests ranged from 0.65-0.87.

Predictor variables. The following variables were proposed predictors in the multiple logistic regression model(s): grade level, gender, and race/ethnicity.

Outcome variable. Dichotomously represented (low vs. not low) TTE was proposed as the criterion or outcome variable for the logistic regression models. The outcome variable was derived from the item-level test data using Wise and Kong's (2005) RTE procedure. Wise and Kong's (2005) RTE index is considered a proxy of TTE, and it represents the proportion of the test items on which the examinee had responded with SB (Wise, 2015). As previously described in the literature review, several replications of Wise and Kong's initial study have provided further evidence for the reliability and validity of this index. In previous studies, coefficient alpha values for RTE scores have ranged from 0.81–0.99, usually exceeding 0.90 (e.g., Kong et al., 2006; Kong et al., 2007; Wise et al., 2006; Wise & DeMars, 2006). Likewise, several researchers have provided evidence for the concurrent validity (e.g., Rios et al., 2014; Swerdzewski et al., 2011) and the discriminant validity (e.g., Kong et al., 2007; Rios et al., 2014; Wise & DeMars, 2010; Wise et al., 2009) of RTE scores. In this study, the derived RTE scores were used to classify each examinee as either exhibiting low TTE on the STAR Reading test (RTE at or below 0.90) or not exhibiting low TTE. To calculate the RTE index, researchers must select a time threshold to represent the boundary between RGB and SB. Responses are classified as SB if the response times are higher than the threshold time and classified as RGB otherwise. Finally, the number of SB exhibited by an examinee is summed and divided by the total number of items, and the resulting proportion is the student's RTE score. Because each administration of a STAR Reading test contained precisely 34 items, there were 35 discrete RTE scores that could be derived from an individual's testing data, ranging from 0.0 (no SB) to 1.0 (all SB).

In this study, the common threshold method was used to set the time threshold for RTE. Of the four conventional methods for setting time thresholds (see Kong et al., 2007), the common threshold is most practical for calculating RTE scores for a CAT like STAR Reading (given the immense size of the item bank used to administer test items). Prior research has shown that the four methods are all comparable for identifying low TTE (Kong et al., 2007). Response times (collected automatically and rounded to the nearest second) were compared to a common three-second threshold. That is, all of the responses submitted in less than four seconds (i.e., 0–3 seconds) were classified as instances of RGB. This procedure was consistent with the common threshold method employed by Wise and colleagues (2004).

After deriving the RTE index, students were identified as exhibiting low TTE if they earned RTE scores below 0.90. Several scholars have suggested 0.90 is a reasonable criterion for identifying low TTE (e.g., Rios et al., 2014; Swerdzewski et al., 2011; Wise & DeMars, 2010). As Swerdzewski and colleagues (2011) stated, "an examinee with an RTE of 0.90 only exerted effort on 90% of the test items. …it is reasonable for examinees to not try on a small portion of items (i.e., 10%) and still be retained in a dataset" (p. 172).

There is evidence that the 0.90 criterion meaningfully differentiates between examinees who had exhibited appropriate levels of TTE and those who had not given adequate effort. Several empirical studies (e.g., Kong et al., 2007) have shown that using a 0.90 criterion for score filtering can result in greater convergent validity correlations, and Wise (2015) found that RTE scores below 0.90 are low enough to distort scores and threaten individual score validity. On this test, RTE scores below 0.90 indicate that four or more of the responses were RGB because students exhibiting four RGB responses had RTE scores of 0.88 (30/34), whereas students exhibiting three RGB responses had RTE scores of 0.91 (31/34).

Data Analyses

Data screening and preliminary analyses. According to Raykov and Marcoulides (2008), data screening and preliminary analyses are conducted "(a) to ensure that the data to be analyzed represent correctly the data originally obtained, (b) to search for any potentially very influential observations, and (c) to assess whether assumptions underlying the method(s) to be applied subsequently are plausible" (p. 61). The values for all predictor and outcome variables across all participants were examined to ensure that the observed values were plausible. Items with negative (i.e., impossible) or unreasonably high response times were likely representative of an instrument malfunction, computer error, or test interruption. If instrumentation malfunction or a coding error is presumed to be the reason for missing data or outliers, Raykov and Marcoulides (2008) suggest this may warrant listwise deletion. Thus, listwise deletion was used to exclude cases for which response times included missing values or for which *all* of the response times were 0 seconds. Note that individual instances of response times coded as 0 were not treated as missing data, as these represented cases in which some responses were submitted in less than 0.5 seconds and rounded down to 0.

Descriptive statistics. To address RQ1, the proportion of students in grades 4 and 8 identified as exhibiting low TTE on a STAR Reading test (i.e., RTE scores below 0.90) was examined, and descriptive statistics for the sample were reported and disaggregated by group.

Comparative analyses. To address RQ2, potential mean RTE score differences by grade, gender, and race/ethnicity were tested using independent samples t-tests. Next, potential differences in the proportion of students flagged with low TTE by grade, gender, and race/ethnicity were tested using chi-squared tests.

Logistic regression analyses. To address RQ2 further, multiple logistic regression was performed to investigate the hypothesized relationships between the student characteristics and the probability of a student being flagged as exhibiting low TTE (described in terms of log odds). The test for significance in a logistic regression model is the Wald test, which tests the null hypothesis that a predictor does not affect the likelihood that the criterion variable is equal to one (Agresti & Finlay, 2009). In multiple logistic regression models, the *F* test of model significance indicates whether the full model had significantly improved the explanatory power of the restricted model. In a statistically significant model, any variables with significant regression weights are significantly associated with the outcome, controlling for all other predictors. If the parameter β for a given predictor variable is significant, it would indicate the increase in the log odds of being identified as exhibiting low TTE for each one-unit increase in the value of the predictor variable, holding all other predictors constant.

CHAPTER IV

RESULTS OF STUDY I

Data Screening and Preliminary Analyses

Testing records were obtained for 572,847 administrations of the STAR Reading test completed in the winter of the 2014–15 academic year. Each test event included 34 items, which resulted in 19,476,798 observed item responses. Preliminary analysis of the item response times for the sample revealed that a small number of responses were coded as being submitted after the 90-second time limit. This issue was presumed to be indicative of a computer error, and so those response times were treated as missing data. For the remaining 19,307,311 item responses, the average submission time was 32.68 seconds. As expected, there was a subset of responses that were submitted rapidly, (i.e., in three seconds or less). Specifically, 1,421 response times were rounded to zero seconds, 23,546 response times were rounded to one second, 67,356 response times were rounded to two seconds, and 79,455 response times were rounded to three seconds.

To support the appropriateness of using the RTE index, the effectiveness of the selected common threshold was evaluated. Results indicated that the overall accuracy for all responses was 68.36%. However, the accuracy of responses submitted in 0–3 seconds was much lower, ranging from 32.44% to 34.73%. Given there were three response options, the low accuracy rates (i.e., roughly one correct response for every two incorrect responses) for items submitted in three seconds or less were nearly equivalent to chance guessing (33.3%). This finding supported the assumption that there was a meaningful difference between responses classified as SB and RGB. Therefore, a three-second common threshold was accepted, meaning responses rounded to 0–3 seconds were considered RGB and responses rounded to 4–89 seconds were considered SB. Responses classified as RGB were submitted in less than 10% of the average time (32.68).

Results indicated that 19,135,533 responses (99.11%) were classified as SB, whereas 171,778 responses (0.89%) were classified as RGB. The accuracy of RGB was 34.00%, whereas the accuracy of SB was 68.67% (i.e., consistent with the overall average for all responses).

Descriptive Statistics

Demographic information for sample. Demographic data for the full sample of students (N = 572,847) are provided in Table 2. The sample included more fourth-grade students (60.40%) than eighth-grade students, and the sample was balanced by student gender. The race/ethnicity of students in the sample was listed most often as White (36.93%), Hispanic (16.98%), Black (16.2%), or Unknown (23.45%).

An unexpectedly low proportion of students (0.09%) had a value of 1 ("Yes") endorsed for the dichotomous variable representing LD status, whereas the remaining students (99.91%) had a default value of 0. Correspondence from the testing company indicated that it was not possible to ascertain whether values of 0 indicated that educators had entered a value of 0 ("No") or that the data were missing (because educators had provided no information about LD status).

RTE scores. The distribution of RTE scores for the sample (N = 571,386) is provided in Table 3. 1,461 students (0.26%) had missing item response times and were therefore excluded from the analyses. The mean RTE score for each subgroup of students is provided in Table 4. The overall mean RTE score was 0.99 (SD = 0.03). Most students (89.50%) had RTE scores equal to 1.0, indicating all 34 responses were classified as SB, whereas the other 10.50% of students had RTE scores falling below 1.0. In total, 96.91% of students had RTE scores above the 0.90 criterion, meaning they exhibited RGB on fewer than four test items. Mean RTE scores for subgroups of students in this sample are disaggregated in Table 5, Table 6, and Table 7.

Students with low TTE. Notably, there was a total of 16,250 students (2.84%) identified as exhibiting low TTE on the STAR Reading test. RTE scores for students flagged for low TTE ranged from 0.38–0.88, meaning some examinees exhibited RGB on as many as 21 of the 34 test items. The proportion of students with low TTE in each subgroup is provided in Table 8, and demographic information for the 16,250 students identified with low TTE is provided in Table 9.

Comparative Analyses

Mean RTE scores by subgroup. The mean RTE score for female students ($M = 0.9943$) was higher than the mean for males ($M = 0.9880$), $t(557{,}612) = 66.138$, $p < 0.001$, $d = 0.178$. The mean RTE score for grade 4 students ($M = 0.9922$) was higher than grade 8 ($M = 0.9897$), $t(444{,}090.47) = 24.624$, $p < 0.001$, $d = 0.069$. The mean RTE score for Asian/Pacific Islander students was highest ($M = 0.9961$), followed by White ($M = 0.9928$), Hispanic ($M = 0.9905$), and Black students ($M = 0.9869$). A one-way ANOVA indicated that the differences in mean RTE scores of students in different race/ethnicity subgroups were statistically significant, $F(437{,}361) = 442.031$, $p < 0.001$. The mean RTE score for students with an LD ($M = 0.9836$) was lower than the mean for students whose LD status was unknown ($M = 0.9912$), $t(486.434) = 3.387$, $p = 0.01$, $d = 0.176$.

Grade and low TTE. The proportion of students in grade 4 ($N = 344{,}945$) identified with low TTE on this test was 2.54 percent, whereas the proportion in grade 8 ($N = 226{,}441$) with low TTE was 3.31 percent. The chi-squared test indicated that the difference between the proportions of students with low TTE (0.77%) was statistically significant, $\chi^2(1) = 298.090$, $p < 0.001$. Students in grade 8 comprised 39.6% of the sample, yet they represented 46.2% of the 16,250 students with low TTE.

Gender and low TTE. The proportion of male students (N = 278,872) identified with low TTE on this test was 3.95%, whereas the proportion of females (N = 278,742) was 1.79%. The chi-squared test indicated this difference in proportions of students with low TTE (2.16%) was statistically significant, $\chi^2(1) = 2{,}388.262$, $p < 0.001$. Although the sample was balanced in regard to gender, male students represented 67.8% of the 16,250 students flagged with low TTE.

Race/ethnicity and low TTE. The chi-squared test indicated statistically significant differences among the proportions of students flagged with low TTE in the six race/ethnicity subgroups, $\chi^2(1) = 32.767$, $p < 0.001$. The proportion of Black students (N = 92,682) identified with low TTE on this test was highest at 4.33%, the proportion of Hispanic students (N = 96,957) with low TTE was 3.08%, the proportion of students with race/ethnicity endorsed as "Other" (N = 10,117) with low TTE was 2.80%, the proportion of American Indian/Alaskan native students (N = 5,798) with low TTE was 2.71%, the proportion of White students (N = 211,070) with low TTE was 2.27%, and the proportion of Asian/Pacific Islander students (N = 20,738) with low TTE was lowest at 1.22%.

Logistic Regression Analyses

Multiple regression model for low TTE. Finally, multiple logistic regression was performed to test the relationships between three predictors (grade, gender, race/ethnicity) and the odds of low TTE. Note that dude to group sizes, only White, Hispanic, Black, and Asian/Pacific Islander students were included in the logistic regression model using binary dummy variables for the subgroups. Because the LD status of most students (99.91%) could not be confirmed, the LD variable was not included in the model. The logistic regression model regressed grade, gender, and race/ethnicity on low TTE, and the full model was statistically significant, $\chi^2(5) = 3{,}292.642$, $p < 0.001$, explaining 3.4% of variance in the odds of low TTE.

Results of the full logistic regression model are provided in Table 10. Each of the five predictors in the model was significantly associated with the odds a student was identified as exhibiting low TTE on a STAR Reading test ($p < 0.001$). As such, the coefficients for the five significant predictor variables were interpreted. Results by ethnicity were more complex due to this variable having more than two categories and requiring binary coding. Within the model:

(a) Male students were 2.303 times as likely to exhibit low TTE as female students, controlling for all other factors.

(b) Eighth-grade students were 1.361 times as likely to exhibit low TTE as fourth-grade students, controlling for all other factors.

(c) Hispanic students were 1.378 times as likely to exhibit low TTE as White students, controlling for all other factors. Black students were 1.976 times as likely to exhibit low TTE as White students, controlling for all other factors. Asian/Pacific Islander students were 0.533 times as likely to exhibit low TTE as White students controlling for all other factors; stated another way, White students were 1.876 times as likely to exhibit low TTE as Asian/Pacific Islander students, controlling for all other factors.

CHAPTER V

METHODS OF STUDY II

Purpose and Design of Study II

The purpose of Study II was to replicate and extend the findings from Study I in order to investigate a) the extent to which students might differ in their test-taking expectancy and value beliefs, and b) the extent to which test-taking expectancy and value beliefs might relate to the likelihood students exhibit low TTE on a STAR Reading test. In Study II, an online survey protocol was used to measure individual student characteristics and test-taking expectancy and value beliefs, and low TTE was measured as described in Study I, using item response time data derived from administrations of the STAR Reading test. The rationale for Study II was that it allowed for the collection of data on test-taking beliefs, which were not available in Study I. In doing so, Study II addressed hypothesized relationships in the demands–capacity model of TTE.

Sampling Procedure

Participants were recruited using a convenience sample of students from a school district that used the STAR Reading test. Two middle schools were targeted for inclusion based on the following characteristics: geographic region, grade levels present, and administrator support for district-wide participation in the study. Recruiting participants from schools already using the STAR Reading test was desired because it allowed the researcher to replicate the analyses from Study I without requiring the participating students to complete any additional testing. Initial recruitment was completed through informal meetings with district administrators. According to student data from the Michigan Department of Education (2018; see www.mischooldata.org), students in the targeted district were mostly White (75%) and Hispanic (14%), not economically disadvantaged (75%), and proficient on the statewide English language arts test (60–70%).

According to the district's curriculum and instruction website, the test was used for the following purposes: complying with state legislation requiring the regular assessment of reading skills, screening students for reading difficulties, grouping students for intervention classes and differentiated instruction, and monitoring reading progress over time. At the middle school level, test scores were used as an indicator of the proportion of students with grade-level proficiency in reading and used to measure progress toward goals outlined in their school improvement plans. Test data were shared with parents during conferences and included in student report cards, and parents were directed to contact their child's language arts teacher for additional information.

Measures

Student Perceptions of Testing Survey. Data on demographic characteristics, test-taking expectancy beliefs, and test-taking value beliefs were derived from participants' responses to a brief online survey (Appendix D) called the "Student Perceptions of Testing Survey" (SPOTS). This 23-item survey included two items measuring student demographic information, two practice items, and nineteen items measuring student beliefs related to the STAR Reading test. Seven predictor variables were coded using the following procedures. (Note that test-taking beliefs were both predictor and outcome variables, depending on the research question being addressed by a given analysis.)

Grade level. One item measured student grade-level (coded grade seven or grade eight).

Gender. One item measured student gender (coded male or female).

Test-taking expectancy beliefs. Test-taking expectancy beliefs were measured using five items adapted from the Academic Efficacy subscale of the Patterns of Adaptive Learning Scales (PALS; Midgley et al., 2000), a set of well-established measures of motivation (Senko, 2016).

The PALS Academic Efficacy subscale consists of five items measuring an individual's perceived competence to do school work, and items are similar to the expectancy belief measures used by Wigfield and Eccles (2000), which have been studied extensively (Wigfield et al., 2016). PALS items are rated using five-point Likert-type scales, anchored at 1 ("Not at all true for me"), 3 ("Somewhat true for me"), and 5 ("Very true for me"). Internal consistency for this subscale has been shown to be adequate (α = 0.78; Midgley et al., 2000). In the SPOTS, five items from the Academic Efficacy scale were adapted to more specifically measure expectancies for success when taking the STAR Reading test (see Appendix E). Test-taking expectancy beliefs subscale scores were derived from the average ratings of the five items and ranged from 1.0–5.0.

Test-taking value beliefs. In Study II, test-taking value beliefs were measured using items adapted from Conley's (2012) four subjective task value subscales, which were based on the work of Eccles, Wigfield, and colleagues (e.g., Wigfield & Eccles, 2000). In the SPOTS, 14 items from Conley's value subscales were adapted to more specifically measure value beliefs related to the STAR Reading test. The value scales were shortened to reduce the overall length of the SPOTS (see Appendix F), and items selected for inclusion in the SPOTS were those judged by the researcher to be most pertinent to test-taking, most appropriate for use with middle school students, and most reliable according to Conley (2012). Each test-taking value subscale scores was derived from the average rating of the items in the adapted scale (ranging from 1.0–5.0).

The first type of task value in the EEVT is attainment value, which refers to perceptions of how a task relates to important aspects of an individual's identity (Wigfield & Eccles, 2000). Conley (2012)'s original Attainment Value scale (α = 0.85) consisted of six Likert-type items (e.g., "Being someone who is good at math is important to me"), and four of the items were selected for administration in the current study and adapted to reference the STAR Reading test.

The second type is intrinsic value, which refers to one's enjoyment or interest in a task (Wigfield & Eccles, 2000). Conley's original Interest Value scale (α = 0.96) consisted of six items (e.g., "I like math."), and four of the items were selected for administration in the current study and adapted to address the intrinsic value of STAR Reading.

The third is utility value, the usefulness of a task for reaching one's goals (Wigfield & Eccles, 2000). Conley's Utility Value scale (α = 0.80) was four items (e.g., "Math will be useful for me later in life.") Four items were selected and adapted to reference STAR Reading.

The last is relative cost value, which represents the belief that one will lose other desired opportunities (Wigfield & Eccles, 2000). Conley's (2012) Cost Value scale (α = 0.70) consisted of two items (e.g., "I have to give up a lot to do well in math."), and both items were adapted.

Low test-taking effort. The outcome variable of interest in Study II, low TTE on a STAR Reading test, was measured using Wise and Kong's (2005) RTE method, as described in the previous chapter. The method for computing the outcome variable in Study II was the same as the method in Study I (i.e., RTE below 0.90 were flagged as indicators of low TTE).

Procedures

In Study II, RTE scores were derived from secondary analyses of data from the existing STAR Assessment database following the procedures for data retrieval described in Study I. As previously stated, districts using STAR Assessments provide informed consent to Renaissance Learning per the Application and Hosting Privacy Policy (Renaissance Learning, 2017). The researcher requested written permission from administrators in the targeted district to receive test data from the STAR Reading database for the purpose of learning more about the use of the test in the district (see Appendix G).

The Star Reading test data for participants in Study II were fully anonymized (i.e., school ID numbers were removed) before the researcher received the data. Instead, the only identifier associated with the STAR Reading test data was a randomly generated Renaissance Learning ID number, which the researcher could not connect to students. For this reason, the Michigan State University institutional review board classified Study II as exempt from review for the protection of human subjects (see Appendix H).

In addition to requesting express written consent from district administrators for students in the targeted school district to participate in the proposed study, the researcher also shared an information letter with all parents of children in the district to inform them about the purpose and scope of the research study (see Appendix I). Parents were informed that a) their child's participation in the online survey was entirely voluntary, b) their child could elect to opt out of the survey at any time, c) information gathered through the research study would not affect their child's grades, instruction, or eligibility for any educational services or supports, d) PII would not be gathered through the study, and e) the researcher would not disclose the survey responses or performance of any individual or class.

After consent for district-wide participation was obtained from the superintendent of the participating district, students were invited to take the online survey during the spring of the 2017–2018 school year (following the spring administration of the STAR Reading test). In their language arts classrooms, students were invited to open and complete the SPOTS. Information about the study and the survey instructions were read aloud to students by the teachers using a standardized protocol (Appendix D). Students were given informed consent documentation and instructed to indicate their assent to participate in the survey by entering their Renaissance ID number (shared with them by their teacher).

After the participating students had provided assent, teachers read each item of the SPOTS aloud, including two practice items (to ensure students knew how to use the scale) and nineteen items measuring their expectancy and value beliefs reading to STAR Reading. Following the completion of the survey, all students in the participating schools received a gift certificate to a local ice cream store (regardless of whether they had completed the survey). School staff were invited to attend a presentation by the researcher related to the study findings.

Data Analyses

In general, the analyses for Study II followed the same methods as Study I. That is, data screening, analyzing item response times, handling missing data, analyzing the accuracy of the response time threshold, calculating RTE scores, and identifying students with low TTE were completed in the same manner as previously described in the Methods chapter for Study I.

Descriptive statistics. As in Study I, demographic information for the sample are presented, and descriptive statistics for RTE scores and the outcome variable (low TTE) are presented and disaggregated by subgroups (grade and gender). To inform RQ1 further, mean RTE scores were reported, and the proportions of students in grades 7–8 who were flagged as exhibiting low TTE on STAR Reading test (i.e., RTE scores below 0.90) were described.

Principal component analysis of SPOTS. Principal component analysis (PCA) was performed using the data from the SPOTS survey to refine the subscales used in subsequent analyses iteratively. As previously stated, the survey measuring test-taking expectancy and test-taking value beliefs was an adaptation of two established measures of academic efficacy and task value beliefs. Given the SPOTS included five subscales designed to measure test-taking expectancy beliefs and four types of test-taking value beliefs, it was essential to examine whether the SPOTS responses fit the anticipated factor structure.

The PCA was conducted following the procedure used by Brown and Hirschfeld (2008) for refining the SCoA. The goodness-of-fit indices for the resulting factor structure were reported, and the descriptive statistics for each identified factor in the model were described. Based on the results of the PCA, the psychometric properties of the resulting subscales were analyzed, and mean responses for items and subscales were provided.

Comparative analyses. To address RQ3, a series of one-way multivariate analysis of variance (MANOVA) was used to examine whether students differed by grade or gender in their test-taking expectancy or test-taking value beliefs.

Logistic regression analyses. To address RQ2 further and to address RQ4, hierarchical logistic regression analyses were performed using the binary variable low TTE as the criterion variable. First, the student demographic variables (grade, gender) were entered into the model. Next, the student test-taking belief (expectancy, attainment, intrinsic, utility, and cost) subscales were entered into the multiple logistic regression model. If results of the *F* test indicated model significance, then the coefficient of determination for the full model would represent the overall proportion of variance in the log odds of being identified as exhibiting low TTE that is explained by all of the predictor variables in the model. Furthermore, if the full model was statistically significant, then any of the predictor variables with significant regression weights would be significantly associated with the outcome variable, controlling for all other predictors. If the parameter β was found to be significant for a predictor variable, then the estimated parameter would indicate the increase in the log odds of being identified as exhibiting low TTE for each one-unit increase in the value of the predictor variable, holding all other predictors constant.

CHAPTER VI

RESULTS OF STUDY II

Data Screening and Preliminary Analyses

Testing records were obtained for 826 administrations of the STAR Reading test which were completed in the spring of the 2017–18 school year by students in grades seven and eight from one Midwestern school district. Item response times were reviewed and any items with response times higher than 90 seconds were removed, which resulted in 28,736 observed item response times. The same three-second common threshold was applied to the data, and the results indicated that there were indeed observable instances of RGB in this dataset.

In Study II, there were 471 responses (1.64%) submitted in three seconds or less. Of those rapid responses, only 157 (33.33%) responses were correct. This accuracy rate was similar to the rate observed in Study I. Additional data screening revealed that 20 test records (2.36%) did not contain precisely 34 items, and those cases were excluded from the sample for all further student-level analyses. To address RQ3 and RQ4, students were only included in analyses if they a) completed the online survey with their Renaissance ID number and b) completed the STAR Reading test such that item-level data were available for matching and secondary analysis.

Descriptive Statistics

Demographic data for the participating students who were included in the final sample for Study II ($N = 675$) are provided in Table 11. The sample included more seventh-grade students than eighth-grade students and included more female students than male students.

RTE scores for sample. The distribution of RTE scores is provided in Table 12. Results indicated that 557 students (82.52%) had RTE scores equal to 1.0, meaning all 34 of their responses were submitted after the three-second threshold and thereby classified as SB.

The remaining 17.48% of students exhibited at least one RGB and had RTE scores below 1.0. In total, 93.8% of students in this sample had RTE scores that fell above the 0.90 criterion. Furthermore, a noteworthy total of 42 students (6.22%) in the Study II sample were identified as exhibiting low TTE on the STAR Reading test. The RTE scores for students flagged with low TTE ranged from 0.41–0.88, meaning some examinees exhibited RGB on as many as 20 of the 34 test items (and therefore showed SB on only 14 of the 34 items). The proportion of students flagged with low TTE in each subgroup is provided in Table 13, and the demographic information for the 42 students flagged with low TTE is provided in Table 14.

Student perceptions of testing survey. A total of 748 students completed the online survey, which indicates that approximately 98 students opted out of taking the survey or did not complete it. Please note that this total is greater than the number of students for whom their survey responses could be linked to their STAR Reading data. Specifically, 35 students entered an ID number that did not correspond with a test ID, 29 students entered an ID that was not unique, and 20 students entered no ID. No students or parents contacted the lead researcher for additional information about the study. Descriptive statistics for each of the 19 items in the SPOTS are provided in Table 15. A PCA was performed on the 19-item SPOTS. The suitability of the PCA was evaluated prior to analysis, and inspection of the correlation matrix showed that all variables had at least one correlation that was greater than 0.3. The Kaiser-Meyer-Olkin measure of sampling adequacy was 0.877, and Bartlett's Test of Sphericity was statistically significant ($p < 0.001$), indicating the survey was appropriate for factoring into components.

The PCA revealed five principal components that had eigenvalues greater than 1.0, and the components explained 16.54%, 15.95%, 15.48%, 14.77%, and 8.12% of the total variance in the SPOTS, respectively. Eigenvalues for the principal components ranged from 1.22 to 6.56.

The five-component solution explained 70.87% of the total variance, and inspection of the scree plot supported a five-factor solution. Varimax orthogonal rotation was employed to aid interpretability, and the factor structure met the interpretability criterion. The respective items for each subscale of the SPOTS loaded on the anticipated factor, which indicated that the survey was consistent with the previously validated expectancy and value subscales from which it was derived. Thus, the five factors were retained, and scale scores for each factor were computed. Descriptive statistics for the resulting subscales of the SPOTS, disaggregated by subgroup, are provided in Table 16 and Table 17. Reliability coefficient alpha exceeded 0.80 for all subscales but cost ($\alpha = 0.67$). In general, participants reported perceiving the STAR Reading test as low in attainment value ($M = 2.91$), interest value ($M = 1.76$), and relative cost ($M = 1.67$). Participants reported moderate ratings for utility value ($M = 3.02$). Students in this sample reported relatively high test-taking expectancy beliefs ($M = 3.25$) when compared to their test-taking value beliefs.

Comparative Analyses

Grade level and test-taking beliefs. A one-way MANOVA was performed to examine the relationships between grade and test-taking expectancy and value beliefs. The results of the MANOVA indicated there was no statistically significant difference between grade levels on the combined dependent variables, $F(5, 669) = 1.878$, $p = .096$; Wilks' $\Lambda = .986$; partial $\eta^2 = .014$. However, post-hoc univariate one-way ANOVA analyses indicated that there were statistically significant differences by grade for utility value, $F(1, 673) = 5.199$, $p = .023$, and attainment value, $F(1, 673) = 3.944$, $p = .047$. Specifically, students in grade 7 reported higher ratings for utility and attainment value ($M = 3.11$, $SD = 1.04$; and $M = 2.99$, $SD = 0.98$, respectively) than those in grade 8 ($M = 2.93$, $SD = 1.02$; and $M = 2.84$, $SD = 1.01$, respectively). Results indicated differences by grade level for expectancy beliefs, interest, and cost were non-significant.

Gender and test-taking beliefs. Next, a one-way MANOVA was performed to examine the relationships between grade level and test-taking expectancy and value beliefs. The results of the MANOVA indicated that there was a statistically significant difference by gender on the combined dependent variables $F(5, 669) = 3.899$, $p = 0.002$, partial $\eta^2 = .028$. Further, post-hoc univariate one-way ANOVA analyses indicated that there was a statistically significant gender difference for cost, $F(1, 673) = 11.228$, $p = .047$. Specifically, females reported lower ratings for cost value ($M = 1.57$, $SD = 0.75$) than males ($M = 1.78$, $SD = 0.92$). Results indicated differences by gender level for expectancy beliefs, interest, utility, and attainment were non-significant.

Logistic Regression Analyses

A hierarchical logistic regression analysis was performed to investigate whether the demographic variables (grade, gender) and motivational variables (expectancy and value beliefs) were associated with the likelihood students exhibited low TTE. The bivariate correlation matrix describing the relationships between predictor and outcome variables is provided in Table 18. The linearity of the continuous variables with respect to the logit of the criterion variable was assessed via the Box-Tidwell (1962) procedure. A Bonferroni correction was applied using all ten terms in the model, resulting in statistical significance being accepted when $p < .005$ (Tabachnick & Fidell, 2019). Based on these preliminary assessments, all of the continuous independent variables were found to be linearly related to the logit of the criterion variable.

The first restricted model regressed grade and gender on the criterion variable of low TTE. The model explained 5.5% of the variance in the criterion variable, and the model was statistically significant, $\chi^2(1) = 13.967$, $p = .001$. The full model regressed all seven predictor variables (grade level, gender, expectancy, intrinsic, attainment, utility, and cost) on the criterion variable of low TTE, and the full model was statistically significant, $\chi^2(7) = 26.224$, $p < .001$.

The full logistic regression model (see Table 19) explained 10.2% of the variance in the criterion variable and correctly classified 93.8% of cases. In the full model, four of the predictors (gender, grade level, attainment value, and cost value) had statistically significant associations with the criterion variable, whereas expectancy, interest value, and utility value did not. As such, the coefficients for the four significant predictor variables were interpreted. Within the model:

(a) Male students were 2.675 times as likely to exhibit low TTE as female students, controlling for all other factors.

(b) Students in grade 7 were 2.070 times as likely to exhibit low TTE as grade 8 students, controlling for all other factors.

(c) For each one-point increase in cost value score (greater perceived cost), students were 1.562 times as likely to exhibit low TTE, controlling for all other factors.

(d) For each one-point increase in attainment value score, students were 0.684 times as likely to exhibit low TTE, controlling for all other factors; stated another way, each one-point *decrease* in attainment value score was associated with being 1.462 times as likely to exhibit low TTE, controlling for all other factors.

CHAPTER VII

DISCUSSION

The purpose of the present investigation was to examine the prevalence of low TTE and to identify student-level correlates of low TTE. This discussion summarizes the major findings from two quantitative studies, notes general limitations, and describes implications for theory, research, and practice in this topic area. The research questions for this study were as follows:

(1) What proportion of students in grades 4–8 exhibit low TTE on a CAT in reading, as determined by RTE (Wise & Kong, 2005)?

(2) To what extent do demographic variables relate to the likelihood students exhibit low TTE on a CAT in reading?

(3) To what extent do students differ in test-taking expectancy and value beliefs by student demographic variables?

(4) To what extent do student test-taking expectancy and value beliefs relate to the likelihood that students exhibit low TTE on a CAT in reading?

Summary of Major Findings

The current findings extended the previous research literature on the prevalence of low TTE and the student-level correlates of low TTE. Specifically, this study made four primary contributions to the existing literature. First, this was one of the first studies to explore TTE using item response time data from K–8 test-takers. This large-scale investigation of student TTE was relatively unique, as few previous studies had measured TTE using item response time data (a feasible and non-obtrusive approach) rather than self-report (Wise, Ma, & Theaker, 2014). Results from Study I revealed that the proportion of test-takers with low TTE was 2.84%, which supported the research hypothesis that the observed prevalence would be in the range of 1–12%.

Second, the study replicated earlier research on the student-level correlates of TTE, which has shown that certain demographic variables are associated with RTE. Specifically, as found in Wise and Cotton (2009), female students had significantly higher RTE scores than male students. The results of Study I supported the hypothesis that grade, gender, and race/ethnicity would be significantly associated with the odds a student was flagged as exhibiting low TTE.

Third, this study advanced knowledge about motivational variables that might inform *why* certain students might be more likely to show low TTE. Results from Study II supported the hypothesis that certain test-taking beliefs would differ by grade and gender. More specifically, younger students reported higher utility and attainment, and female students reported lower cost.

Fourth, much of the work on student TTE has been contextualized within the EEVT, but only a few empirical studies (e.g., Cole et al., 2008) had directly tested whether motivational variables from the EEVT are associated with student TTE. Results of Study II provided support for the hypothesis that certain test-taking beliefs would be associated with the odds of being identified with low TTE. Specifically, higher cost value beliefs were positively associated with low TTE, whereas higher attainment value beliefs were negatively associated with low TTE.

Interpretation of Results

Prevalence of low TTE (RQ1). A primary goal of this study was to measure the prevalence of low TTE in K–12 schools using response time data. There is evidence to suggest that this study effectively replicated previous research using the RTE approach. Consistent with previous work by Wise and colleagues (e.g., 2004; 2005; 2012), the accuracy rates of items classified as RGB were comparable to those expected by chance guessing, which suggests that item responses classified as RGB were almost certainly invalid responses. In this study, the accuracy rate of items classified as RGB was approximately 33% (for items with three choices).

Results from Study I indicated that 2.84% of students were identified with low TTE, and subgroup analyses showed that prevalence rates for various demographic groups ranged from 1.22–4.33%. Indeed, a majority of students in Study I (97.16%) had RTE scores above 0.90, a generally accepted indicator of adequate effort. Still, results indicated that 8,749 fourth-grade and 7,501 eighth-grade students exhibited low TTE on a STAR Reading benchmark assessment. These figures might seem negligible compared to the total number of fourth-grade and eighth-grade test-takers (which exceeded 572,000), but those statistics represent 16,250 individual students whose test scores were likely not valid due to low TTE. Notably, results from Study II revealed an even higher rate of low TTE, with 6.22% of seventh-grade and eighth-grade students from the participating school district having RTE scores that were below 0.90.

In general, the results of Study I were consistent with findings from previous studies by Wise and colleagues (e.g., 2004, 2010, 2012, 2016), which had applied Wise and Kong's (2005) RTE method in K–12 contexts. For instance, Wise and colleagues (2010) reported average RTE scores by grade for a MAP test in reading; the mean RTE score for fourth-grade was 0.994, and the mean RTE for eighth-grade was 0.983. This study yielded similar results, with fourth-grade and eighth-grade students having mean RTE scores of 0.992 and 0.990, respectively. The total prevalence of low TTE in Study I was also relatively consistent with prior studies using different measures of TTE, such as Nering, Bay, and Meijer's (2002) finding that 1.1% of K–12 students could be identified with low TTE using pattern-based (e.g., ABCDABCD) person-fit statistics. By contrast, the observed prevalence of low TTE in this study was considerably lower than those found in some other RTE studies with K–12 and college-age samples. For instance, Wise (2015) found over 11.5% of students in ninth grade were flagged with low TTE on a MAP reading test, which was much higher than the 3.3% of students in eighth grade with low TTE in Study I.

One noteworthy difference between Wise's (2015) investigation and the present study was the method used for measuring the criterion variable. Whereas the current study employed a common three-second threshold, Wise (2015) used the normative threshold method introduced by Wise and Ma (2012), which the authors claimed might be more appropriate for classifying items as SB or RGB. The normative threshold method could not be applied in this study, and therefore it is possible the prevalence of low TTE would have been higher using another method.

As expected, the prevalence of low TTE was much lower than those in college studies. However, the reason for this discrepancy is unclear, and no studies have directly compared the TTE of K–12 and college students. One explanation is that college students have more control over how they spend their time (and less oversight over their behavior), which could suggest the opportunity cost of testing might be greater for college students. Another possibility is that the design of the test or testing context could result in higher perceived RD for college students. Previous studies (e.g., Swerdzewski et al., 2011) have involved assessment batteries consisting of multiple tests taken over several hours. By contrast, tests like STAR Reading typically are completed in fewer than thirty minutes. As such, it is possible that the higher rates of RGB in college contexts could be explained by mental fatigue that students experienced on exceptionally lengthy tests. A final explanation could be that many of the previous research studies of college students were based on student responses on an *optional* test, which many students might have viewed as particularly unimportant. On the contrary, school-age students are typically *required* to take tests like STAR Reading, which was the situation for the testing data used in the current study. Even though the reasons for such high rates of RGB in college settings is currently uncertain, the present study was important because it helped to further clarify the prevalence of low TTE in K–12 settings. Findings were strengthened by the use of a large, national dataset.

Student-level correlates of low TTE (RQ2). The second contribution of this study was that it addressed the relationships between student demographic variables and the odds a student was identified with low TTE. Notably, gender was the strongest student-level predictor of low TTE in Study I, with male students being 2.28 times as likely to exhibit low TTE as female students. Because the majority of students in the sample had RTE scores of 1.0, results indicated a small effect size when comparing mean RTE scores by gender ($d = 0.178$). Still, the results of logistic regression analyses did reveal a significant gender difference in the log odds of being flagged with low TTE. Specifically, 68% of the students flagged with low TTE in Study I were male, even though the large sample was balanced by gender. Results suggested that male students were more than twice as likely to exhibit low TTE, controlling for grade and race/ethnicity. The results from Study II were similar, indicating that 71% of students with low TTE were male, despite the sample also being balanced by gender. These findings were convergent with previous research on gender and TTE, which has shown lower TTE in male students (Eklöf, 2007; Wise & DeMars, 2010; Wise et al., 2004). This investigation extended previous research because it was the first study to statistically test mean differences in the RTE scores of male and female students in elementary or middle school, whereas previous studies (e.g., Wise et al., 2010) had reported only descriptive statistics by group.

Regarding student grade level, the predicted relationship between grade and TTE was supported, with the log odds of exhibiting low TTE being 1.317 times as high for eighth-grade students compared to fourth-grade students. Again, the effect size was small for differences in mean RTE score by grade ($d = 0.069$), but the proportion of students flagged with low TTE was significantly higher for eighth-grade students than fourth-grade students. These findings were consistent with prior research that has shown a decrease in TTE as students go through school.

The observed proportion of students with low TTE in grades four (2.54%) and eight (3.31%) in Study I were lower than those observed by Hauser and Kingsbury (2009) for grades three (6.9%) and nine (7.7%), although those authors had used a different flag for identifying low TTE. Wise and colleagues (2010) had also observed higher rates of low TTE in eighth-grade students compared to fourth-grade students. When considered together, the results of the present study seem to support previous findings that student TTE tends to decline as students age, which would be consistent with previously documented declines in general academic motivation over time (Archambault et al., 2010; Durik et al., 2006; Jacobs et al., 2002; Wigfield et al., 1997). It could be that elementary teachers more consistently communicate the importance of giving one's best effort on a test, whereas middle school teachers might assume that their students have already internalized the rationale for doing one's best on the test. Another possible explanation could be that elementary teachers monitor their students more closely while they take the tests.

Furthermore, there were also significant differences by race/ethnicity in the proportion of students who were found to exhibit low TTE. The demographic subgroups in Study I with the highest proportion of students exhibiting low TTE were Black students (4.33%) and Hispanic students (3.08%). This trend was consistent with prior research findings that Black and Hispanic students may be at higher risk for academic disengagement when compared to White students (Graham & Taylor, 2002). Conversely, Asian/Pacific Islander students had the highest average RTE score and were the least likely to be identified as exhibiting low TTE. Previous research has suggested Asian American/Pacific Islander students tend to report greater overall educational expectations from their families compared to European American students (Mello, 2009), and so it is possible the students from that subgroup may internalize those high expectancy beliefs.

Taken together, the results of this study revealed several subgroup differences in RTE score and the likelihood of being identified with low TTE on a STAR Reading test, and all of the significant relationships were in the predicted directions. Students with the highest likelihood of exhibiting low TTE were male, Black, and eighth-grade students, and students with the lowest log odds were female, Asian/Pacific Islander, and fourth-grade students. These results supported the research hypothesis that different subgroups of students would show varying levels of TTE on a low-stakes reading test. Future research should continue to address why young, female, White, and Asian/Pacific Islander students tend to show higher TTE than other student groups.

Psychological correlates of low TTE (RQ3 and RQ4). As previously noted, several findings from Study II represented meaningful contributions to the previous research literature on student beliefs about academic test-taking and student TTE.

First, the current findings provided support for the hypothesis that some test-taking *value* beliefs would differ by demographic group. Results suggested that two subtypes of value beliefs (utility and attainment) varied by grade, and one type of value belief (cost) varied by gender. Specifically, seventh-grade students in this sample reported perceiving the test as more useful (higher utility value) and more relevant to their identities (higher attainment value) on average than eighth-grade students; comparisons by gender indicated male students reported perceiving the test as a greater loss of other opportunities (higher cost value) on average than females.

The significant relationships between grade and test-taking value beliefs were consistent with survey research by Paris and colleagues (1991, 2000), who found that older students were less likely to describe achievement tests as valuable to them. However, neither interest value nor cost value was significantly related to low TTE. The relationship between grade and interest value may have been nonsignificant because students reported low interest in the test overall.

Further, one type of value belief (cost) was significantly associated with gender, and this relationship had the largest effect size of all comparative analyses ($d = 0.257$). This result could suggest cost value beliefs might be particularly crucial for explaining differences in TTE. Still, it remains unclear why male students reported higher cost value beliefs in regard to taking this test.

Second, results did not support the hypothesis that test-taking *expectancy* beliefs would differ significantly by grade or by gender. This finding was unexpected given previous research showing age and gender differences in academic self-efficacy. Instead, the majority of students reported moderate expectancies for success the next time they would take a STAR Reading test. One possible explanation for the lack of variability in test-taking expectancy beliefs could be related to the adaptive nature of the test. That is, the item-selection algorithm for the STAR Reading test is designed such that all students receive items at a level that is determined to be of moderate difficulty given their earlier individual performance results, which might explain why the students reported neutral test-taking expectancy beliefs overall. On the other hand, it is also possible that the nonsignificant differences in expectancy beliefs have pointed to a meaningful finding about TTE in school-age students. Specifically, the results could suggest that test-taking *value* beliefs are more likely to vary among students than test-taking *expectancy* beliefs about low-stakes test, which would potentially make the value component of the EEVT particularly critical for a better understanding of test-taking perceptions and student TTE in general. Further research on expectancy beliefs in low-stakes testing settings could help address this question.

Third, the results of Study II provided mixed support for the hypothesis that test-taking *value* beliefs would be significantly associated with low TTE. Specifically, the current findings suggested that two types of value beliefs (attainment and cost) were significantly associated with the likelihood of being identified with low TTE (whereas interest and utility were not).

The results of Study II suggested that perceptions that taking a STAR Reading test had a higher relative cost (i.e., limited them from participation in some other preferred activities) were associated with poorer TTE. Controlling for other factors, a one-point increase in relative cost beliefs was associated with being 1.55 times as likely to exhibit low TTE. Likewise, a one-point decrease in attainment value was associated with students being 1.435 times as likely to exhibit low TTE, which was consistent with previous research on relationships between value beliefs and TTE. Cole and colleagues (2008) gave a self-reported value belief scale to college students who had taken general education tests, and they found that two distinct types of test-taking value beliefs (perceived utility value and attainment value) were significant predictors of self-reported TTE. This study provided further evidence that test-taking *attainment* value beliefs are related to TTE, but results did not support the relationship between test-taking *utility* value and TTE.

Fourth, results did not support the hypothesis that test-taking *expectancy* beliefs would be significantly associated with low TTE. This unexpected finding is especially noteworthy because it suggests that the test-taking *value* component of the current model might be particularly crucial for our understanding of low TTE on this test, whereas the test-taking *expectancy* component of the model might not have the same explanatory power. Previous studies (e.g., Wise et al., 2009) have shown that item difficulty is unrelated to the likelihood a student displays RGB or SB, and so these findings might suggest student perceptions of their ability to pass a test item have less influence on TTE than their perceptions about whether it would be valuable to try their best.

Altogether, this study advanced our understanding of TTE, but researchers will need to continue investigating the phenomenon of TTE to extend our understanding of the reasons why low TTE occurs. Given the remaining questions in this topic area, the present findings have shed light on some essential lines of inquiry for researchers to address through future empirical work.

Implications for Theory and Research

The current findings provided evidence that demographic characteristics and motivational variables were significant correlates of low TTE, which has important implications for Wise and Smith's (2011) demands–capacity model of TTE. Surprisingly, results suggested *value* beliefs, but not *expectancy* beliefs, predicted the likelihood a student exhibited low TTE. However, the students in the current sample reported limited variability in their expectancy beliefs related to the STAR Reading test, with students reporting moderate to positive expectancy for success. Future research with a nationally representative and randomly selected sample is needed to corroborate the nonsignificant relationship between expectancy beliefs and the odds of low TTE.

Conversely, this finding could highlight the relative importance of the value component of an EEVT-informed model of TTE and suggest that theoretical models of TTE should account for the specific subtypes of task value. As such, future researchers could further investigate how attainment value and cost value relate to TTE. An essential next step might be to use more advanced statistical modeling to examine whether changes in attainment value over time might be a factor that explains the observed decline in TTE as students age (e.g., Wise et al., 2010). It also seems critical to examine further why male students reported perceiving this sort of reading test as having a higher cost compared to female students. Relatedly, the current investigation demonstrated distinct differences in the TTE of students from different racial/ethnic groups, but there remain questions about why these differences may exist. Given the significantly elevated risk of disengagement from test-taking for Black and Hispanic students, it seems especially important for future researchers to explore the motivational patterns of students from different racial/ethnic groups and uncover why these disparities exist. Doing so could be an important first step toward developing differentiated strategies to support the TTE of these student subgroups.

Indeed, evolving models of TTE should reflect the complex nature of engagement, which has behavioral, emotional, and cognitive components (Fredricks, Blumenfeld, & Paris, 2004). Wise and Smith's (2011) model suggests that TTE results from interactions among student-level, test-level, and contextual variables, and previous research on school engagement suggests that engagement is influenced by "family, community, culture, and educational context" (Fredricks et al., 2004, p. 73). To extend upon the current findings, it might be critical for future research to explore the contextual variables that might be malleable antecedents of low TTE. As Wigfield and colleagues (2016) stated, "experiences with different activities, parent and teacher feedback about the importance and usefulness of different activities, and children's comparison of interest in different activities to those of their peers may all influence children's valuing of activities" (p. 62). As such, future research might focus on specific behaviors by parents and teachers that could affect student perceptions about test-taking. For instance, teacher recognition when students give their best effort has been found to support their motivation and engagement (Schunk & Pajares, 2009). It is possible that students receive little to no feedback on their effort when they are taking CBTs, which could diminish the attainment value of the test. Although teachers report communicating the importance of learning (Brophy, 1999), it is unclear whether they convey the value of doing one's best on tests that don't count toward a grade. Therefore, future research could explore how teachers present the purpose and importance of TTE before administering the test. Newmann and colleagues (1992) asserted that authentic academic tasks are "meaningful, valuable, significant, and worthy of one's effort, in contrast to those considered nonsensical, useless, contrived, trivial, and therefore unworthy of effort (p. 23). Knowing more about how teachers connect the testing process to the "real world" could shed light on the mechanism by which students develop positive or negative perceptions about a test.

Furthermore, future research should address contextual factors that may help explain the subgroup differences in TTE observed in the current student. Previous research has revealed numerous factors that could explain differences in school engagement by students from minority groups, but the reason for group differences in TTE still remains unclear. It is possible that some students from certain minority groups may demonstrate behavioral disengagement as a conscious response to an education system they view as unjust (Ogbu, 1992). Others may underperform in school because they fear that giving their best effort could distance them from their peer group (Fordham & Ogbu, 1986). Other scholars have found that stress from stereotypes about the underachievement of a minority group results in poorer performance for students from that group (Osborne, 1997; Steele, 1997). Parental expectations for children's achievement are associated with student engagement among students from racial and ethnic minority groups (Murray, 2009), so it might be helpful for researchers to explore whether parental beliefs about the importance of test-taking are correlates of student TTE. In sum, future research can build on the current study by expanding what is known about how cultural, familial, and environmental variables affect the development of test-taking beliefs, which would be expected to proximally influence TTE.

Next, the current findings pointed to the importance of further studying student *cost beliefs* associated with test-taking. Recently, scholars have argued for further consideration of cost in models of motivation and achievement (e.g., Jiang, Rosenzweig, & Gaspard, 2018), given cost beliefs have been a "historically neglected" component of the EEVT in prior research (Flake, Barron, Hulleman, McCoach, & Welsh, 2015, p. 232). A defining feature of cost beliefs in the EEVT is that perceived cost is considered a negative component of motivation, whereas intrinsic value, utility value, and attainment value are each considered positive indicators.

As such, some scholars have suggested that cost might need to be conceptualized as a separate construct, which has resulted in the recent development of *expectancy-value-cost* models (Jiang et al., 2018). Indeed, empirical research suggests cost is more complex than previously thought, and scholars have proposed multi-factor conceptual models of cost value (e.g., Battle & Wigfield, 2003; Perez, Cromley, & Kaplan, 2014). For example, Flake and colleagues (2015) defined four distinct subtypes of cost: *task effort* cost ("negative appraisals of time, effort, or amount of work put forth to engage in the task"), *outside effort* cost ("negative appraisals of time, effort, or amount of work put forth for task other than the task of interest"), *loss of valued alternatives* cost ("a negative appraisal of what is given up as a result of engaging in the task of interest") and *emotional* cost ("negative appraisals of a psychological state that results from exerting effort for the task") (p. 237).

If cost perceptions are truly multi-dimensional as posited by Flake and colleagues (2015), it is possible the measure of cost in the current study did not adequately measure this construct. The two cost items derived from Conley (2012) most closely align with the *loss of valued alternatives cost* dimensions, but it is possible students vary more in *task effort cost* or *emotional cost* perceptions. The demands–capacity model suggests the RD of a test item is a primary determinant of TTE, and research has shown perceived mental taxation of an item is negatively associated with the amount of TTE students exhibit (Wolf et al., 1995). Thus, researchers should consider exploring the relationships between individual sub-types of cost beliefs and student TTE. It is possible an expectancy–value–cost model of TTE could have even greater explanatory power than an expectancy–value model, given that cost has been found to predict more variance in school disengagement than expectancies and values alone (e.g., Perez et al., 2014).

Because low TTE is an example of a maladaptive academic outcome, a greater focus on cost (which has a negative valence) might be an important next step. Further research on the emotional costs of testing would be consistent with Wise and Smith's (2011) proposition that test anxiety might be a key determinant of a student's EC during test-taking. Indeed, we know that test anxiety is a multi-faceted constellation of maladaptive affective, behavioral, and cognitive responses to test-taking (Pekrun & Stephens, 2015), and it could be critical for future researchers to consider whether disengagement from test-taking could be an example of avoidance-oriented coping (see Spangler, Pekrun, Kramer, & Hofmann, 2002) resulting from an individual's inability to regulate negative emotions experienced during a test (i.e., emotional cost).

Finally, another critically important area for future empirical work would be applied research on the potential effectiveness of motivational interventions that are designed to promote student TTE and prevent low TTE from occurring on low-stakes assessments. Given the current finding that value beliefs were the only motivational variables significantly associated with low TTE, further research may need to focus on the effects of value-oriented interventions specific to academic test-taking. In fact, there have been several empirical studies that have documented the effectiveness of targeted motivational interventions that are rooted in the contemporary EEVT (see Lazowski & Hulleman, 2016). Recently, motivational interventions have been designed to specifically target the various subtypes of perceived task value (e.g., Canning et al., 2018; Gaspard et al., 2015). Value-based interventions have been found to reduce achievement gaps between different racial/ethnic groups (Harackiewicz, Canning, Tibbetts, Priniski, & Hyde, 2016) and differentially support student achievement according to gender (Rozek, Hyde, Svoboda, Hulleman, & Harackiewicz, 2015).

Given the clear need for strategies that support student TTE, researchers should study whether motivationally supportive interventions could be effective for reducing the proportion of K–12 who exhibit low TTE. Research has shown brief, psychosocial interventions can change student beliefs about themselves as learners and can have long-term benefits for their academic and social-emotional wellbeing (Walton & Cohen, 2011). These value interventions have included reflections about the personal relevance of course content and brief essay-writing assignments (e.g., Harackiewicz et al., 2016), prompts to generate a list of a few successful peers (e.g., Walton & Cohen, 2007), videos and course discussions related to academic motivation and performance (e.g., Struthers & Perry, 1996), direct communication about the utility value of an activity (e.g., Durik & Harackiewicz, 2007), and sharing online resources designed to normalize feelings of anxiety about beginning a new academic program (e.g., Walton & Cohen, 2007). Experimental research has demonstrated the efficacy of such value interventions for improving the self-reported intrinsic value, utility value, and attainment value of college participants, and there is evidence to suggest these interventions are especially helpful for students from lower SES families and students from racial/ethnic minority groups (Harackiewicz et al., 2016). The existing interventions have been focused on general academic motivation, but the findings from this study point to a critical need for instructional strategies and motivational interventions that are specifically designed to target *test-taking motivation* in order to reduce disparities in TTE.

Implications for Practice

From a practical perspective, there are two critical implications for test developers and educators: 1) prevention of low TTE and 2) identification and response to low TTE. The current findings suggested that middle school students generally find reading tests like STAR Reading to be uninteresting, unenjoyable, moderately useful, and costly compared to other activities.

In light of these findings, practitioners must consider ways to support student's beliefs about the value of test-taking. First, test developers would benefit from considering the issue of student TTE more carefully when designing and evaluating new tests. Although best practice in educational testing indicates the importance of validity evidence based on response patterns (AERA et al., 2014), this type of validity evidence is rarely reported by test developers. The current study further demonstrated how Wise and Kong's (2005) RTE approach could be applied as an efficient and non-obtrusive method for evaluating if a CBT is vulnerable to frequent RGB by test-takers. Accordingly, future educational assessments could include features intended to help prevent students from exhibiting low TTE. Wise and colleagues (2006) showed that a brief warning message triggered by multiple instances of RGB could successfully prevent students from RGB on subsequent items. This effort-monitoring feature could be incorporated into other CBTs as a strategy for supporting TTE while students are still completing the test. Similarly, tests could be designed to flag individual test scores that are potentially invalid immediately. Like the validity indices automatically calculated for computer-based rating scales for emotional and behavioral disorders, it would be reasonable to consider using RTE as a validity indicator for CBTs. Doing so could help their teachers determine whether additional testing might be needed in order to gather more accurate data. In fact, one test system adopted this approach and created guidelines for how teachers can respond in the moment to disengaged students (NWEA, 2018).

Second, the current findings suggest that students might benefit from having the utility value of testing be made more salient to them, given that students reported low utility value beliefs in general. Previous research has shown that modifying test instructions to emphasize the importance of demonstrating good effort on the test results in higher scores and lower rates of low TTE (Brown & Walberg, 1993; Liu et al., 2015).

Sessoms and Finney (2015) recommended that students should be made aware of the usage and value of a test in the weeks before the actual test administration, and so providing motivationally supportive instructions might be a simple preventative strategy. Relatedly, there may be a need for teachers to provide direct "motivational instruction" (Liu et al., 2015) to support the test-taking motivation and subsequent performance of students, particularly in the case of low-stakes testing. Research suggests many college students do not report understanding the purpose of standardized tests they took during their K–12 education (Zilberberg et al., 2014), and so it is questionable whether teachers are making their students aware of how testing is a useful part of their education. As previously noted, it is possible that motivational interventions could potentially help to increase the perceived value of test-taking. Teachers might explicitly share how the test will be used to help students with their academic growth, invite students to reflect on the personal relevance of the tested skills for their future academic and professional goals, include students in personal goal-setting related to their test scores, and encourage students to articulate the importance of giving their best effort in school.

Finally, this study highlights the importance of Wise's (2015) ISV process. Given the observed prevalence of RGB on low-stakes academic testing, educators must pay close attention to the potential adverse effects of low TTE on test validity. ISV is a systematic approach for identifying and responding to low TTE, and practitioners should consider applying Wise and Kong's (2005) RTE approach when using CBT data to make important educational decisions. Research has shown that practical applications of RTE can inform data-based decisions related to program evaluation (Wise & DeMars, 2010), test fraud (Wise, Ma, & Theaker, 2014), teacher evaluation (Wise et al., 2012), and estimation of growth scores (Wise, 2015).

Given that educators do not always make accurate inferences about the reasons why students achieve low test scores (VanDerHeyden, Witt, & Naquin, 2003), practitioners should consider Wise's (2015) assertion that addressing TTE is an issue of professional ethics. Critical evaluation of the distinction between "can't do" and "won't do" problems can help educators to make better decisions based on student testing data (VanDerHeyden, Witt, & Gilbertson, 2007). This study suggests that the ISV process might be a reasonable way to inform such decisions.

Limitations

Several limitations to the current investigation must be acknowledged, among which are issues concerning the research design, external validity, measurement, and statistical analyses in the present empirical study. It will be necessary for future research to address these limitations.

First, the correlational nature of this research study precludes making inferences about causality. Specifically, the significant associations among student demographic characteristics, test-taking beliefs, and the likelihood of exhibiting low TTE may have been confounded by other variables that were not examined in this study. Indeed, the EEVT would imply that several additional factors might contribute to the presence of low TTE, including cultural and family characteristics, behaviors of key socializers, individual goals and self-schemata, and affective reactions to previous achievement outcomes. In particular, no variable was readily available as a control for general academic ability or previous reading achievement. As such, it is possible that group differences in achievement could have mediated or moderated relationships between demographic variables and low TTE. Specifically, the smaller group size for students in grade 8 could potentially suggest that many of the eighth-grade students who took this test were lower performing students who were receiving supplemental reading support. Thus, the correlational findings must be interpreted with caution, recognizing the potential for confounding variables.

Second, questions remain about the external validity of the current investigation. Due to the limited number of empirical studies of TTE in K–12 schools, it is unclear how the current findings might generalize to different grade levels or different assessments. Additionally, the current operationalization of the outcome variable of interest, low TTE, was selected based on the previous literature, but there is currently no empirical justification for using the 0.90 criterion as a flag for low TTE. Thus, one must exercise caution when comparing the current results with those of other studies of student TTE that operationalized low TTE another way. Also, Study II used a convenience sample of students from a specific geographical region, rather than a random, nationally representative sample. Therefore, the findings of Study II might not reflect the beliefs reported by students from different schools or with different demographic characteristics.

Third, measurement limitations must be acknowledged. One potential issue was the use of self-report in Study II. It is possible students demonstrated social desirability bias when taking the survey, and it is also reasonable to question whether students who showed low TTE on the STAR Reading test might have also answered quickly or carelessly when completing the SPOTS (which also had no consequences for them). Another concern is that the sample size for Study II was limited by the number of survey responses that could not be linked to a corresponding STAR Reading test ID and were excluded from subsequent analyses. Because all PII had been removed and teachers were not contacted for their input, the researcher could not further investigate why some of the participants entered ID numbers that were not unique or entered no ID number. It is possible these students entered their ID number incorrectly or that some students indeed had an identical ID number. Regardless, it would be reasonable to question whether some of the students whose data could not be matched had shown low TTE, in which case the subset of participants retained in the dataset might not have been representative of the whole sample.

Another concern related to measurement was that the items in the survey were adapted from the original measures, and so the validity of the SPOTS has yet to be proven. Specifically, the measure of test-taking expectancy beliefs was adapted from an academic efficacy scale. Given that expectancy beliefs and general academic self-efficacy are distinct constructs, it is unclear if the SPOTS items measured test-taking perceptions specific to the STAR Reading test (as opposed to general perceptions of academic ability or reading skill). A related concern is whether the scale measured *current* expectancy beliefs rather than attributions for *past* successes or failures. It is uncertain whether students were truly thinking about taking the test in the future. Further empirical work focused on the measurement of test-taking perceptions is warranted.

Next, the current study was limited by the statistical procedures that were employed. The significance threshold of 0.05 was selected based on the conventional *p*-value used in the extant research literature and recommended by experts in statistical methods for the social sciences (Agresti & Finlay, 2009). However, other scholars have recently argued that the default *p*-value should be decreased (e.g., Benjamin, Berger, Johannesson, Nosek, Wagenmakers, et al., 2018). As such, a more stringent criterion for statistical significance would have yielded nonsignificant results for some of the relationships investigated in Study II. Additionally, the statistical analyses used in Study II did not allow for a complete investigation of the hypothesized relationships between student-level characteristics, test-taking beliefs, and low TTE. That is, this study did not address potential mediation effects or employ structural equation modeling to analyze the extent to which the predictor variables might be directly or indirectly related to the outcome variable of interest. Thus, further exploration of these relationships will add to our understanding of how the components described in the Wise–Smith (2011) model of TTE could relate to one another.

A final limitation of this investigation warrants further discussion. Specifically, the data suggested that one of the primary independent variables of interest, LD status, did not appear to be measured reliably based on the data provided by the test company. According to the National Center for Learning Disabilities (Cortiella & Horowitz, 2014), about 2.4 million students in the United States (5%) have an identified LD. However, the results from Study I indicated that fewer than 0.2% of students in the sample had the LD status variable endorsed as 1 ("Yes"). The LD status of the remaining students (N = 570,899) could not be identified with an adequate certainty.

Accordingly, there are several reasons why the findings related to LD status and TTE should be interpreted with caution. First, the LD status variable in the dataset did not specify whether students had a specific LD in the area of reading (as opposed to writing or mathematics). The research hypothesis that students with LD would be more likely to exhibit low TTE on the STAR Reading test was predicated on the notion students with identified difficulties in reading would be more likely to hold maladaptive motivational beliefs related to reading tests, resulting in lower TTE. However, it is possible that some of the students in the LD group had a specific LD in another academic area, in which case it is possible that reading was actually an area of strength for those students. This is merely speculation given that additional information about the specific educational disabilities of students in this group could not be accessed. For this reason, it is unclear to what extent the present findings about LD status might be generalizable to the larger population of students with LD in elementary and middle schools across the country. Next, the estimated prevalence of low TTE for students with an LD was derived from a relatively small sample of students (N = 487) when compared to the estimates for other groups. If data on LD status had been provided for a higher number of students, then there would have been stronger evidence that the observed prevalence (5.13%) accurately represented this subgroup of students.

Finally, the LD status variable was excluded from all of the logistic regression analyses because more than 99% of cases were treated as having missing data for the LD status variable. As such, the present study did not directly test whether LD status is significantly associated with the odds of exhibiting low TTE when controlling for other demographic variables. If students with LD indeed have a higher likelihood of being identified with low TTE, then it might be especially critical to consider TTE when making educational decisions based on the testing data from SWD. As future research continues to inform the Wise–Smith (2011) demands–capacity model of TTE, researchers are encouraged to consider disability status as a potentially important student-level correlate of TTE. Additional research on this issue will be necessary.

Conclusion

Cronbach (1946) cautioned test users to consider what the response sets of test-takers might suggest about the validity of their scores, and this issue remains true seven decades later. With this in mind, researchers and practitioners will need to be innovative about how to promote student TTE and optimally use test data. The current study extended previous research on TTE by examining the relationships among student-level variables, test-taking beliefs, and TTE on a low-stakes test. As expected, students from various subgroups showed differences in their TTE, and certain malleable motivational variables were significantly related to the odds of exhibiting low TTE. Overall, results pointed to the potential importance of theoretically-driven motivational strategies focused on increasing the perceived value of testing and reducing the perceived cost.

APPENDICES

APPENDIX A.

TABLES AND FIGURES.

Table 1.

Studies Measuring Test-Taking Effort Using Response Time Effort

Study	*N*	Grade	*M* (RTE)	Low TTE[a]
Studies in higher education contexts				
Wise & Kong (2005)	472	College	—	7.4
Wise, Bhola, & Yang (2006)	435	College	—	11.2
Wise & DeMars (2006)	524	College	0.94	12.2
Kong, Wise, Harmes, & Yang (2006)	714	College	0.52–0.71[b]	—
DeMars (2007)	981	College	0.85–1.00	—
Kong, Wise, & Bhola (2007)	488	College	0.93–0.95	9.4–11.9
Wise & Cotten (2009)	802	College	0.94	9.0
Wise, Pastor, & Kong (2009)	386	College	0.90	24.1
Wise & DeMars (2010)	706	College	0.943–0.996	0.6–11.0
Swerdzewski, Harmes, & Finney (2011)	303	College	—	35.6
Setzer, Wise, van den Heuvel, & Ling (2013)	10,004	College	0.987	2.9
Rios, Liu, & Bridgeman (2014)	132	College	—	23.5
Studies in K–12 contexts				
Wise, Kingsbury, Thomason, & Kong (2004)	2382	6–10	0.996	1.1
Wise, Ma, Kingsbury, & Hauser (2010)	711,831	3–9	0.971–0.998	0.2–11.9
Wise & Ma (2012)	573,951	3–9	0.960–0.993	—
Wise, Ma, & Theaker (2014)	26,879	3–8	0.987–0.997	1.4–7.3
Wise (2015)	18,039	9	—	11.5–11.9
Wise & Kingsbury (2016)	285,230	2–12	0.99	1.95

Note. Mean RTE represents the total proportion of solution behavior across all item responses. [a]Low TTE represents the percentage of examinees with RTE scores at or below 0.90. [b]A range of values indicates results were disaggregated by subgroup, assessment, or RTE threshold.

Table 2.

Demographic Information for Sample (Study I)

Variable	*n*	Percent
Full Sample	572,847	100.00
Grade		
4th	345,971	60.40
8th	226,876	39.60
Gender		
Male	279,608	48.81
Female	279,443	48.78
Unknown	13,796	2.41
Race/Ethnicity		
White	211,548	36.93
Hispanic	97,261	16.98
Black	92,952	16.23
Asian/Pacific Islander	20,803	3.63
American Indian/Alaskan Native	5,806	1.01
Other Race/Ethnicity	10,147	1.77
Unknown	134,330	23.45
Learning Disability Status		
Student with Learning Disability	491	0.09
Unknown	572,356	99.91

Table 3.

Distribution of RTE Scores for Sample (Study I)

RTE Score	Number of SB	RTE Score	*n*	Percent
Total			572,847	100.00
Missing			1,461	0.26
RTE > 0.90			555,136	96.91
	34	1.0	512,688	89.50
	33	0.97	24,047	4.20
	32	0.94	11,314	1.98
	31	0.91	7,087	1.24
RTE < 0.90			16,250	2.84
	30	0.88	4,723	0.82
	29	0.85	3,273	0.57
	28	0.82	2,430	0.42
	27	0.79	1,749	0.31
	26	0.76	1,226	0.21
	25	0.74	887	0.15
	24	0.71	663	0.12
	23	0.68	442	0.08
	22	0.65	309	0.05
	21	0.62	210	0.04
	20	0.59	137	0.02
	19	0.56	89	0.02
	18	0.53	42	0.01
	17	0.50	46	0.01
	16	0.47	15	0.00
	15	0.44	6	0.00
	14	0.41	1	0.00
	13	0.38	2	0.00
	12	0.35	0	0.00
	< 12	–	–	–

Table 4.

Mean RTE Scores by Subgroup (Study I)

Variable	*n*	Mean	*SD*
Full Sample	572,847	0.9912	0.0357
Grade			
4th	345,971	0.9922	0.0340
8th	226,876	0.9897	0.0382
Gender			
Male	279,608	0.9880	0.0419
Female	279,443	0.9943	0.0280
Race/Ethnicity			
White	211,548	0.9928	0.0317
Hispanic	97,261	0.9905	0.3743
Black	92,952	0.9869	0.0442
Asian/Pacific Islander	20,803	0.9961	0.0224
American Indian/Alaskan Native	5,806	0.9922	0.0327
Other Race/Ethnicity	10,147	0.9912	0.0356
Learning Disability Status			
Student with Learning Disability	487	0.9836	0.0494
Unknown	570,899	0.9912	0.0357

Table 5.

RTE Scores by Grade and Gender (Study I)

Variable	*n*	Mean	*SD*
Full Sample	572,847	0.9912	0.0357
4th Grade			
Male	126,432	0.9897	0.0393
Female	125,379	0.9947	0.0274
8th Grade			
Male	82,127	0.9853	0.0458
Female	82,813	0.9938	0.0290

Table 6.

RTE Scores by Grade and Race/Ethnicity (Study I)

Variable	*n*	Mean	*SD*
Full Sample	572,847	0.9912	0.0357
4th Grade			
White	125,500	0.9940	0.0293
Hispanic	59,887	0.9917	0.0349
Black	56,202	0.9878	0.0432
Asian/Pacific Islander	13,088	0.9966	0.0208
8th Grade			
White	85,500	0.9911	0.0347
Hispanic	36,982	0.9886	0.0411
Black	36,370	0.9856	0.0455
Asian/Pacific Islander	7,647	0.9953	0.0248

Table 7.

RTE Scores by Gender and Race/Ethnicity (Study I)

Variable	*n*	Mean	*SD*
Full Sample	572,847	0.9912	0.0357
Male			
White	104,439	0.9902	0.0374
Hispanic	48,304	0.9868	0.0443
Black	45,447	0.9825	0.0510
Asian/Pacific Islander	10,369	0.9946	0.0265
Female			
White	102,908	0.9955	0.0244
Hispanic	48,234	0.9942	0.0286
Black	46,759	0.9911	0.0358
Asian/Pacific Islander	20,660	0.9977	0.0171

Table 8.

Proportion Identified with Low TTE by Subgroup (Study I)

Variable	*n*	RTE < 0.90	Percent
Full Sample	571,386	16,250	2.84
Grade			
4th	344,945	8,749	2.54
8th	226,441	7,501	3.31
Gender			
Male	278,872	11,016	3.95
Female	278,742	4,930	1.79
Race/Ethnicity			
White	211,070	4,784	2.27
Hispanic	96,957	2,982	3.08
Black	92,682	4,012	4.33
Asian/Pacific Islander	20,738	252	1.22
American Indian/Alaskan Native	5,798	157	2.71
Other Race/Ethnicity	10,117	283	2.80
Learning Disability Status			
Student with Learning Disability	487	25	5.13
Unknown	570,899	16,225	2.84

Table 9.

Demographic Information for Students with Low TTE (Study I)

Variable	*n*	Percent
Full Sample	16,250	100.00
Grade		
4th	8,749	53.84
8th	7,501	46.16
Gender		
Male	11,016	67.79
Female	4,930	30.34
Missing	304	1.87
Race/Ethnicity		
White	4,784	29.44
Hispanic	2,982	18.35
Black	4,012	24.69
Asian/Pacific Islander	252	1.55
American Indian/Alaskan Native	157	0.97
Other Race/Ethnicity	283	1.74
Missing	3,780	23.26
Disability Status		
Student with Learning Disability	25	0.15
Unknown	16,225	99.85

Table 10.

Results of Multiple Logistic Regression Model (Study I)

	B	*SE*	Wald	*df*	Sig.	Exp(*β*)	95% *CI*
Constant	-4.396	0.022	39,248.388	1	0.000	0.012	–
Gender (Male)	0.834	0.020	1,724.773	1	0.000	2.303	2.214–2.395
Race/Ethnicity (Hispanic)	0.321	0.024	181.347	1	0.000	1.378	1.315–1.444
Race/Ethnicity (Black)	0.681	0.022	963.624	1	0.000	1.976	1.893–2.063
Race/Ethnicity (Asian/Pacific Islander)	-0.628	0.065	92.693	1	0.000	0.533	0.469–0.606
Grade (8th)	0.309	0.019	272.183	1	0.000	1.361	1.312–1.412

Table 11.

Demographic Information for Sample (Study II)

Variable	*n*	Percent
Full Sample	675	100.00
Grade		
7^{th}	330	48.89
8^{th}	345	51.11
Gender		
Male	319	47.26
Female	356	52.74

Table 12.

Distribution of RTE Scores for Sample (Study II)

RTE Score	Number of SB	RTE Score	*n*	Percent
Total			675	100.00
Missing				
RTE > 0.90			633	93.78
	34	1.0	557	82.52
	33	0.97	41	6.07
	32	0.94	17	2.52
	31	0.91	18	2.67
RTE < 0.90			42	6.22
	30	0.88	15	2.22
	29	0.85	7	1.04
	28	0.82	3	0.44
	27	0.79	4	0.59
	26	0.76	2	0.30
	25	0.74	1	0.15
	24	0.71	2	0.30
	23	0.68	2	0.30
	22	0.65	1	0.15
	21	0.62	0	0.00
	20	0.59	2	0.30
	19	0.56	0	0.00
	18	0.53	0	0.00
	17	0.50	1	0.15
	16	0.47	0	0.00
	15	0.44	0	0.00
	14	0.41	2	0.30
	< 14	0.38	0	0.00

Table 13.

Proportion Identified with Low TTE by Subgroup (Study II)

Variable	*n*	RTE < 0.90	Percent
Full Sample	675	42	6.22
Male	319	30	9.40
7th	154	18	11.69
8th	165	12	7.27
Female	356	12	3.37
7th	176	8	4.55
8th	180	4	2.22

Table 14.

Demographic Information for Students with Low TTE (Study II)

Variable	*n*	Percent
Full Sample	42	100
Grade		
7th	26	61.90
8th	16	38.10
Gender		
Male	30	71.43
Female	12	28.57

Table 15.

Descriptive Statistics for SPOTS Items (Study II)

Student Perceptions of Testing Survey Item	*M*	*SD*
Expectancy Item		
1. I'm certain I can answer the questions correctly next time when I take the STAR Reading test.	3.24	0.92
6. I'm certain I can figure out how to answer the most difficult questions next time I take a STAR Reading test.	2.93	1.05
11. I can answer almost all the questions next time I take the STAR Reading test.	3.78	1.10
17. Even if the questions are hard when I take the STAR Reading test, I can answer them correctly.	2.97	0.97
19. I can answer even the hardest questions when I take the STAR Reading test if I try.	3.35	1.14
Interest Value Item		
2. I like taking the STAR Reading test.	2.02	1.03
7. Taking the STAR Reading test is really exciting to me.	1.57	0.90
12. I enjoy taking the STAR Reading test.	1.78	1.02
15. I enjoy taking tests like the STAR Reading test.	1.75	1.00
Attainment Value Item		
3. Being someone who does well on the STAR Reading test is important to me.	3.53	1.11
8. Being good at tests like the STAR reading test is an important part of who I am.	2.55	1.26
13. It is important for me to be someone who is good at taking tests like the STAR Reading test.	3.05	1.20
16. Doing well on tests like the STAR Reading test is an important part of who I am.	2.54	1.25
Utility Value Item		
4. Being good at taking I graduate tests like the STAR Reading test will be useful for what I want to do after and go to work.	3.14	1.23
9. Taking tests like the STAR Reading test will be useful for me later in life.	2.88	1.16
14. Taking tests like the STAR Reading test is valuable because it will help me in the future.	2.92	1.19
18. Being good at taking tests like the STAR Reading test will be important when I get a job or go to college.	3.07	1.21
Relative Cost Item		
5. I have to give up a lot of things I like to do when I take the STAR Reading test.	1.72	1.00
10. Success on the STAR Reading test requires that I give up other activities I enjoy.	1.63	0.95

Table 16.

Descriptive Statistics for SPOTS Subscales (Study II)

SPOTS Subscale	Coefficient α	*M*	*SD*
Test-Taking Expectancy Beliefs	0.808	3.25	0.75
Test-taking Interest Value	0.909	1.76	0.86
Test-Taking Attainment Value	0.865	2.91	1.00
Test-Taking Utility Value	0.891	3.02	1.03
Test-Taking Relative Cost	0.668	1.67	0.84

Table 17.

Mean SPOTS Subscale Scores by Subgroup (Study II)

Variable	Expectancy		Interest		Utility		Attainment		Cost	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Grade										
7th	3.27	0.74	1.80	0.85	3.11	1.04	2.99	0.98	1.72	0.82
8th	3.23	0.77	1.73	0.87	2.93	1.02	2.84	1.01	1.62	0.85
Gender										
Male	3.27	0.78	1.80	0.91	3.06	1.74	2.96	0.98	1.78	0.92
Female	3.24	0.72	1.73	0.82	2.99	0.99	2.86	1.02	1.57	0.75

Table 18.

Bivariate Correlation Matrix for Variables (Study (II)

	Low TTE	Efficacy	Interest	Utility	Attainment	Cost
Low TTE	–					
Efficacy	0-.014 (ns)	–				
Interest	0.027 (ns)	0.355 ($p < 0.001$)	–			
Utility	-0.040 (ns)	0.295 ($p < 0.001$)	0.421 ($p < 0.001$)	–		
Attainment	-0.070 (ns)	0.346 ($p < 0.001$)	0.374 ($p < 0.001$)	0.569 ($p < 0.001$)	–	
Cost	0.105 ($p = 0.006$)	-0.047 (ns)	0.009 (ns)	0.054 (ns)	0.084 ($p = 0.029$)	–

Table 19.

Results of Multiple Logistic Regression Model (Study II)

	B	*SE*	Wald	*df*	Sig.	Exp(β)	95% *CI*
Constant	-3.492	0.617	32.081	1	0.000	0.030	
Gender (Male)	0.985	0.357	7.661	1	0.006	2.677	1.330–5.387
Cost Value	0.441	0.170	6.726	1	0.009	1.554	1.114–2.169
Attainment Value	-0.361	0.169	4.539	1	0.033	0.697	0.500–0.972
Grade (7th)	0.725	0.338	4.599	1	0.032	2.064	1.064–4.004

Figure 1.

Conceptualization of TTE in Demands–Capacity Model of Test-Taking Effort

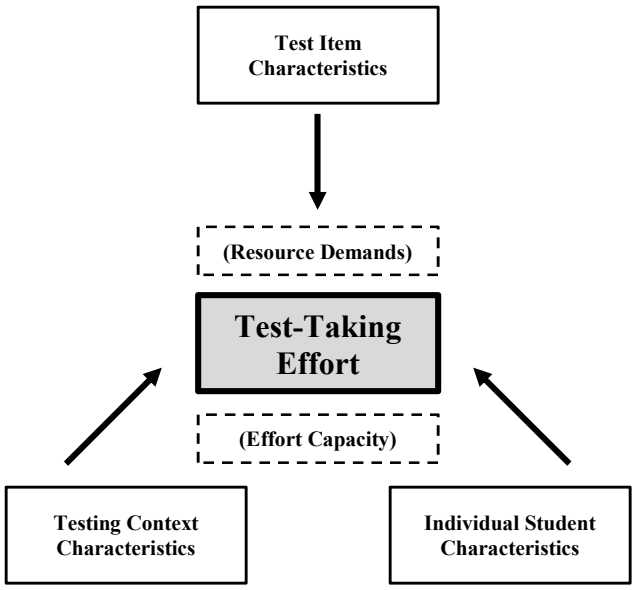

Figure 1. *Conceptualization of TTE in "Demands-capacity model of test-taking effort."* Adapted from "A Model of Examinee Test-Taking Effort" by S. L. Wise and L. F. Smith, 2011, in "High-Stakes Testing in Education: Science and Practice in K–12 Settings" by J. A. Bovaird, K. F. Geisinger, and C. W. Buckendahl (Eds.), p. 149.

Figure 2.

Relationships from EEVT Examined in Current Study.

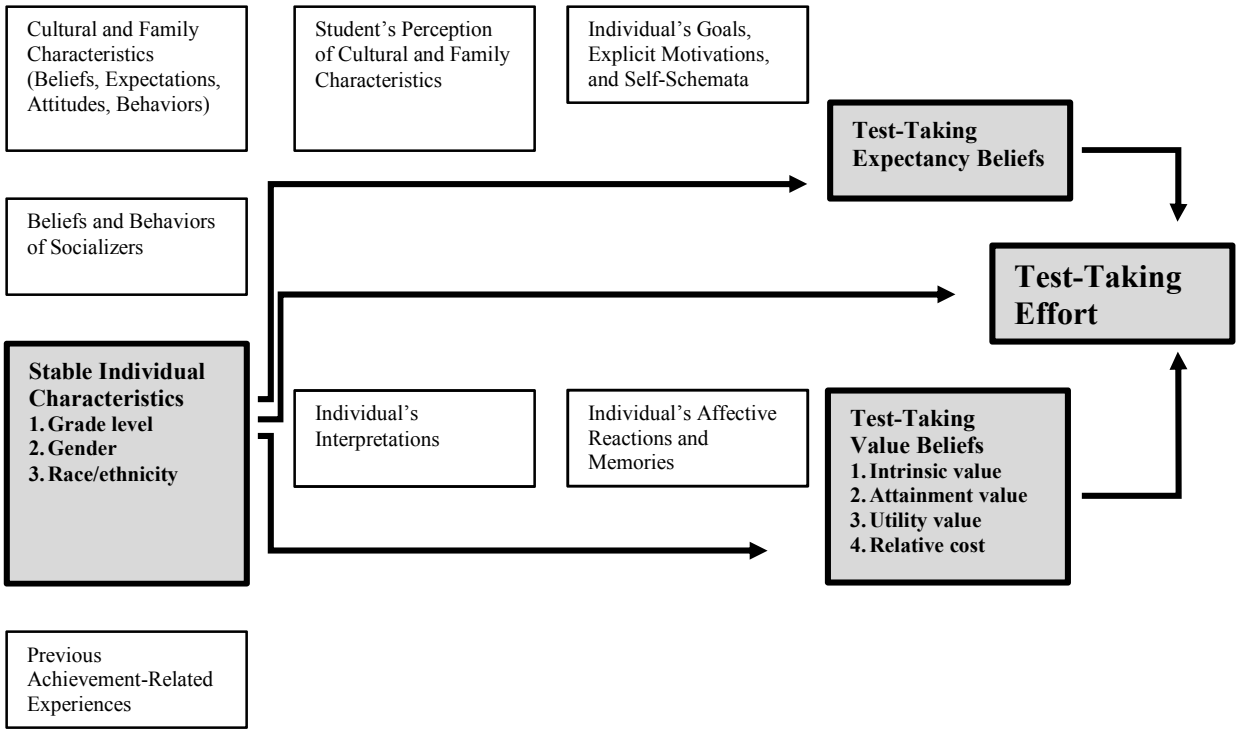

Figure 2. *Relationships from EEVT Examined in Current Study*. Informed by the Eccles et al. expectancy–value model from "Part I Commentary: So What Is Student Engagement Anyway?" by J. Eccles and M.-T. Wang, 2012, in "Handbook of Research on Student Engagement" by S. L. Christenson, A. L. Reschly, and C. Wylie (Eds.), p. 143.

APPENDIX B.

LETTER TO TEST DEVELOPERS.

MICHIGAN STATE
UNIVERSITY

[Date]

James Los, M.A.
Michigan State University
620 Farm Ln., Rm. 447
East Lansing, MI 48824

Dear [Addressees]

My name is James Los, and I am a doctoral student in the School Psychology program at Michigan State University. I am completing a doctoral dissertation, and I am writing to request your permission to conduct a secondary data analysis using STAR Reading assessment data collected during the 2016–2017 and 2017–2018 school years. The purpose of my dissertation research study is to investigate the prevalence and correlates of low test-taking effort (TTE) on a computerized-adaptive reading test. In doing so, I hope to contribute to what we currently know about 1) the extent to which low TTE might be apparent in low-stakes academic testing systems, and 2) the demographic variables and internal motivational variables associated with low TTE.

With a better understanding of *which students* are most likely to show low TTE and the factors related to *why* students exhibit low TTE, it may be possible to develop targeted strategies for promoting more effortful responding during low-stakes testing in schools. Therefore, the primary goals of my doctoral dissertation study are 1) to identify the prevalence of low TTE in students in grades 3–8, 2) to examine whether any demographic subgroups of students are particularly likely to show low TTE, and c) to investigate the extent to which motivational variables (i.e., test-taking expectancy beliefs and value beliefs) relate to low TTE. The specific research questions for my study of TTE on STAR Reading tests are as follows:

1. What proportion of students in grades 3–8 are identified as exhibiting low TTE on a computer adaptive reading test, as determined by Response Time Effort (RTE) score?
2. To what extent do student demographic variables (grade, gender, race/ethnicity) relate to the likelihood that students are identified as exhibiting low TTE on a computer adaptive reading test (as determined by RTE score)?
3. Do students differ in their test-taking expectancy and/or value beliefs based on student demographic variables (grade, gender)?
4. To what extent do student test-taking expectancy and/or value beliefs relate to the likelihood that students are identified as exhibiting low TTE on a computer adaptive reading test (as determined by RTE score)

Study I Purpose and Research Design

The purpose of the first stage of my research study will be to describe the proportion of students in grades 3–8 who exhibit low TTE on a STAR Reading assessment and to examine the relationships between four demographic characteristics and the likelihood of exhibiting low TTE. The four predictor variables are grade level, gender, and race/ethnicity, derived from demographic information entered by educators using the STAR Assessments Renaissance Data Integrator (RDI) service. The dependent variable *low TTE* will be a binary categorical variable defined as Response Time Effort (RTE) derived from a STAR Reading assessment falling at or below 0.90. RTE is an index that represents the proportion of test items on which the examinee exhibited "rapid-guessing behavior" (RB) by submitting a response extremely quickly (i.e., in less three seconds or less). An RTE score of 0.0 would mean all of the student's responses were classified as RGB, whereas an RTE score of 1.0 would mean none of the student's responses were classified as RGB. Therefore, "low TTE" (RTE at or below 0.90) is operationally defined as 10% or more of the student's responses being classified as RGB.

Requested Data for Study I

Using data from administrations of STAR Reading assessments during the 2016–2017 school year, a targeted sample of item-level testing data is requested for inclusion in this study. More specifically, data are requested for a randomly selected sample of 2,500 students in each grade 3–8 (N = 15,000) who took a STAR Reading interim test in the winter of the 2016–2017 school year and for whom the following demographic information and testing data are available:

- Grade level; Gender; Race/ethnicity
- STAR Reading data from a winter screening assessment that includes both item response times and the associated item-level scores (i.e., correct vs. incorrect), and overall score
- Student identification number (non-personally identifiable information)

Study II Purpose and Research Design

The purpose of the second stage of my research study will be to extend upon Study I by allowing researchers to apply the Eccles et al. expectancy–value theory (EEVT) to an empirical study of TTE. The general assumption of the EEVT is student's achievement-related behaviors are guided by their expectancies for success on a task and the extent to which they value the task. This theory will guide a quantitative study testing the relationships between student demographic characteristics, expectancy beliefs, value beliefs, and TTE. Specifically, the second stage of the study is designed to 1) examine the extent to which students differ in their test-taking expectancy and value beliefs by grade and gender and 2) examine the extent to which test-taking expectancy and value beliefs relate to the likelihood students exhibit low TTE on a STAR Reading test.

In my research study, an online survey protocol will be used to gather information on the independent and dependent variables (grade, gender, expectancy beliefs, and value beliefs). The survey will be administered to students from a school district that administers the STAR Reading assessment to students three times per year as part of a district-wide MTSS initiative.

In Study II, I will replicate the method for measuring the dependent variable, *low TTE*, that I will use in Study I (as described above). As in the first stage of the my investigation, TTE will be measured using RTE scores derived from item response times on a STAR Reading assessment. The primary rationale for the second stage of the my research study is that it will allow for the collection of data on student perceptions of test-taking. In doing so, this study may help educators (as well as the test developers) to better understand motivational factors that might relate to the odds a student disengages from low-stakes testing.

Requested Data for Study II

Using data from administrations of STAR Reading assessments during the 2017–2018 school year, a targeted sample of item-level testing data is requested for inclusion in this study. More specifically, data are requested for all students in grades 3–8 in the "City" Public School District in "City," Michigan who complete a STAR Reading interim assessment in the winter of the 2017–2018 school year and for whom the following testing data are available:

- STAR Reading data from a winter screening assessment that includes both item response times and the associated item-level scores (i.e., correct vs. incorrect), and overall score
- Student identification number (non-personally identifiable information)

To allow for my proposed data-matching without disclosing any personally identifiable student information to the researchers, I will request that students' unique school ID numbers be included as variable in the de-identified dataset. This will allow me to match the student survey data with their STAR Reading data by having students enter their unique school ID number prior to completing the online survey about their test-taking perceptions.

Data Retrieval and Storage

Data will be downloaded and stored using high performance computer storage at MSU's Institute for Cyber-Enabled Research (iCER) High Performance Computing Center (HPCC).

If these arrangements meet with your approval, please sign the letter where indicated below and return it to me in the enclosed return envelope. Thank you very much.

Sincerely,

James Los, M.A.
School Psychology Program
Michigan State University

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

_____
[Name of addressee]

Date: _____

APPENDIX C.

RENAISSANCE LEARNING (2014) PRIVACY POLICY NOTICE.

RENAISSANCE LEARNING

Frequently Asked Questions About Student Information in our Software Products

Question: Why does our child's school use Renaissance Learning software products and what does it mean for student information?

Answer: Your child's school has chosen to partner with Renaissance Learning to help improve your child's learning. We are a leading provider of educational solutions to tens of thousands of schools. Our products are used to practice academic skills, guide the learning process, and provide timely student progress information that educators use to improve and personalize your child's instruction. Accelerating learning for all students is our company philosophy and mission, and our products are designed to maintain the privacy of your child's personally identifiable information (PII).

Question: What personal information about my child is stored on your system?

Answer: Our software products only require student name, user name, and school name to function. Some schools input additional demographic data. Renaissance Learning does not collect any personal information from your child. It is collected by your child's school and input into the system by the school. Sometimes we upload this information for the school, if requested.

Question: Does Renaissance Learning store my child's PII in the cloud?

Answer: The "cloud" is a general term that most often relates to services provided over the Internet. Like most software companies today, Renaissance provides access to our products over the Internet. We have a long legacy of operating secure datacenters, and comprehensive security and privacy measures protecting your child's information.

Question: What security measures are in place to keep student information secure?

Answer: Renaissance Learning employs extensive technological and operational measures to ensure security and privacy. A few of these include: advanced security systems technology, regular security audits, physical access controls, privacy training for employees, monitoring of all systems, and segregation of PII into a separate database for each educational institution that purchases our products.

Question: Does Renaissance Learning use my child's personal information for any purpose other than to provide services to his/her school?

Answer: Renaissance Learning does not use your child's PII for any purpose other than to provide services to your child's school. Combined information that has been stripped of PII, and therefore not traceable to any student, is used for research and development so we can continuously improve our products and accelerate learning for all students.

Question: Does Renaissance Learning give away, share with, or sell student information to any third-party organization?

Answer: We do not give away, share, or sell PII. Data sharing (if any) is completely at the control of the educational institutions that purchase our products.

Question: Can parents see or request that their child's records be removed from or changed on a Renaissance Learning system?

Answer: Renaissance Learning provides services under contract to your child's school. We do not own or directly manage any student information. Managing student information is completely in the hands of your child's school. All access to student information is strictly controlled; even Renaissance Learning employees are not authorized to view student information unless requested by the school for customer service purposes.

Question: Who can parents contact about their child's information?

Answer: You must contact your child's school or district directly about your child's information. Renaissance Learning cannot disclose, delete, or make changes to educational records without authorization from the school.

Renaissance Learning™
PO Box 8036 | Wisconsin Rapids, WI 54495-8036
www.renaissance.com

R57975.1114

APPENDIX D.

STUDENT PERCEPTIONS OF TESTING SURVEY.

Part I. Student Assent

Purpose of Research

You are being asked to participate in an online survey of what students think about the STAR Reading test you take at your school. Your school was selected as possible participants in this study because it is one of the schools in Michigan that uses STAR Reading. From this study, the researchers hope to learn about student beliefs about this type of reading test. Your participation in this study will take you about ten minutes.

What You Will Do

What you will do to participate in this study is complete a survey on the computer. You do not need to complete any other testing or school work to participate in this study.

Your Rights to Participate, Say No, or Withdraw

Participation in this research project is completely voluntary. You have the right to say no. You may change your mind at any time and withdraw. You may choose not to answer specific questions or to stop participating at any time. Whether you choose to participate or not will have no effect on your grade or evaluation.

Costs and Compensation for Being in the Study

To thank you for participating in the survey, your class will receive a $50 gift card for a free lunch from the researchers.

Contact Information for Questions and Concerns

If you have concerns or questions about this study, please contact the lead researcher James Los (Address: 620 Farm Ln, East Lansing, MI, 48824; Email: losjames@msu.edu). If you have questions or concerns about your role and rights as a research participant, would like to obtain information or offer input, or would like to register a complaint about this study, you may contact, anonymously if you wish, the Michigan State University's Human Research Protection Program at 517-355-2180, Fax 517-432-4503, or email irb@msu.edu or regular mail at 4000 Collins Rd, Suite 136, Lansing, MI, 48910.

Part II. Documentation of Assent

By entering your student ID number below, you voluntarily agree to participate in this survey.

Part III. Questions About STAR Reading Test

First, please answer a few questions about yourself.

Please select your grade.
- o 7th Grade
- o 8th Grade

Please select your gender.
- o Female
- o Male
- o Prefer not to say

Now we will ask you some questions. For each one, you will answer how true it is for you, using a scale from 1 to 5.

1 means "Not at all true" for you.

3 means "Somewhat true" for you.

5 means "Very true" for you.

Let's practice a few questions.

1	2	3	4	5
Not at all true		Somewhat true		Very true

I like eating pizza.
Playing basketball is fun.
I can travel to the moon after school today.

Now we will ask you some questions about the STAR Reading test. For each one, you will answer how true it is for you, using the same scale from 1 to 5.

1 means "Not at all true" for you.
3 means "Somewhat true" for you.
5 means "Very true" for you.

Remember to think about the STAR Reading test when you answer each of these questions.

1	2	3	4	5
Not at all true		Somewhat true		Very true

1. I'm certain I can answer the questions correctly next time I take the STAR Reading test.
2. I like taking the STAR Reading test.
3. Being someone who does well on the STAR Reading test is important to me.

4. Being good at taking tests like the STAR Reading test will be useful for what I want to do after I graduate and go to work.
5. I have to give up a lot of things I like to do when I take the STAR Reading test.
6. I'm certain I can figure out how to answer the most difficult questions next time I take the STAR Reading test.
7. Taking the STAR Reading test is exciting to me.
8. Being good at tests like the STAR Reading test is an important part of who I am.
9. Taking tests like the STAR Reading test will be useful for me later in life.
10. Success on the STAR Reading test requires that I give up other activities I enjoy.
11. I can answer almost all the questions next time I take the STAR Reading test if I don't give up.
12. I enjoy taking the STAR Reading test.
13. It is important for me to be someone who is good at taking tests like the STAR Reading test.
14. Taking tests like the STAR Reading test is valuable because it will help me in the future.
15. I enjoy taking tests like the STAR Reading test.
16. Doing well on tests like the STAR Reading is an important part of who I am.
17. Even if the questions are hard when I take the STAR Reading test, I can answer them correctly.
18. Being good at taking tests like the STAR Reading test will be important when I get a job or go to college.
19. I can answer even the hardest questions when I take the STAR Reading test if I try.

APPENDIX E.

EXPECTANCY BELIEFS ORIGINAL AND ADAPTED ITEMS.

Patterns of Adaptive Learning Scales (PALS) Academic Efficacy (Midgley et al., 2000)

(Original items)	**Adapted items (5)**
(1. I'm certain I can master the skills taught in class this year.)	**1. I'm certain I can answer the questions correctly next time I take the STAR Reading test.**
(11. I'm certain I can figure out how to do the most difficult class work.)	**6. I'm certain I can figure out how to answer the most difficult questions next time I take STAR Reading test.**
(52. I can do almost all the work in class if I don't give up.)	**11. I can answer almost all the questions next time I take the STAR Reading test if I don't give up.**
(56. Even if the work is hard, I can learn it.)	**17. Even if the questions are hard when I take the STAR Reading test, I can answer them correctly.**
(58. I can do even the hardest work in this class if I try.)	**19. I can answer even the hardest questions when I take the STAR Reading test if I try.**

APPENDIX F.

VALUE BELIEFS ORIGINAL AND ADAPTED ITEMS.

Subjective Task Value Scales (Conley, 2012)

Interest Value

(Original items)	Adapted items (4)
(How much do you like doing math?)	—
(I like math.)	**2. I like taking the STAR Reading test.**
(Math is exciting to me.)	**7. Taking the STAR Reading test is exciting to me.**
(I am fascinated by math.)	—
(I enjoy doing math.)	**12. I enjoy taking the STAR Reading test.**
(I enjoy the subject of math.)	**15. I enjoy taking tests like the STAR Reading test.**

Attainment Value

(Original items)	Adapted items (4)
(Being someone who is good at math is important to me.)	**3. Being someone who does well on the STAR Reading test is important to me.**
(I feel that, to me, being good at solving problems which involve math or reasoning mathematically is (*not at all important* to *very important*).	—
(Being good at math is an important part of who I am.)	**8. Being good at tests like the STAR Reading test is an important part of who I am.**
(It is important for me to be someone who is good at solving problems that involve math.)	**13. It is important for me to be someone who is good at taking tests like the STAR Reading test.**
(It is important for me to be a person who reasons mathematically.)	—
(Thinking mathematically is an important part of who I am.)	**16. Doing well on tests like the STAR Reading test is an important part of who I am.**

Utility Value

(Original items)	Adapted items (4)
(How useful is learning math for what you want to do after you graduate and go to work?)	**4. Being good at taking tests like the STAR Reading test will be useful for what I want to do after I graduate and go to work.**
(Math will be useful to me later in life.)	**9. Taking tests like the STAR Reading test will be useful for me later in life.**
(Math concepts are valuable because they will help me in the future.)	**14. Taking tests like the STAR Reading test is valuable because it will help me in the future.**
(Being good at math will be important when I get a job or go to college.)	**18. Being good at taking tests like the STAR Reading test will be important when I get a job or go to college.**

Cost Value

(Original items)	Adapted items (2)
(I have to give up a lot to do well in math.)	**5. I have to give up a lot of things I like to do when I take the STAR Reading test.**
(Success in math requires that I give up other activities I enjoy.)	**10. Success on the STAR Reading test requires that I give up other activities I enjoy.**

APPENDIX G.

LETTER TO SCHOOL ADMINISTRATORS.

MICHIGAN STATE
UNIVERSITY

[Date]

James Los, M.A.
Michigan State University
620 Farm Ln., Rm. 447
East Lansing, MI 48824

Dear [Addressees]

My name is James Los, and I am a doctoral student in the School Psychology program at Michigan State University. I am completing a doctoral dissertation, and I am writing to request your permission to conduct a secondary data analysis using STAR Reading assessment data collected during the 2017–2018 school years. The purpose of my dissertation research study is to investigate the prevalence and correlates of low test-taking effort (TTE) on a computerized-adaptive reading test. In doing so, I hope to contribute to what we currently know about 1) the extent to which low TTE might be apparent in low-stakes academic testing systems, and 2) the demographic variables and internal motivational variables that are associated with low TTE.

In my proposed study, I intend to collect data about student test-taking beliefs, focusing on student perceptions of their ability to be successful on the test, as well as the extent to which they value succeeding on the test. To measure student TTE, I plan to analyze item response time data from the STAR Reading assessment. Previous research on tests like this has shown us that a small subset of test takers tends to respond with extremely low TTE by submitting their answers rapidly. However, we currently don't know how many students show this type of disengagement from tests they take on the computer at school. This is an important question for educators and researchers alike, as we know that test scores are only meaningful if they reflect students' actual skills or knowledge. For this reason, my goal is to learn more about student test-taking behaviors in assessment contexts that may be perceived as "low stakes" (i.e., no personal consequences) for students. In addition to learning more about how frequently this issue might occur on tests like the STAR Reading, my study is designed to investigate two more questions: 1) *which students* are most likely to show inappropriate effort during testing, and 2) what may be the *reasons why* these students disengage? Answering these questions may help to inform the development of targeted strategies that are designed to prevent low TTE from occurring during academic testing.

Requested Data for Study

With your participation in this study, I would request that the developers of the STAR Assessments (Renaissance Learning) provide me a dataset of test data for all students in grades 6–8 in "City" Public Schools who take a STAR Reading test in the winter of the 2017–2018 school year and for whom the following demographic information and testing data are available:

- STAR Reading data from a winter screening assessment that includes both item response times and the associated item-level scores (i.e., correct vs. incorrect), and overall score
- Randomly generated identification number (non-personally identifiable information)

In addition to requesting the STAR Reading test data, I would also request that I could come to classrooms in grades 6–8 to administer the online survey to students. For me to match student survey data with their STAR Reading test data (to test my research questions) without any personally identifiable information being disclosed, I would request that students enter a unique student ID number prior to their completion of the survey. This way, I could match the de-identified STAR Reading data to student survey responses by linking the datasets with the ID.

Data Retrieval and Storage

Data will be downloaded and stored using high performance computer storage at MSU's Institute for Cyber-Enabled Research (iCER) High Performance Computing Center (HPCC).

Student Perceptions of Testing

In order to examine why some students may show quick, low-effort responses on this test, a survey of student perceptions of testing will be administered to all students in the district who complete this assessment. Understanding why students may show behavioral disengagement during testing is essential for researchers to optimally inform improvements to testing systems. An online survey will be administered to students on the computer in October, after completion of the STAR Reading™ Enterprise fall assessment and before the winter assessment. The items on the survey will measure student perceptions about their perceived interest in testing, perceived importance of testing, perceived usefulness of testing, perceived relative cost of testing, and their expectations for success. The survey will take students approximately ten minutes to complete.

Risks of Participation in Research

There are minimal foreseeable risks associated with participation in this research study. Because no personally identifiable student information (i.e., names, school ID numbers) will be associated with student data, it is anticipated that this study will be approved for an expedited review by the Michigan State University (MSU) Social Science / Behavioral / Educational Institutional Review Board (SIRB). Student participation would be completely voluntary, and student assent would be gathered prior to the students participating in the online survey.

Compensation and Benefits of Participation in Research

There are no direct benefits to students associated with participating in this study. For participating in the study, each class will be compensated with a $50 gift card for a class lunch.

Additionally, the results of this study will be a valuable contribution to the existing research on academic testing. Currently, few studies have directly measured student engagement during testing at the elementary or middle school levels, so this study will be important for informing future research on assessment validity.

Key findings from this study would be presented by the lead researcher to any interested administrators, support staff, and/or classroom teachers as part of an informal meeting or a formal professional development session. This presentation would include the following topics:

- Descriptive statistics about student behavioral engagement during testing overall, as well as data on student engagement during testing at the school, grade, and classroom levels
- Descriptive statistics about student perceptions of testing
- Information about student perceptions associated with disengagement during testing
- Strategies for identifying disengaged test examinees
- Strategies for encouraging engagement during testing
- Strategies for promoting the valid use of academic testing systems

If these arrangements meet with your approval, please sign the letter where indicated below and return it to me in the enclosed return envelope. Thank you very much.

Sincerely,

James Los, M.A.

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

_____

[Name of addressee]

Date: _____

APPENDIX H.

IRB EXEMPT DETERMINATION LETTER.

MICHIGAN STATE
UNIVERSITY

EXEMPT DETERMINATION

March 28, 2018

To: Sara Elizabeth Witmer

Re: **MSU Study ID:** STUDY00000086
Principal Investigator: Sara Elizabeth Witmer
Category: Exempt 1
Exempt Determination Date: 3/28/2018

Title: An Investigation of Test-Taking Effort in a Computer-Adaptive Test of Reading

This project has been determined to be exempt under 45 CFR 46.101(b) 1.

Principal Investigator Responsibilities: The Principal Investigator assumes the responsibilities for the protection of human subjects in this project as outlined in Human Research Protection Program (HRPP) Manual Section 8-1, Exemptions.

Continuing Review: Exempt projects do not need to be renewed.

Modifications: In general, investigators are not required to submit changes to the Michigan State University (MSU) Institutional Review Board (IRB) once a research study is designated as exempt as long as those changes do not affect the exempt category or criteria for exempt determination (changing from exempt status to expedited or full review, changing exempt category) or that may substantially change the focus of the research study such as a change in hypothesis or study design. See HRPP Manual Section 8-1, Exemptions, for examples. If the project is modified to add additional sites for the research, please note that you may not begin the research at those sites until you receive the appropriate approvals/permissions from the sites.

Change in Funding: If new external funding is obtained for an active human research project that had been determined exempt, a new initial IRB submission will be required, with limited exceptions.

Reportable Events: If issues should arise during the conduct of the research, such as unanticipated problems that may involve risks to subjects or others, or any problem that may increase the risk to the human subjects and change the category of review, notify the IRB office promptly. Any complaints from participants that may change the level of review from exempt to expedited or full review must be reported to the IRB. Please report new information through the project's workspace and contact the IRB office with any urgent events. Please visit the Human Research Protection Program (HRPP) website to obtain more information, including reporting timelines.

Office of Regulatory Affairs
Human Research Protection Program

4000 Collins Road
Suite 136
Lansing, MI 48910

517-355-2180
Fax: 517-432-4503
Email: irb@msu.edu
www.hrpp.msu.edu

Personnel Changes: After determination of the exempt status, the PI is responsible for maintaining records of personnel changes and appropriate training. The PI is not required to notify the IRB of personnel changes on exempt research. However, he or she may wish to submit personnel changes to the IRB for recordkeeping purposes (e.g. communication with the Graduate School) and may submit such requests by submitting a Modification request. If there is a change in PI, the new PI must confirm acceptance of the PI Assurance form and the previous PI must submit the Supplemental Form to Change the Principal Investigator with the Modification request (http://hrpp.msu.edu/forms).

Closure: Investigators are not required to notify the IRB when the research study is complete. However, the PI can choose to notify the IRB when the project is complete and is especially recommended when the PI leaves the university.

For More Information: See HRPP Manual, including Section 8-1, Exemptions (available at https://hrpp.msu.edu/msu-hrpp-manual-table-contents-expanded).

Contact Information: If we can be of further assistance or if you have questions, please contact us at 517-355-2180 or via email at IRB@ora.msu.edu. Please visit hrpp.msu.edu to access the HRPP Manual, templates, etc.

Exemption Category. This project has qualified for Exempt Category (ies) 1. Please see the appropriate research category below from 45 CFR 46.101(b) for the full regulatory text. [123]

Exempt 1. Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as (i) research on regular and special education instructional strategies, or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.

Exempt 2. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

Exempt 3. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior that is not exempt under paragraph (b)(2) of this section, if:
(i) the human subjects are elected or appointed public officials or candidates for public office; or (ii) federal statute(s) require(s) without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter.

Exempt 4. Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

Exempt 5. Research and demonstration projects which are conducted by or subject to the approval of department or agency heads, and which are designed to study, evaluate, or otherwise examine: (i) Public benefit or service programs; (ii) procedures for obtaining benefits or services under those programs; (iii) possible changes in or alternatives to those programs or procedures; or (iv) possible changes in methods or levels of payment for benefits or services under those programs.

Exempt 6. Taste and food quality evaluation and consumer acceptance studies, (i) if wholesome foods without additives are consumed or (ii) if a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural chemical or environmental contaminant at or below the level found to be safe, by the Food and Drug Administration or approved by the Environmental Protection Agency or the Food Safety and Inspection Service of the U.S. Department of Agriculture.

[1]Exempt categories (1), (2), (3), (4), and (5) cannot be applied to activities that are FDA-regulated.
[2] Exemptions do not apply to research involving prisoners.
[3] Exempt 2 for research involving survey or interview procedures or observation of public behavior does not apply to research with children, except for research involving observations of public behavior when the investigator(s) do not participate in the activities being observed.

APPENDIX I.

LETTER TO PARENTS.

MICHIGAN STATE
UNIVERSITY

[Date]

James Los, M.A.
Michigan State University
620 Farm Ln., Rm. 447
East Lansing, MI 48824

Dear parents,

Students in your child's classroom have been invited to participate in a research study being conducted by a doctoral student from Michigan State University's College of Education.

Researchers are interested in learning more about how schools use the STAR Reading test and how students in elementary and middle schools view these tests. The purpose for studying this topic is to help test developers and educators understand how best to use reading tests in schools.

Because your school district already requires students to complete this assessment, no additional testing will be necessary. Instead, the researchers have requested permission from administrators to collect and analyze student testing data (with no identifying information) from Renaissance Learning™, the test developers. Again, no personally identifiable information will be disclosed.

The second part of this study involves inviting students to complete a brief, online survey about their perceptions of the STAR Reading test. **Participation is voluntary, and the survey is completely anonymous.** Students will be asked to enter their individual school ID number before they complete the survey, and the researchers will never be able to connect this number to your child's identifying information. Students' responses to these questions will not be shared with anyone except the researchers who are gathering this information. If there are any questions that students do not want to answer, they may choose not to respond. Again, students are not required to complete this survey, and they can choose to withdraw from the survey at any time.

To thank students for participating in the survey, your child's class will receive a free lunch from the researchers.

On the following page, please find attached a copy of the assent form your child will be read prior to being asked to take the survey. If you have any questions about the study, please contact the lead researcher, James Los, at the email address provided below. Thank you very much!

Sincerely,

James Los, M.A

losjames@msu.edu

Purpose of Research

You are being asked to participate in an online survey of what students think about the STAR Reading test you take at your school. Your school was selected as possible participants in this study because it is one of the schools in Michigan that uses STAR Reading. From this study, the researchers hope to learn about student beliefs about this type of reading test. Your participation in this study will take you about ten minutes.

What You Will Do

What you will do to participate in this study is complete a survey on the computer. You do not need to complete any other testing or school work to participate in this study.

Your Rights to Participate, Say No, or Withdraw

Participation in this research project is completely voluntary. You have the right to say no. You may change your mind at any time and withdraw. You may choose not to answer specific questions or to stop participating at any time. Whether you choose to participate or not will have no effect on your grade or evaluation.

Costs and Compensation for Being in the Study

To thank you for participating in the survey, your class will receive a $50 gift card for a free lunch from the researchers.

Contact Information for Questions and Concerns

If you have concerns or questions about this study, please contact the lead researcher James Los (Address: 620 Farm Ln, East Lansing, MI, 48824; Email: losjames@msu.edu). If you have questions or concerns about your role and rights as a research participant, would like to obtain information or offer input, or would like to register a complaint about this study, you may contact, anonymously if you wish, the Michigan State University's Human Research Protection Program at 517-355-2180, Fax 517-432-4503, or email irb@msu.edu or regular mail at 4000 Collins Rd, Suite 136, Lansing, MI, 48910.

REFERENCES

REFERENCES

Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools, 45*(5), 369–386. doi:10.1002/pits.20303

Archambault, I., Eccles, J. S., & Vida, M. N. (2010). Ability self-concepts and subjective value in literacy: Joint trajectories from grades 1 through 12. *Journal of Educational Psychology, 102*(4), 804–816. doi:10.1037/a0021075

Archambault, I., Janosz, M., Morizot, J., & Pagani, L. (2009). Adolescent behavioral, affective, and cognitive engagement in school: Relationship to dropout. *Journal of School Health, 79*(9)*,* 408 – 415. doi:10.1111/j.1746-1561.2009.00428.x

Bailey, S. M. (1993). The current status of gender equity research in American schools. *Educational Psychologist, 28*(4), 321–339. doi:10.1207/s15326985ep2804_3

Baird, G. L., Scott, W. D., Dearing, E., & Hamill, S. K. (2009). Cognitive self-regulation in youth with and without learning disabilities: Academic self-efficacy, theories of intelligence, learning vs. performance goal preferences, and effort attributions. *Journal of Social and Clinical Psychology*, *28*(7), 881–908. doi:10.1521/jscp.2009.28.7.881

Baker, L., & Wigfield, A. (1999). Dimensions of children's motivation for reading and their relations to reading activity and reading achievement. *Reading Research Quarterly, 34*(4)*,* 452–477. doi:10.1598/RRQ.34.4.4

Barnard, J. (2015). Implementing a CAT: The AMC experience. *Journal of Computerized Adaptive Testing, 3*(1), 1–12. doi:10.7333/15100301001

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342–363. doi:10.1080/15305058.2010.508569

Battle, J. (1979). Self-esteem of students in regular and special classes. *Psychological Reports, 44*(1), 212–214. doi:10.2466/pr0.1979.44.1.212

Battle, A., & Wigfield, A. (2003). College women's value orientations toward family, career,

and graduate school. *Journal of Vocational Behavior, 62*(1), 56–75. doi:10.1016/s0001-8791(02)00037-4

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441–462. doi:10.1007/bf03173192

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6-10. doi:10.1038/s41562-017-0189-z

Brophy, J. (1999). Toward a model of the value aspects of motivation in education: Developing appreciation for particular learning domains and activities. *Educational Psychologist, 34*, 75–85. doi:10.1207/s15326985ep3402_1

Brown, G. T. L., & Hirschfeld, G. H. F. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy & Practice, 15*(1), 3–17. doi:10.1080/09695940701876003

Brown, G. T. L., Irving, S. E., Peterson, E. R., & Hirschfeld, G. H. F. (2009). *Students' Conceptions of Assessment–Version V*. PsycTESTS Dataset. doi:10.1037/t03968-000

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *The Journal of Educational Research, 86*(3), 133–136. doi:10.1080/00220671.1993.9941151

Bong, M., Cho, C., Ahn, H. S., & Kim, H. J. (2012). Comparison of self-beliefs for predicting student motivation and achievement. *The Journal of Educational Research, 105*(5), 336–352. doi:10.1080/00220671.2011.627401

Box, G. E. P., & Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics, 4*(4), 531. doi:10.2307/1266288

Butkowsky, I. S., & Willows, D. M. (1980). Cognitive-motivational characteristics of children varying in reading ability: Evidence for learned helplessness in poor readers. *Journal of Educational Psychology, 72*(3), 408–422. doi:10.1037/0022-0663.72.3.408

Canning, E. A., Harackiewicz, J. M., Priniski, S. J., Hecht, C. A., Tibbetts, Y., & Hyde, J. S. (2018). Improving performance and retention in introductory biology with a utility-value intervention. *Journal of Educational Psychology, 110*(6), 834–849. doi:10.1037/edu0000244

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*(2), 300–310. doi:10.1037/0021-9010.82.2.300

Chapman, J. W., & Boersma, F. J. (1979). Academic self-concept in elementary learning disabled children: A study with the student's perception of ability scale. *Psychology in the Schools, 16*(2), 201–206. doi:10.1037/h0081208

Christenson, S., Reschly, A., & Wylie, C. (Eds.). (2012). *Handbook of research on student engagement*. New York, NY: Springer Science. doi:10.1007/978-1-4614-2018-7

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*(4), 609–624. doi:10.1016/j.cedpsych.2007.10.002

Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy–value perspectives. *Journal of Educational Psychology, 104*(1), 32–47. doi:10.1037/a0026042

Cortiella, C., & Horowitz, S. (2014). *The state of learning disabilities: Facts, trends, and emerging issues*. New York: National Center for Learning Disabilities. Retrieved from: https://www.ncld.org/wp-content/uploads/2014/11/2014-State-of-LD.pdf

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*(4), 475–494. doi:10.1177/001316444600600405

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harper & Row.

Cronin, J., Bontempo, B., Kingsbury, G. G., Hauser, C., McCall, M., & Houser, R. (2005). *Using item response time and accuracy on a computer adaptive test to predict deflated estimates of performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*(1), 23–45. doi:10.1080/10627190709336946

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research and Practice in Assessment, 8*(2), 69–82.

Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: How individual interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology, 99*(3), 597–610. doi:10.1037/0022-0663.99.3.597

Durik, A. M., Vida, M., & Eccles, J. S. (2006). Task values and ability beliefs as predictors of high school literacy choices: A developmental analysis. *Journal of Educational Psychology, 98*(2), 382–393. doi:10.1037/0022-0663.98.2.382

Eccles (Parsons), J. S. (1984). Sex differences in achievement patterns. In T. Sonderegger (Ed.),

Nebraska Symposium on Motivation (Vol. 32, pp. 97–132). Lincoln, NE: University of Nebraska Press.

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

Eccles, J. S., & Wang, M-Te. (2012). Part 1 commentary: So what is student engagement anyway? In S. Christenson, A. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 133–145). New York, NY: Springer.

Eccles, J. S., Wigfield, A., Flanagan, C. A., Miller, C., Reuman, D. A., & Yee, D. (1989). Self-concepts, domain values, and self-esteem: Relations and changes at early adolescence. *Journal of Personality, 57*(2), 283–310. doi:10.1111/j.1467-6494.1989.tb00484.x

Eccles (Parsons), J. S., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development, 64*(3)*,* 830–847. doi:10.1111/j.1467-8624.1993.tb02946.x

Eccles, J. S., Wigfield, A., & Schiefele, U. (1998). Motivation to succeed. In N. Eisenberg (Ed.), W. Damon (Series Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 1051–1071). New York: Wiley.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*(3), 311–326. doi:10.1080/15305050701438074

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice 17*(4), 345–356. doi:10.1080/0969594X.2010.516569

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015).

Flake, J. K., Barron, K. E., Hulleman, C., McCoach, B. D., & Welsh, M. E. (2015). Measuring cost: The forgotten component of expectancy-value theory. *Contemporary Educational Psychology, 41*, 232–244. doi:10.1016/j.cedpsych.2015.03.002

Fordham, S., & Ogbu, J. U. (1986). Black students' school success: Coping with the "burden of 'acting white.'" *The Urban Review, 18*(3), 176–206. doi: 10.1007/bf01112192

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. doi:10.3102/00346543074001059

Graham, S. (1994). Motivation in African Americans. *Review of Educational Research, 64*(1), 55–117. doi:10.3102/00346543064001055

Graham, S., & Taylor, A. Z. (2002). Ethnicity, gender, and the development of achievement values. In A. Wigfield & J. S. Eccles (Eds.), *A Vol. in the educational psychology series. Development of achievement motivation* (pp. 121-146). San Diego, CA, US: Academic Press. doi:10.1016/B978-012750053-9/50007-3.

Grolnick, W. S., & Ryan, R. M. (1990). Self-perceptions, motivation, and adjustment in children with learning disabilities: A multiple group comparison study. *Journal of Learning Disabilities, 23*(3), 177–184. doi:10.1177/002221949002300308

Guthrie, J. T., Wigfield, A., Barbosa, P., Perencevich, K. C., Taboada, A., Davis, M. H., et al. (2004). Increasing reading comprehension and engagement through concept-oriented reading instruction. *Journal of Educational Psychology, 96*, 403–423. doi:10.1037/0022-0663.96.3.403

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27. doi:10.1111/j.1745-3992.2004.tb00149.x

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology, 111*(5), 745–765. doi:10.1037/pspp0000075

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns. Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18*(3), 133–146. doi:10.1111/j.1745-3984.1981.tb00848.x

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). *Student testing in America's great city schools: An inventory and preliminary analysis*. Washington, DC: Council of the Great City Schools. Retrieved from http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing%20Report.pdf

Hauser, C., & Kingsbury, G. G. (2009). *Individual score validity in a modest-stakes adaptive educational testing setting*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Higgins, E. T. (2007). Value. In A. Kruglanski & E. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 454–472). New York: Guilford Press.

Jacobs, J. E., Lanza, S., Osgood, D., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development, 73*(2)*,* 509–527. doi:10.1111/1467-8624.00421

Jiang, Y., Rosenzweig, E. Q., & Gaspard, H. (2018). An expectancy-value-cost approach in predicting adolescent students' academic motivation and achievement. *Contemporary Educational Psychology, 54*, 139–152. doi:10.1016/j.cedpsych.2018.06.005.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*(4), 606–619. doi:10.1177/0013164406294779

Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006). *Motivational effects of praise in response-time based feedback: A follow-up study of the effort-monitoring CBT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education. *Review of Educational Research, 86*(2), 602–640. doi:10.3102/0034654315617832

Licht, B. G., & Kistner, J. A. (1986). Motivational problems of learning-disabled children: Individual differences and their implications for treatment. In J. K. Torgesen & B. W. L. Wong (Eds.) *Psychological and educational perspective on learning disabilities* (pp. 225–255). Orlando, FL: Academic.

Liu, O., L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment, 20*(2), 79–94. doi: 10.1080/10627197.2015.1028618

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Meece, J. L., Glienke, B. B., & Askew, K. (2009). Gender and motivation. In K. Wentzel & A. Wigfield (Eds.), *Handbook on motivation at school* (pp. 411-432). New York: Routledge, Taylor, and Francis.

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its consequences for young adolescents' course enrollment intentions and performances in mathematics. *Journal of Educational Psychology, 82*(1), 60–70. doi.org/10.1037//0022-0663.82.1.60

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*(1), 72–87. doi:10.1037/1082-989x.8.1.72

Mello, Z. R. (2009). Racial/ethnic group and socioeconomic status variation in educational and

occupational expectations from adolescence to adulthood. *Journal of Applied Developmental Psychology, 30*(4), 494-504. doi:10.1016/j.appdev.2008.12.029.

Michigan Department of Education (2018). *MI School Data*. Lansing, MI: Michigan Department of Education. Retrieved from: https://www.mischooldata.org.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS*). Ann Arbor: University of Michigan.

Murray, C. (2009). Parent and teacher relationships as predictors of school engagement and functioning among low-income urban youth. *The Journal of Early Adolescence, 29*(3), 376–404. doi:10.1177/0272431608322940

National Joint Committee on Learning Disabilities. (2008). *Adolescent Literacy and Older Students With Learning Disabilities* [Technical Report]. Retrieved from www.asha.org/policy

Nering, M. L., Bay, L. G., & Meijer, R. R. (2002). Identifying pattern markers in a large-scale assessment program. *Measurement and Evaluation in Counseling and Development 35*(3), 182–195.

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

Northwest Evaluation Association (2018). *About rapid-guessing and test disengagement*. Portland, OR: NWEA. Retrieved from: https://community.nwea.org/docs/DOC-2964.

Ogbu, J. U. (1992). Understanding cultural diversity and learning. *Educational Researcher, 21*(8), 5–14. doi:10.2307/1176697

O'Neil, Jr., H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress Mathematics performance. *Educational Assessment, 3*(2), 135–157. doi:10.1207/s15326977ea0302_2

Osborne, J. W. (1997). Race and academic disidentification. *Journal of Educational Psychology, 89*(4), 728–735. doi:10.1037/0022-0663.89.4.728

Paris, S. G., Lawton, T. A., Turner, J. C., & Roth, J. L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher, 20*(5), 12–20. doi:10.2307/1176397

Paris, S. G., Turner, J. C., & Lawton, T. A. (1990). *Students' views of standardized achievement tests*. Paper presented at the American Educational Research Association, Boston, MA.

Pekrun, R., & Stephens, E. J. (2015). Test anxiety and academic achievement. *International Encyclopedia of the Social & Behavioral Sciences*, 244–249. doi:10.1016/b978-0-08-097086-8.26064-9

Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology, 106*(1), 315–329. doi:10.1037/a0034027

Pollard, D. S. (1993). Gender, achievement, and African-American students' perceptions of their school experience. *Educational Psychologist*, *28*(4), 341–356. doi.org/10.1207/s15326985ep2804_4

Ratelle, C. F., Guay, F., Larose, S., & Senécal, C. (2004). Family correlates of trajectories of academic motivation during a school transition: A semiparametric group-based approach. *Journal of Educational Psychology, 96*(4)*,* 743–754. doi:10.1037/0022-0663.96.4.743

Raykov, T., & Marcoulides, Ga. A. (2008). *An introduction to applied multivariate analysis*. New York, NY: Routledge Press.

Renaissance Learning. (2014). *Frequently asked questions about student information in our software products*. Wisconsin Rapids, WI: Renaissance Learning.

Renaissance Learning. (2016). *STAR Reading™ Technical Manual*. Wisconsin Rapids, WI: Renaissance Learning.

Renaissance Learning. (2017). *Application and hosting privacy policy (US applications)*. Wisconsin Rapids, WI: Renaissance Learning.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research, 2014*(161), 69–82. doi:10.1002/ir.20068

Robinson, J. P., & Lubienski, S. T. (2010). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*(2), 268–302. doi:10.3102/0002831210372249

Rozek, C. S., Hyde, J. S., Svoboda, R. C., Hulleman, C. S., & Harackiewicz, J. M. (2015). Gender differences in the effects of a utility-value intervention to help parents motivate adolescents in mathematics and science. *Journal of Educational Psychology, 107*(1), 195–206. doi:10.1037/a0036981

Salvia, J. S., Ysseldyke, J. E. & Bolt, S. (2013). *Assessment in special and inclusive education* (12th ed.). Boston, MA: Wadsworth/Cengage Publications.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state

mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232. doi:10.1111/j.1745-3984.1997.tb00516.x

Schunk, D. H., Meece, J. L., & Pintrich, P. R. (2014). *Motivation in education: Theory, research, and applications* (4th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Schunk, D. H., Pajares, F. (2009). Self-efficacy theory. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 35–53). New Work: Routledge/Taylor & Francis Group.

Senko, C. (2016). Learning environments and motivation. In K. R. Wentzel & D. Miele (Eds.), *Handbook of motivation at school* (2nd ed.). (pp. 55–74). New York, NY: Routledge.

Sessoms, J., & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing, 15*(4), 356–388. doi:10.1080/15305058.2015.1034866

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. doi:10.1080/08957347.2013.739453

Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly, 30*(4), 470–487. doi:10.1037/spq0000116

Skinner, E. A., Kindermann, T. A., Connell, J. P., & Wellborn, J. G. (2009). Engagement and disaffection as organizational constructs in the dynamics of motivational development. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 223–245). New Work: Routledge/Taylor & Francis Group.

Skinner, E. A. & Pitzer, J. R. (2012). *Developmental dynamics of student engagement, coping, and everyday resilience*. In S. Christenson, A. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 133–145). New York, NY: Springer.

Smith, L. F., & Smith, J. K. (2002). Relation of test-specific motivation and anxiety to test performance. *Psychological Reports, 91*(3), 1011–1021. doi:10.2466/pr0.2002.91.3.1011

Spangler, G., Pekrun, R., Kramer, K., & Hofmann, H. (2002). Students' emotions, physiological reactions, and coping in academic exams. *Anxiety, Stress, & Coping, 15*(4), 413–432. doi:10.1080/1061580021000056555

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613–629. doi:10.1037/0003-066x.52.6.613

Struthers, C. W., & Perry, R. P. (1996). Attributional style, attributional retraining, and

inoculation against motivational deficits. *Social Psychology of Education, 1*(2), 171–187. doi:10.1007/bf02334731

Sundre, D. L. (1999). *Does examinee motivation moderate the relationship between test consequences and test performance?* (Report No. TM029964). Harrisonburg, Virginia: James Madison University. Retrieved from https://eric.ed.gov/?id=ED432588

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*(1), 6–26. doi:10.1016/s0361-476x(02)00063-2

Sundre, D. L., & Wise, S. L. (2003). *Motivation filtering: An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes Assessment context. *Applied Measurement in Education, 24*(2), 162–188. doi:10.1080/08957347.2011.555217

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Boston: Pearson.

U. S. Department of Education (2015). *Fact sheet: Testing action plan*. Washington, DC: U.S. Department of Education. Retrieved from: https://www.ed.gov/news/press-releases/fact-sheet-testing-action-plan.

U.S. Department of Education: National Center on Response to Intervention (2010). *Tools Charts*. Washington, DC: U.S. Department of Education. Retrieved from: https://rti4success.org/resources/tools-charts.

VanDerHeyden, A. M., Witt, J. C., & Naquin, G. (2003). The development and validation of a process for screening referrals to special education. *School Psychology Review, 32*, 204–227.

VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. (2007). A multi-year evaluation of the effects of a Response to Intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45*(2), 225–256. doi:10.1016/j.jsp.2006.11.004

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology, 92*(1), 82–96. doi:10.1037/0022-3514.92.1.82

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331*(6023), 1447–1451. doi:10.1126/science.1198364

Watt, H. M. (2004). Development of adolescents' self-perceptions, values, and task perceptions according to gender and domain in 7th-through 11th-grade Australian students. *Child development*, *75*(5), 1556–1574. doi.org/10.2307/1131221

Weiner, B. (1992). *Human motivation: Metaphors, theories, and research*. Newbury Park, CA: Sage Publications.

Weiss, D. J. (1983). *Introduction*. In D. J. Weiss (ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York, NY: Academic Press, Inc.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1–27. doi:10.2458/azu_jmmss.v2i1.12351

Wentzel, K. R., & Brophy, J. E. (2014). *Motivating students to learn*. New York, NY: Routledge.

Wentzel, K. R., & Miele, D. B. (Eds.). (2016). *Handbook of motivation in school* (2nd ed.). New York, NY: Routledge.

Wiest, D. J., Wong, E. H., Cervantes, J. M., Craik, L., & Kreil, D. A. (2001). Intrinsic motivation among regular, special, and alternative education high school students. *Adolescence, 36*(141), 111–126.

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review, 12*(3), 265–310. doi:10.1016/0273-2297(92)90011-p

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. doi:10.1006/ceps.1999.1015

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 92–120). San Diego: Academic Press.

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Changes in children's competence beliefs and subjective task values across the elementary school years: A three-year study. *Journal of Educational Psychology, 89*(3), 451–469. doi:10.1037//0022-0663.89.3.451

Wigfield, A., Tonks, S., & Klauda, S. L. (2009). Expectancy–value theory. In K. R. Wentzel, & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 55–75). New York, NY: Routledge/Taylor & Francis Group.

Wigfield, A., Tonks, S., M., & Klauda, S., L. (2016). Expectancy–value theory. In K. R. Wentzel

& D. Miele (Eds.), *Handbook of motivation at school* (2nd ed.). (pp. 55–74). New York, NY: Routledge.

Wigfield, A., & Wentzel., K. R. (2007). Introduction to motivation at school: Interventions that work. *Educational Psychologist, 42*(4), 191–196. doi: 10.1080/00461520701621038

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. doi:10.1207/s15324818ame1902_2

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education, 58*(3), 152–166. doi:10.1353/jge.0.0042

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, *2*(3), 1–17. doi:10.7333/1401-02010001

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*(3), 237–252. doi:10.1080/08957347.2015.1042155

Wise, S. L., Bhola, D. S., & Yang, S.-T. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25*(2), 21–30. doi:10.1111/j.1745-3992.2006.00054.x

Wise, S. L., & Cotten, M. R. (2009). Test-taking effort and score validity. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 187–205). Charlotte, NC: Information Age.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. doi:10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38. doi:10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & DeMars, C. E. (2009). A clarification of the effects of rapid guessing on coefficient alpha: A note on Attali's reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement*, *33*, 488–490. doi:10.1177/0146621607304655

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*(1), 27–41. doi:10.1080/10627191003673216

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement 53*(1), 86–105. doi:10.1111/jedm.12102

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

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., Ma, L., Kingsbury, G. G., & 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.

Wise, S. L., Ma, L., & Theaker, R. A. (2014). Identifying non-effortful student behavior on adaptive tests: Implications for test fraud detection. In N. M. Kingston & A. K. Clark (Eds.) *Test fraud: Statistical detection and methodology* (pp. 175–185). New York, NY: Taylor & Francis/Routledge.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*(2), 185–205. doi:10.1080/08957340902754650

Wise, S. L., & Smith, L. F. (2011). *A model of examinee test-taking effort*. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (pp. 139–153). Washington, DC: American Psychological Association.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*(3), 227–242. doi:10.1207/s15324818ame0803_3

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education*, *8*(4), 341–351. doi:10.1207/s15324818ame0804_4

Ysseldyke, J. E., Burns, M., Dawson, P., Kelley, B., Morrison, D., Ortiz, S., Rosenfield, S., & Telzrow, C. (2006). *School psychology: A blueprint for training and practice III*. Bethesda, MD: National Association of School Psychologists.

Zilberberg, A., Anderson, R. D., Finney, S. J., & Marsh, K. R. (2013). American college students' attitudes toward institutional accountability testing: Developing measures: *Educational Assessment*, *18*(3), 208–234. doi.org/10.1080/10627197.2013.817153

Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. D. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing, 14*(4), 360–384. doi:10.1080/15305058.2014.928301