Summer 2016

Applying solution behavior thresholds to a noncognitive measure to identify rapid responders: An empirical investigation

Mary M. Johnston
James Madison University

Applying Solution Behavior Thresholds to a Noncognitive Measure to Identify Rapid Responders: An Empirical Investigation

Mary M. Johnston

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

August 2016

FACULTY COMMITTEE:

Committee Chair: Dena A. Pastor

Committee Members:

Sara J. Finney

Christine E. DeMars

Acknowledgements

I would like to thank my doctoral advisor and dissertation chair, Dena Pastor. Thank you for your unwavering support, guidance, and encouragement over the course of the last three years and for helping me grow both academically and professionally. I am very grateful to have had the opportunity to work with you. I would also like to thank my dissertation committee members, Sara Finney and Christine DeMars, for the time, effort, and invaluable feedback they provided while working on this project. In addition, I would like to thank the staff and faculty members within the Center for Assessment and Research Studies for their support and guidance. In particular, I want to thank Jeanne Horst and Monica Erbacher Smith. Thank you for making me laugh and feel welcome. Finally, I would like to thank my family and friends. Thank you for keeping me focused and helping me achieve this goal. I could not have done it without your support. Thank you for being there for me at every step.

Table of Contents

List of Tables

List of Figures

Abstract

Noncognitive measures are increasingly being used for accountability purposes in higher education (e.g., O. L. Liu, Frankel, & Roohr, 2014). Because these measures are often collected under low-stakes conditions, there is a concern students do not put forth their best effort when responding, which is problematic given previous research has found noneffortful responding can negatively impact the validity of results (e.g., Barry & Finney, 2009; Meade & Craig, 2012; Swerdzewski, Harmes, & Finney, 2011). Subsequently, there is a need to identify students displaying low effort on low-stakes noncognitive measures. One method, which is based on response time and can discreetly assess student effort at the item level, is the solution behavior (SB) index (Kong, Wise, & Bhola, 2007). A challenging task in using the SB index is the identification of an appropriate time threshold that can meaningfully distinguish responses made with effort (i.e., solution behavior responses) from responses made without effort (i.e., rapid responses). Thus, the purpose of the current study was to examine if the SB index could be used with low-stakes noncognitive measures to distinguish responses – and ultimately students – exhibiting solution behavior from responses made without any effort. In particular, eight different time threshold calculation methods were used to classify responses to a noncognitive measure assessing the construct meaningful life. The resulting time thresholds and SB classification indices were compared and external validity evidence for the resulting SB classification indices was gathered. Results of the study found support for four of the eight threshold calculation methods. In particular, support was found for defining the time thresholds by (a) visually inspecting items' response time distributions, (b) visually inspecting items' response time distributions

with a known group of rapid responders added to the sample, (c) using a normative threshold that was 30% of the average response time to an item, and (d) using lognormal mixture modeling. Practical implications and limitations of the results are discussed.

CHAPTER ONE

Introduction

Chapter Overview

Noncognitive measures are increasingly being used for accountability purposes in higher education (e.g., O. L. Liu, Frankel, & Roohr, 2014). Because these measures are often collected under low-stakes conditions, there is a concern students do not put forth their best effort when responding (Haladyna, & Downing, 2004; O. L., Liu et al., 2014). Although multiple methods have been developed to identify examinees displaying low effort on low-stakes cognitive tests (e.g., self-report effort scores; Sundre & Moore, 2002), less attention has been paid to methods that can detect low effort on low-stakes noncognitive measures. The purpose of the current study was to examine if a method based on response time known as the solution behavior (SB) index could be used to identify responses made without effort on a low-stakes noncognitive measure. Prior to describing the study, the current chapter reviews (a) why noncognitive measures are increasingly being used for accountability purposes, (b) problems associated with noneffortful responding, (c) methods used to detect noneffortful responding, (d) research using response time in the survey literature, (d) research using response time in the low-stakes cognitive literature, and (e) research using the SB index with noncognitive measures. Finally, the need for the current study is addressed.

Accountability in Higher Education

Over the last decade, the general public, policymakers, and other stakeholders have become increasingly concerned about the quality and affordability of higher education in the United States (U.S.). A series of reports and news briefs released in the

beginning of the 21st century spurred the beginning of what is now referred to as the "era of accountability." Describing a series of disturbing trends, the reports indicated the quality of education U.S. students were receiving was deteriorating although the cost of tuition and the amount of time students spent pursuing a degree were substantially increasing. For example, a report released in 2006 by the U.S. Department of Education indicated an increasing number of college-educated adults did not possess basic reading, writing, and mathematical skills prior to graduating college (U.S. Department of Education, 2006).

One year later, a report entitled *America's Perfect Storm* described how changes in the U.S. economy induced by technology and globalization, paired with the increasing demographic diversification of the workforce had prompted employers to note an increasing number of employees did not possess the knowledge, skills, or abilities necessary to enter the workforce – a concept referred to as "workforce ready" (Kirsch, Braun, Yamamoto, & Sum, 2007). According to *America's Perfect Storm* (Kirsch et al., 2007) and other similar reports (e.g., Giffi et al., 2015; Society for Human Resource Management, 2015), one of the major skills employees fail to possess were "soft skills" (i.e., noncognitive skills), such as communication, teamwork, and critical thinking. Soft skills are increasingly becoming necessary in the workforce and are considered by many employers as more important for success than "hard," cognitive skills (Kirsch et al., 2007; Kyllonen, 2013; Markle, Brenneman, Jackson, Burrus, & Robbins, 2013; Naemi et al., 2012; Robles, 2012). Similarly, soft skills such as critical thinking and perseverance are also considered important predictors for success in school (Kyllonen, 2000, 2013; Markle et al., 2013).

Given these revelations, policy makers and the general public began to question the utility of pursuing and obtaining a college degree and demanded higher education institutions be held accountable for student learning. In response, colleges and universities began developing and assessing student learning outcomes (if they had not been doing so already). According to one survey in 2013, of 1,202 accredited colleges and universities, 84% had adopted student learning outcomes whereas only 74% had in 2009 (Kuh, Jankowski, Ikenberry, & Kinzie, 2014). In addition, colleges and universities began adopting noncognitive student learning outcomes despite having historically only focused on developing students' cognitive knowledge and skills (Kyllonen, 2013; Pascarella & Terenzini, 2005). As a result, noncognitive assessments are now increasingly being used for accountability purposes in higher education institutions (Markle et al., 2013; Naemi et al., 2012; Schuh & Gansemer-Topf, 2010). For example, some noncognitive constructs that are commonly assessed include critical thinking skills, intercultural competence, and personal well-being (O. L. Liu et al., 2014; Schuh & Gansemer-Topf, 2010; Torney-Purta, Cabrera, Crotts Roohr, Liu, & Rios, 2015).

Testing Stakes, Motivation, and the Problem of Noneffortful Responding

Many assessments administered in higher education, including noncognitive measures, are typically administered under low-stakes conditions. Noncognitive measures are considered to be low stakes when the results are used to make inferences about student learning and development outcomes and are not used to make decisions about the students themselves. That is, although others (e.g., teachers, administrators) may be impacted by the results, students completing the noncognitive measure are not directly impacted by the results. As a result, students completing low-stakes measures may be less

likely to effortfully respond to items than they would if they were completing high-stakes tests. Previous research has found noneffortful responses on noncognitive measures are essentially meaningless – they "are missing data that is not actually missing" (Curran, 2015, p. 1). Specifically, noneffortful responses can distort item-level and composite-level scores, attenuate or inflate relationships with other variables, attenuate internal consistency estimates, and impact the factor structure of a measure (Barry & Finney, 2009; Conway, 2002; Huang, Liu, & Bowling, 2015; Kam & Meyer, 2015; MacKenzie & Podsakoff, 2012; Meade & Craig, 2012; Swerdzewski, Harmes, & Finney, 2011). For instance, if a student effortfully responds to a noncognitive measure, then the resulting composite score should reflect the level of the trait measured. However, if a student does not put forth effort when responding to the measure, then the resulting composite score will not adequately reflect the level of the trait measured (Swerdzewski et al., 2011). Consequently, any inferences based on the resulting composite scores would be incorrect. Moreover, because results are often aggregated across students and used to make inferences about the effectiveness of educational programs, the resulting conclusions would also be inaccurate. For example, Swerdzewski et al. (2011) examined the impact low motivation had on the validity of results and found low examinee effort significantly and practically inflated the composite scores of two subscales measuring worrisome thinking and amotivation. Because both of these traits are maladaptive, lower composite scores are desirable. Thus, had the results been used and examinee effort not been taken into account, administrators would have concluded that the programming was ineffective given students displayed higher levels of the maladaptive traits.

Unfortunately, a large majority of researchers mistakenly believe noneffortful responses on noncognitive measures only attenuate, or underestimate, relationships (Huang et al., 2015). Falsely believing noneffortful responses only attenuate relationships "may lead unsuspecting researchers to be complacent about the need to screen" (Huang et al., 2015, p. 838) the data for noneffortful responding and as a result, researchers may unknowingly make Type I errors (i.e., incorrectly reject the null hypothesis even though there is no true difference). For example, if the effectiveness of a student affairs program was evaluated by comparing students who attended the program to students who did not attend, researchers who do not screen for noneffortful responding may incorrectly conclude the two groups were different even though in reality they were not. This is particularly problematic because if a study with Type I errors is published any future replication attempts may yield results that either confirm or conflict with the original results. Thus, without clearly understanding noneffortful responding is the cause of discrepant results across studies, the "conflicting results may take volumes of time and effort to untangle" (Curran, 2015, p. 2).

In summary, the presence of noneffortful responding on low-stakes assessments is a major problem. Given the increasing use of noncognitive measures for accountability purposes, there is a concern decisions will be made in error if students' motivation is not taken into account. As argued by Wise (2015), "the issue of test-taking effort becomes a matter of professional ethics" (p. 250). That is, given results are used to make decisions such as program effectiveness, researchers have an obligation to identify and address noneffortful test-taking when present (Wise, 2015). However, an inherent challenge in

addressing noneffortful test-taking is first being able to identify students responding without effort on low-stakes noncognitive measures.

Identifying Noneffortful Responses

Identifying students responding without effort enables researchers to examine the extent to which low motivation is a problem, improve the quality of the results by removing unmotivated respondents, and study motivation and its relationship with other factors. Given the utility of identifying noneffortful responses, several methods have been developed to detect noneffortful responding (Curran, 2015; DeSimone, Harms, & DeSimone, 2015). For example, self-report measures of effort ask respondents to indicate how much effort was put forth on the substantive measure of interest (e.g., Swerdzewski et al., 2011; Wise & DeMars, 2010). Although seemingly useful, a primary disadvantage of using self-report measures of effort is they only measure the *overall* amount of effort respondents exhibited as opposed to the amount of effort respondents exhibited on each item (Sundre & Moore, 2002). To convey the problem with overall scale-level measures of effort, consider two respondents who have the same moderately high scale-level effort score. The first respondent provided moderately-high effort on all of the items, whereas the second respondent provided high effort on some items but not on others. Although both of these respondents received the same overall effort score, they differ by how much effort they displayed throughout the assessment. That is, the first respondent displayed a constant amount of moderately-high effort on every item, whereas the second respondent did not. As demonstrated by this example, although overall scale-level measures of effort are informative, they have limited utility in conveying whether levels of effort changed during an assessment and how they varied if so.

Another disadvantage of using self-report measures of effort is their susceptibility to respondent bias effects such as social desirability responding. That is, given self-report measures of effort are direct measures of effort and respondents know their effort is being monitored, it is possible respondents will falsify their answers to appear as though they put forth effort when in reality they did not. Moreover, it is not likely students who completed the substantive measure of interest without effort will suddenly display effort while completing the self-report measure of effort.

Given these disadvantages associated with self-report measures of effort, there is a need for a measure of respondent effort at the item level and for one that is covert. These two needs will be addressed in turn. First, identifying noneffortful responses at the item level provides practitioners with a wealth of information that cannot be obtained from scale-level measures. Specifically, item-level measures of effort allow practitioners to examine whether effort changes or remains stable across items. In addition, item-level measures of effort can also be used to calculate overall scale-level measures of effort and can be used to study item characteristics that are related to effort. Moreover, item-level measures of effort can serve as a red flag to scale developers for items needing modification. For example, if a large proportion of respondents answered an item without effort, a review of the item may indicate the wording of the item was ambiguous, thus signaling to practitioners the item needs to be modified. Second, in addition to needing a measure of effort at the item level, there is also a need for a covert or discreet measure of effort: if respondents are not aware their effort is being monitored, they are less likely to provide false or inaccurate answers for deception purposes.

One identification method that can measure noneffortful responding at the item level and covertly is based on the amount of time respondents use to answer an item, referred to as response time. An advantage of using response time to measure effort is it can be collected at any level (e.g., webpage, entire assessment), including the item level. In addition, response time can measure effort discreetly. That is, because respondents are not aware their response time is being collected and used to measure effort, it is not susceptible to respondent bias as other identification methods are; the only way respondents can fake motivation is to spend more time on an item. A disadvantage in using response time to identify noneffortful responses is it requires administrating the assessment via a computer. However, given the increasing availability and use of computers, this is not as much of a problem as it once was.

Although response time is a promising covert measure of item-level effort, it has rarely been used with noncognitive measures administered for accountability purposes. However, because response time has been used to measure effort in the survey research and low-stakes cognitive testing domains, the literature relevant to these domains is reviewed below.

Uses of Response Time in the Survey Literature

Response time has been used in a variety of ways in the survey literature to identify noneffortful responses. Although some survey researchers have used response time as a continuous variable (e.g., Maniaci & Rogge, 2014; Meade & Craig, 2012), the majority of survey researchers have used response time as a dichotomous variable to distinguish responders who rapidly respond to items from those who did not (e.g., Huang, Curran, Keeney, Poposki, & DeShon, 2012). To create this dichotomy, researchers

identify a minimum amount of time required to complete an item or webpage and then classify respondents' motivation based on this time threshold: respondents who complete the item or webpage faster than the defined time threshold are flagged as responding without effort (Huang et al., 2012; M. Liu, Bowling, Huang, & Kent, 2013; Meade & Craig, 2012; Zhang & Conrad, 2014). Comparing a respondent's response time on an item to a predetermined time threshold distinguishes those who may have taken the time to thoughtfully respond to the item from those who did not put forth any effort in responding to the item. Although this method does not capture those respondents who took a longer amount of time to respond without effort, it is effective at capturing those respondents who assuredly did not put forth any effort in responding at all.

When treated as a dichotomy to differentiate responses made with effort from those made without, researchers have used a variety of methods to define time thresholds. For instance, some researchers have defined time thresholds a priori based on the expected time it took to read an item (e.g., 300 milliseconds per word; Zhang & Conrad, 2014), whereas others used an "educated guess" and applied the same time threshold to every webpage (e.g., Huang et al., 2012, p. 106). In contrast, other researchers have defined time thresholds after collecting data by using the average response time of the sample under study (e.g., Meade & Craig, 2012).

Although the survey research literature has utilized response time in a variety of ways to measure respondent effort, to my knowledge, no studies have focused on how best to use response time for this purpose. That is, various methods have been used to define time thresholds in the survey research literature, but no studies exist with the explicit purpose of evaluating the utility of the various time threshold calculation

methods. In addition, because studies in the survey research literature typically use the response time of a webpage or entire survey, the benefits associated with the use of response time at the item level have not been fully explored. Because research in the low-stakes cognitive literature has paid relatively more attention to the use of item-level measures of effort and the calculation of time thresholds, the following section reviews how response time has been used when measuring examinee effort in low-stakes cognitive testing environments.

Uses of Response Time in the Low-Stakes Cognitive Literature

In addition to being used in the survey research literature, response time has also been used with low-stakes cognitive tests administered for accountability purposes to identify examinees responding to items without effort (e.g., Wise & DeMars, 2010). The most common way response time has been used in this context is in the creation of the solution behavior (SB) index. Specifically, the SB index classifies each item-examinee response based on an examinee's response time in comparison to a predetermined time threshold (Wise & Kong, 2005). If the response time exceeds the time threshold, the response is classified as a solution behavior response. In contrast, if the examinee answers the item faster than the amount of time necessary to read and thoughtfully respond to the item (i.e., respond without effort), and the response is faster than the time threshold, the response is classified as a rapid response (e.g., Wise & DeMars, 2010; Wise, Pastor, & Kong, 2009).

Interestingly, although survey researchers often dichotomize response times in an attempt to measure respondent effort, the resulting classification variables are never referred to as SB indices. In fact, with the exception of Huang et al. (2012), references to

studies using low-stakes cognitive tests that created and use the SB index are absent in the survey literature. Given the lack of overlap between the survey literature and cognitive testing literature, it is not surprising different methods for setting time thresholds emerged in the cognitive domain. Some of the more commonly used methods to define time thresholds in the low-stakes cognitive literature include visually inspecting an item's response time distribution (Wise, 2006), fitting lognormal mixture models (Kong, Wise, & Bhola, 2007), and calculating time thresholds as a percentage of the average response time for a sample (Wise & Ma, 2012). For example, the visual inspection method is based on the assumption an item's response time distribution will appear bimodal if motivated and unmotivated respondents are present. Specifically, a smaller mode occurring almost immediately at the low end of the distribution represents those responding without effort and a larger mode occurring above the median response time represents those responding with effort. In general, previous research empirically comparing the various time threshold calculation methods has found the time threshold calculation methods do not substantially differ from one another when applied to low-stakes cognitive tests (Kong et al., 2007; Pastor, Strickman, & Ong, 2015; Wise & Ma, 2012).

As previously reviewed, response times in the survey literature are typically collected at the webpage or survey level and are rarely collected at the item level. In contrast, response times used in the low-stakes cognitive literature are typically collected at the item level explicitly for the purpose of creating the SB index. The stronger emphasis on item-level measurement of effort in low-stakes cognitive testing has led to a variety of different applications using the SB index, which are briefly outlined here and

are more fully described in Chapter Two. Specifically, within the low-stakes cognitive literature, the SB index has been used to study examinee behavior at the item level (e.g., Wise, 2006), study item and examinee characteristics related to low effort (e.g., Wise et al., 2009), and has been incorporated into measurement models to account for respondents exhibiting rapid-guessing behavior (e.g., DeMars, 2007). In addition, the SB index has also been used to determine if there are groups of examinees who demonstrate similar response patterns across a test that were distinctly different from the patterns of other groups of examinees (e.g., Pastor et al., 2015), and has been used to create other measures to study examinee behavior (Wise & DeMars, 2005; Wise, 2006), such as a test-level measure of effort known as Response Time Effort (Wise & Kong, 2005). Finally, the SB index has also been used to evaluate the use of other measures of noneffort such as self-report measures of effort (Swerdzewski et al., 2011).

Use of the SB Index with Noncognitive Measures

Despite the extensive utility of the SB index, since it was first applied to low-stakes tests in 2005, the SB index has primarily been used to identify unmotivated examinees completing low-stakes cognitive tests. To my knowledge, only one study has used the SB index to identify noneffortful responses made on low-stakes noncognitive measures. Specifically, Swerdzewski et al. (2011) used the SB index to identify respondents who were rapidly responding to items on four different noncognitive measures assessing students' attitudes towards learning, academic motivation and beliefs, level of worry, and appreciation for diverse experiences. The SB index was also used in the study to measure effort put forth on low-stakes cognitive tests administered in the same testing session.

Given this is the only study known of to apply the SB index to low-stakes noncognitive measures administered for accountability purposes, it is important to note how the solution behavior time thresholds were calculated. Swerdzewski et al. (2011) defined the time thresholds for each item by visually inspecting items' response time distributions. Recall, the visual inspection method is based on the assumption an item's response time distribution will appear bimodal if motivated and unmotivated respondents are present. Swerdzewski and his colleagues (2011) defined the time threshold as the point where the two distributions crossed; the time thresholds were then cross-validated by comparing the defined time thresholds to the minimum amount of time required to read the items. After classifying examinees according to their response behavior, the researchers then used the item-level SB index values to calculate measures of effort at both the test-level and testing session-level; the results were then compared to corresponding self-reported measures of the examinees' effort.

It is important to note the focus of the Swerdzewski et al. (2011) study was not on the use of the SB index with noncognitive measures; instead, the authors were interested in the correspondence between test-level and testing-session level measures of effort as measured using response time and self-report scales and any differences in the resulting test scores when the various measures of effort were used to filter or remove unmotivated examinees from the data. Thus, although this is the first study known of to use the SB index with noncognitive measures to identify noneffortful responses, the study did not thoroughly evaluate the use of the SB index with noncognitive measures.

Considerations in the Use of the SB Index with Noncognitive Measures

Despite the advantages of using the SB index to identify noneffortful responses, questions remain regarding whether the SB index can be effectively used to measure noneffortful responding on noncognitive measures. Although Swerdzewski and his colleagues (2011) were able to calculate the time thresholds using the visual inspection method and acquired some validity evidence for measures of effort based on the SB index, their focus was not on the use of the SB index with noncognitive measures per se. Before the SB index is adopted for use with other noncognitive measures, more research is needed to provide guidance on the utility and validity of using different methods to calculate the time thresholds (e.g., visual inspection, lognormal mixture modeling) which are used to calculate the SB index.

Prior research in the low-stakes cognitive testing literature has found minor differences among various time threshold calculation methods (e.g., Kong et al., 2007). However, it is inappropriate to assume these findings will generalize to time thresholds when applied to noncognitive measures. In particular, differences between cognitive tests and noncognitive measures may make some of the time threshold calculation methods more difficult or impossible to use when applied to noncognitive measures. Specifically, cognitive tests assess knowledge using dichotomously scored items whereas noncognitive measures assess attitudinal traits using items answered by a rating scale. Moreover, item stems and response options on cognitive tests are typically longer in length and more complex than item stems and response options on noncognitive measures. In particular, response options on rating scales used by noncognitive measures typically do not vary item to item.

Based on these differences, it is likely response times for items on a cognitive test are longer and have more variability between examinees than response times for items on a noncognitive measure. Moreover, given response times for items on a noncognitive measure will probably be shorter than response times for items on cognitive tests, it is likely the response time distributions for noncognitive items will not exhibit a clear bimodal pattern, which would suggest the solution behavior time thresholds for those items using the visual inspection calculation method could not be calculated. Therefore, a potential challenge to using the SB index with noncognitive measures is defining an appropriate time threshold that distinguishes noneffortful responses from those made with effort. If the response time distributions for items on noncognitive measures are much shorter and less variable than for items on cognitive tests, then time threshold calculation methods that rely on the bimodal distribution assumption such as the visual inspection method will not work. Although several threshold calculation methods have been developed and studied in the low-stakes cognitive literature (and are reviewed in detail in Chapter Two), the threshold calculation methods have not been applied or studied in either the survey literature or the noncognitive assessment literature. Based on these considerations and given Swerdzewski et al. (2011) is the first study known of to apply the SB index to low-stakes noncognitive measures administered for accountability purposes, more research is needed to determine if the SB index can be applied to low-stakes noncognitive measures.

Need for study

Given the increasing use of noncognitive measures in educational settings, the negative impact responding without effort has on the validity of results, and the

advantages of using the SB index over other identification methods, more research using the SB index to identify respondents rapidly responding to items on low-stakes noncognitive measures is needed. The general purpose of the current study was to examine if the SB index could be used with noncognitive measures. Specifically, the purpose of the current study was three-fold.

First, it was of interest to contribute to the literature and determine if the SB index could be calculated using various time threshold calculation methods and identify students rapidly responding to items on a low-stakes noncognitive measure without effort. In addition, it was also of interest to examine whether including responses from a known group of rapid responders would affect the calculation of the time thresholds and subsequent solution behavior classification indices.

Second, if the solution behavior time thresholds could be defined using the various time threshold calculation methods, then the second purpose of the study was to compare the time thresholds and resulting SB index values across the threshold calculation methods at the item level.

Finally, the third purpose of the study was to gather validity evidence for the time thresholds and resulting SB classification indices to determine if the time thresholds were meaningful and if there was support for using one threshold calculation method over another. In addition, it was of interest to determine if the external validity evidence, when considered in conjunction with the results from the second purpose of the study, supported the use of one threshold calculation method over another.

Given the dearth of research applying the SB index to low-stakes noncognitive measures administered for accountability purposes, it was important to thoroughly review

the research that has been conducted using low-stakes cognitive tests. Thus, prior to further describing the current study, research using the SB index with low-stakes cognitive tests will be reviewed in detail in Chapter Two. Specifically, Chapter Two will (a) thoroughly review the empirical research using the SB index with low-stakes cognitive tests, (b) identify and examine various time threshold calculation methods used to calculate the SB index, (c) review the empirical research using the SB index with noncognitive measures, and (d) present the purpose of the study in detail.

CHAPTER TWO

Literature Review

Chapter Overview

The purpose of the current chapter was to review the empirical research applying the SB index to low-stakes cognitive tests. Given the dearth of research examining the application of the SB index to low-stakes noncognitive measures, reviewing the research conducted in the low-stakes cognitive literature will serve two aims. First, the literature review will show how the SB indices have previously been used when applied to low-stakes cognitive tests and how the SB indices could potentially be used with low-stakes noncognitive measures if the methods for defining time thresholds are successful. Second, the review of the literature will also provide a review of the relationships with external variables used to gather external validity of the indices when applied to cognitive tests and which potentially could be used to provide validity evidence of the SB indices when applied to low-stakes noncognitive measures. Addressing these aims, Chapter Two (a) reviews the empirical research using the SB index with low-stakes cognitive tests, (b) identifies and examines various time threshold calculation methods commonly used to calculate the SB index when applied to low-stakes cognitive tests, (c) reviews the empirical research using the SB index with noncognitive measures, and (d) presents the purpose of the study in detail.

Solution Behavior

A substantial body of research has emerged over the last decade exploring the detrimental impact low examinee motivation has on low-stakes cognitive test performance (DeMars, 2007; Eklöf, 2010; Finn, 2015; Thelk, Sundre, Horst, & Finney,

2009; Wise & DeMars, 2005; Wise & DeMars, 2006; Wise & DeMars, 2010; Wise & Kong, 2005; Wise, 2006, 2015). Specifically, low examinee motivation introduces construct-irrelevant variance into test scores which in turn significantly impacts the validity of test score interpretations and their subsequent uses, especially when the data are collected under low-stakes testing contexts (Eklöf, 2010; Haladyna & Downing, 2004; Wise, 2015). Recognizing this impact, various methods have been developed to identify examinees who demonstrate low effort on low-stakes cognitive tests. For example, in 2005, Wise and Kong applied a method originally developed to identify examinees who started to rapidly answer items as they began to run out of time on a high-stakes speeded test (Schnipke & Scrams, 1997).

Schnipke and Scrams (1997) demonstrated examinees completing high-stakes speeded cognitive tests would exhibit either solution or rapid-guessing behavior when responding to items depending on how much time remained during the testing session. Specifically, *solution behavior* (SB) refers to the behavior an examinee exhibits when trying to correctly answer an item whereas *rapid-guessing behavior* refers to an examinee's response to an item that occurs so rapidly there was not enough time for the examinee "to fully consider the item" (Wise, 2006, p. 97). When completing high-stakes speeded tests, Schnipke and Scrams (1997) found examinees exhibited solution behavior on the majority of test items until they began to run out of time, at which point examinees would strategically switch response strategies and exhibit rapid-guessing behavior.

Although Schnipke and Scrams (1997) developed the SB index for use with high-stakes speeded tests, Wise and Kong (2005) recognized its utility in identifying examinees responding to items without effort on low-stakes unspeeded cognitive tests

administered for institutional accountability purposes. They hypothesized and found unmotivated examinees completing low-stakes tests would exhibit rapid-guessing behavior throughout the testing session and not just towards the end of a test as time ran out. Specifically, the solution behavior index, SB_{ij}, is a dichotomous index that assesses the amount of effort examinee *j* puts forth answering item *i* on a low-stakes unspeeded test (Wise & Kong, 2005). Based on the amount of time (in seconds) it takes examinees to answer an item, the SB index is calculated by comparing the response time, RT_{ij}, of examinee *j* on item *i* to an identified time threshold, T_i, for item *i*

$$SB_{ij} = \begin{Bmatrix} 1\ if\ RT_{ij} \geq T_{i,} \\ 0\ otherwise. \end{Bmatrix}. \quad (1)$$

Conceptually, the SB index classifies examinees into one of two categories: examinees who are assumed to meaningfully respond by taking time to try to correctly answer an item versus examinees who are assumed to meaninglessly respond by rapidly selecting an answer to an item faster than the amount of time required to read and correctly answer an item (Swerdzewski et al., 2011). Thus, examinees who are assumed to meaningfully respond to an item are classified as exhibiting solution behavior, whereas examinees who are assumed to meaninglessly respond to an item without effort are classified as exhibiting rapid-guessing behavior. For example, consider an item with a defined threshold, T_i, of 15 seconds. If an examinee responds to this item in 20 seconds (i.e., above the item's time threshold), then the examinee is classified as exhibiting solution behavior (SB_{ij} = 1). In contrast, if an examinee responds to the item in 3 seconds, then the examinee is classified as exhibiting rapid-guessing behavior (SB_{ij} = 0). Classifying an examinee's response to an item as a solution behavior response does not necessarily indicate the examinee actually put forth effort trying to correctly answer that

item; it only indicates the examinee did not rapidly respond to that item (Wise & Smith, 2011).

A challenge to using the SB index is to select an appropriate time threshold for items so examinees' responses are appropriately classified as either solution behavior or not (Wise & Kingsbury, 2015). Based on this challenge, various calculation methods have been developed to define time thresholds for items on cognitive tests. Some methods commonly used to define time thresholds for items on cognitive tests are based on researchers' judgments whereas other methods are based on statistical techniques (Kong et al., 2007; Wise, 2006). In addition, some methods will always calculate a time threshold whereas other methods may not always be able to calculate a time threshold (Wise & Ma, 2012). Although various methods have been developed to calculate the time thresholds, the majority of research using the SB index has been conducted with the purpose of studying examinee behavior, rather than empirically evaluating the effectiveness of the index. Given the purpose of the current study is to apply the SB index to a noncognitive measure and given the difference between items on cognitive tests and noncognitive measures, it is important to review previous research applying the SB index and evaluating its effectiveness when applied to low-stakes cognitive tests. Thus, prior to discussing the various time threshold calculation methods previously used with low-stakes cognitive tests, the following section will first review how the SB index has been used with low-stakes cognitive tests and what information has been gathered to support the validity of the SB index when applied to low-stakes cognitive tests.

Applications of the Solution Behavior Index

Since the SB index was first applied to low-stakes unspeeded tests in 2005, it has primarily been used with cognitive tests to create other measures of examinee effort or used on its own. The primary focus of these studies has been on examining and studying examinee behavior rather than focusing on the index itself. Given the SB index has been used with low-stakes cognitive tests in various ways and in order to explore how it can be used with noncognitive measures, the following sections review (a) how the SB index has been used to create test-level measures of effort and (b) how the SB index has been used on its own.

Using the solution behavior index to create RTE. Interestingly, researchers studying examinee test-taking effort in low-stakes contexts have primarily used the SB index to create other measures. For example, although the SB index was first applied to low-stakes cognitive tests in Wise and Kong's (2005) seminal paper, the primary focus of the paper was another measure of effort created from the set of SB index values known as Response Time Effort (RTE). RTE is a test-level measure of examinee effort created from a set of SB index values (Wise & Kong, 2005). Specifically, RTE reflects the proportion of test items on which an examinee exhibited solution behavior and is calculated as

$$RTE_j = \frac{\sum SB_{ij}}{k}, \tag{2}$$

where the term in the numerator is the sum of SB index values for examinee *j* across all items and *k* is the total number of items on the test. RTE_j values range from 0 to 1; higher values indicate examinees exhibited solution behavior on more items on the test whereas lower values indicate examinees exhibited rapid-guessing behavior to more items on the test. For example, an examinee with a RTE_j value of .95 indicates the examinee engaged

in solution behavior on 95% of the test items whereas a RTE_j value of .65 indicates the examinee engaged in solution behavior on only 65% of the test items.

Since its development in 2005, RTE has been used for several purposes including (a) describing examinees' test-taking motivation, (b) examining how examinee characteristics are related to test-taking motivation, (c) exploring how examinee effort changes during a testing session, and (d) exploring the impact low examinee effort has on low-stakes test scores. These four applications are described in further detail below.

Describing RTE. By itself, RTE is a useful test-level measure that gauges how often examinees exhibit rapid-guessing behavior on low-stakes tests. This information is useful for test users who may suspect examinees put forth low effort on a test but "have little empirical evidence concerning the degree to which low effort was actually present" (Wise & Kong, 2005, p. 180). For example, Wise and his colleagues (2009) found evidence indicating 386 upperclass college students completing a 64-item low-stakes test assessing their quantitative and scientific reasoning skills exhibited solution behavior on 90% of the items, on average, which in turn indicates students exhibited rapid-guessing behavior on 10% of the test items, on average. When examined further, 53% of the examinees exhibited solution behavior on every item (RTE = 1.00). In contrast, 23% examinees exhibited solution behavior on at least 90% of the items ($.90 \leq RTE < 1.00$) and 24% examinees exhibited solution behavior on less than 90% of the items ($RTE < .90$; Wise et al., 2009).

Similarly, Wise and Kong (2005) found that out of 472 freshmen college students completing a low-stakes test assessing their information literacy skills, 63.3% of the examinees exhibited solution behavior on all of the test items (RTE = 1.00). In contrast,

29.2% of the examinees exhibited solution behavior towards at least 90% of the test items ($.90 \leq RTE < 1.00$) whereas 7.4% of the examinees exhibited solution behavior on less than 90% of the test items ($RTE < .90$). The variability demonstrated by RTE indicates (a) examinees do exhibit rapid-guessing behavior on low-stakes unspeeded cognitive tests and (b) examinees vary from one another by the degree to which they exhibit rapid-guessing behavior. It should also be noted these results indicate the majority of examinees do exhibit solution behavior on the majority of test items, a trend that has been displayed in several other studies as well (e.g., DeMars, 2007; Kong et al., 2007; Pastor et al., 2015; Setzer, Wise, van den Heuvel, & Ling, 2013).

RTE and examinee characteristics. RTE has also been used to explore the relationship between test-taking motivation and examinee characteristics. For example, in regards to gender, female college students tend to exhibit more effort on low-stakes tests than males (e.g., DeMars, Bashkov, & Socha, 2013; Setzer et al., 2013; Wise & DeMars, 2010). Specifically, Wise and DeMars (2010) found female upperclass college students completing a low-stakes oral communication test exhibited solution behavior on a higher proportion of items, on average, than their male counterparts ($RTE_{females} = .966$, $RTE_{males} = .896$, respectively). Similarly, Setzer and his colleagues (2013) found female college students completing a low-stakes major field test in business exhibited solution behavior on a significantly higher proportion of items, on average, than male college students ($RTE_{females} = 0.991$, $RTE_{males} = 0.985$, $d = .10$). DeMars and her colleagues (2013) also found men exhibited less effort than women, on average, across four tests administered to upperclass college students majoring in business. Conversely, however, Wise et al. (2009) found RTE and gender of upperclass college students completing a low-stakes

quantitative and scientific reasoning skills test were not related (r = -0.02; Wise et al., 2009).

Another examinee characteristic frequently studied in relation to RTE is academic ability. In general, independent measures of examinees' academic ability (i.e., independent of the low-stakes test of interest) have displayed nil to low correlations with RTE (e.g., Kong et al., 2007; Rios, Liu, & Bridgeman, 2014; Wise & DeMars, 2010; Wise & Kong, 2005; Wise et al., 2009). For example, RTE on a low-stakes test assessing college students' knowledge of their major field in business was not related to their overall or major GPA (r = .01, r = .02, respectively; Setzer et al., 2013). Similarly, RTE was not significantly related to SAT-Verbal or SAT-Quantitative scores for incoming freshmen college students completing a low-stakes information literacy test (r = .06, r = -.02, respectively; Wise & Kong, 2005), for upperclass college students completing a low-stakes oral communications test (r = .09, r = .02, respectively; Wise & DeMars, 2010), or for senior college students completing a low-stakes proficiency profile test on critical thinking, reading, writing, and mathematics skills (r = .09, r = -.03, respectively; Rios et al., 2014). In addition, four RTE scores (that were calculated using four different methods to define the SB time thresholds) were not related to SAT-Verbal (r = .07 to r = .08) or SAT-Quantitative scores (r = -.04 to r = -.06) for upperclass college students completing a low-stakes information literacy test (Kong et al., 2007). In contrast to the nil correlations just reviewed, RTE scores for upperclass college students completing a low-stakes test assessing their quantitative and scientific reasoning skills did exhibit a correlation small in magnitude with a combined SAT-Verbal and SAT-Quantitative score (r = .19; Wise et al., 2009).

Two other examinee characteristics that have been examined (albeit once) are class status (e.g., freshmen, sophomore) and race. Specifically, when divided by class status, Wise and DeMars (2010) found college students entering their freshmen year exhibited solution behavior towards a larger proportion of test items than did upperclass students in their sophomore or junior year of college ($RTE_{Freshmen}$ = .996, $RTE_{Upperclass}$ = .943, respectively). In another study that considered race, Setzer and colleagues (2013) found examinees who classified themselves as White had significantly higher RTE scores, on average, than did Non-White examinees (RTE_{White} = 0.991, $RTE_{Non\text{-}White}$ = 0.980, d = .19).

RTE and changes in effort. RTE has also been used to explore how effort fluctuates across multiple testing sessions. For example, DeMars (2007) examined if examinee effort changed across a series of low-stakes tests administered over the course of four weeks. DeMars (2007) found examinees who completed two tests every week across four weeks exhibited rapid-guessing behavior more often on tests administered towards the end of the testing period, on average, than on tests administered at the beginning of the testing period, on average. Specifically, when tests were administered during the first week, examinees displayed solution behavior on 97% to 100% of the test items, on average. In contrast, when tests were administered on the last week, examinees exhibited solution behavior on 85% to 95% of test items, on average (DeMars, 2007). In addition, DeMars (2007) also found rapid-guessing behavior occurred more frequently "on the same test when the test was given later in the series" (p. 40).

Swerdzewski et al. (2011) also used RTE to examine how motivation changed across tests. Instead of examining changes in effort across multiple testing sessions like

DeMars (2007) did, the researchers examined how effort changed during a single testing session when examinees were administered a series of low-stakes cognitive tests and noncognitive measures. Moreover, whereas DeMars (2007) examined how average RTE scores changed across time, Swerdzewski et al. (2011) examined how the percentage of students who were classified as exhibiting effort on a test (RTE ≥ .90) versus the percentage of people who were not classified as exhibiting effort on a test (RTE < .90) changed across tests. Given these considerations, Swerdzewski et al. (2011) found examinee effort did not systematically decrease across tests when a battery of tests were administered in a single testing session. In other words, the proportion of students who were classified as not exhibiting effort on the test (RTE < .90) did not substantially increase as the testing session progressed.

RTE and its impact on test scores. The fourth way RTE has been applied has been to examine the impact rapid-guessing behavior has on test scores. Although previous research has shown RTE is not related to independent measures of academic ability (e.g., SAT scores; Wise & Kong, 2005), research has found RTE is related to examinees' performance on the test for which RTE is measured. For example, Wise and DeMars (2010) found RTE was positively related to upperclass college students' performance on a low-stakes oral communication test (r = .73). Similarly, Wise and Kong (2005) also found RTE exhibited a moderate positive relationship with upperclass college students' performance on a low-stakes test assessing their information literacy skills (r = .54). Wise and Kong (2005) further examined the relationship between effort and test scores by dividing examinees into the following three groups based on their RTE scores: examinees with RTE scores less than .80, examinees with RTE scores between .80 and

.90, and examinees with RTE scores greater than .90. Recall, RTE reflects the proportion of items an examinee exhibited solution behavior, so higher scores are desirable because lower scores indicate examinees exhibited rapid-guessing behavior more frequently. Thus, the group of examinees with RTE scores less than .80 exhibited the highest rate of rapid-guessing across items whereas the group of examinees with RTE scores greater than .90 exhibited the highest rate of solution behavior across items. Wise and Kong (2005) compared the three groups' performance on the low-stakes information literacy test and found examinees who exhibited solution behavior on at least 90% of the test items (i.e., RTE > .90) performed significantly better than examinees in the other two groups ($F(2, 469)$, $p < .001$, $\omega^2 = .26$).

Given the relationship between RTE and examinees' test performance, researchers have used RTE to identify and filter out examinees displaying low effort on low-stakes cognitive tests in order to examine the impact rapid-guessing behavior has on scores. Previous research has demonstrated rapid-guessing behavior on low-stakes cognitive tests attenuates examinees' performance on low-stakes tests (Wise, Bhola, & Yang, 2006; Wise & DeMars, 2006, 2010; Wise & Kingsbury, 2015; Wise & Kong, 2005), inflates estimates of internal consistency (e.g., Wise, 2015; Wise & DeMars, 2005, 2009, 2010; Wise & Kong, 2005), and attenuates relationships with theoretically related variables (e.g., DeMars, 2007; Wise & DeMars, 2006, 2010; Wise & Kong, 2005). For example, Wise and Kong (2005) examined what effect filtering out examinees with varying levels of RTE had on test scores and psychometric properties of a low-stakes information literacy test. The researchers found as examinees displaying increasing amounts of rapid-guessing behavior were removed from the data, average scores on the

information literacy test increased, variability in the test scores decreased, coefficient alpha decreased, the correlation between the test score and SAT-Verbal (an independent measure of examinee proficiency) increased, and examinees' average SAT-Verbal score did not significantly change. These results indicate removing examinees with the lowest RTE scores does not remove only examinees with the lowest ability level.

Similarly, Wise and DeMars (2010) investigated the impact rapid-guessing behavior had on test scores. The researchers found rapid-guessing behavior on a low-stakes oral communication test attenuated test scores, inflated coefficient alpha, and attenuated correlations between the test scores and convergent validity evidence such as SAT scores. For example, coefficient alpha for the low-stakes oral communication test was .84. However, when restricting the data to only those examinees who exhibited solution behavior on 90% or more of the test items (RTE $\geq$ .90), coefficient alpha decreased to .66. The fact rapid-guessing behavior on low-stakes tests spuriously inflates coefficient alpha is concerning because "practitioners may unwittingly perceive a false sense of security regarding the reliability of their test scores" (Wise & DeMars, 2010, p. 36; Wise & DeMars, 2009).

In addition to resulting in attenuated test scores and inflated coefficient alpha estimates, recent research using RTE has shown rapid-guessing behavior on low-stakes tests also impacts growth scores (i.e., gain scores; Wise, 2015; Wise & Ma, 2012). Specifically, rapid-guessing behavior has been found to increase the presence of non-credible growth scores (Wise, 2015; Wise & Ma, 2012). Non-credible growth scores are unrealistic growth scores that are either extremely negative or positive in value: extreme negative growth scores indicate examinees' knowledge substantially decreased over time

beyond what would be expected by chance or measurement error, whereas extreme growth scores indicate examinees' knowledge unrealistically increased over time (Wise, 2015; Wise & Ma, 2012). In a recent study, Wise (2015) identified and removed examinees who exhibited solution behavior on less than or equal to 90% of items (i.e., RTE $\leq$ .90) on a low-stakes test assessing ninth-graders academic progress in reading between the fall and spring of an academic school year calendar. By filtering examinees who displayed rapid-guessing behavior, 76% of the extreme scores flagged as exhibiting negative growth (-20 points or more) were removed because they displayed excessive rapid-guessing in the spring testing session. In addition, 62% of the scores flagged as exhibiting extreme positive growth (20 points or higher) were removed because they displayed excessive rapid-guessing behavior in the fall testing session. These results indicate extreme growth scores "can be often attributable to instances of students violating the universal assumption of effort" (Wise, 2015, p. 249).

Wise and DeMars (2010) also found rapid-guessing behavior differentially impacted subgroups' gain scores. That is, if examinees who displayed rapid-guessing behavior were not accounted for, then the results indicated only upperclass female examinees' performance on a low-stakes test significantly and practically improved from when they were originally tested as incoming freshmen (d = .62); upperclass male examinees' performance did not significantly change over time. In contrast, when examinees who exhibited solution behavior on less than 90% of the items (RTE < .90) were removed from the data, the results indicated both female and male upperclass examinees' scores significantly and practically increased over time (d_{females} = .81, d_{males} = .60, respectively; Wise & DeMars, 2010). As reviewed, RTE has been extensively used

to examine how low examinee effort can "seriously distort test score-based inferences" (Wise, 2015, p. 245).

Using the solution behavior index on its own. In addition to creating test-level measures of examinee effort such as RTE, the SB index has also been used on its own to study examinee motivation. One of the primary advantages of using the SB index to identify examinees not putting forth effort on low-stakes cognitive tests is it provides information about examinee effort at the item level. Other methods created from the SB index such as RTE only provide information about examinee effort at the test level, which requires test users and measurement practitioners to make an implicit assumption examinees exhibit constant effort throughout a testing session (Setzer et al., 2013; Wise, 2015; Wise & Kingsbury, 2015). This assumption is unrealistic to make, especially when mentally demanding cognitive tests are administered in low-stakes settings. When considered individually, the SB index provides a wealth of information for test users, test developers, and measurement practitioners. Specifically, the SB index has been used for multiple purposes including (a) describing examinee behavior at the item level, (b) studying item and examinee characteristics related to solution behavior, (c) examining how examinee behavior fluctuates during a test, (d) monitoring examinee behavior during a test, and (e) modifying measurement models in order to mitigate the effect low examinee motivation has on test score validity. These uses are reviewed in detail below.

Describing solution behavior. In order to describe how much effort items on a test receive from examinees, researchers have often calculated summary statistics of the SB index. For example, Wise and DeMars (2006) found that out of 31,440 total item responses (524 examinees x 60 items), only 5.8% of item responses were classified as

rapid guesses. Similarly, Setzer et al. (2013) found only 1.3% of 1,200,480 total item responses (10,004 examinees x 120 items) were classified as rapid guesses. Another way researchers have described the amount of effort individual test items received has been to calculate an item-level measure of effort known as response time fidelity (RTF). RTF is an item-level characteristic that reflects the proportion of examinees who exhibited solution behavior towards an item (Wise, 2006). RTF is calculated by summing the SB index values for item *i* across all examinees and dividing by the total number of examinees, *N*,

$$RTF_i = \frac{\sum_{j=1}^{N} SB_{ij}}{N}. \quad (3)$$

RTF values range from 0 to 1. Higher RTF values indicate a greater proportion of examinees exhibited solution behavior on an item, whereas smaller values indicate more examinees exhibited rapid-guessing behavior on an item. For example, an RTF value of .90 for item *i* indicates that 90% of examinees exhibited solution behavior while responding to item *i*.

Researchers have found the degree to which solution behavior responses occur varies across items on a test; that is, some items are answered with more effort than others. For example, Wise (2006) found RTF scores for entering freshmen college students completing a low-stakes information literacy test ranged from .907 to .988, whereas RTF scores for upperclass college students completing a shortened version of the same test ranged from .898 to .996. Similarly, Setzer et al. (2013) found RTF scores on a low-stakes major field test in business ranged from .898 to .998. Wise et al. (2009) also found the proportion of upperclass college students answering items on a quantitative and

scientific reasoning skills test with effort varied greatly, with RTF scores ranging from .78 to 1.00.

***Item and examinee characteristics*.** In an attempt to better understand examinee motivation on low-stakes tests and why it varies across items, researchers have studied various item and examinee characteristics related to solution behavior. For example, after finding rapid-guessing behavior varied across items in the first part of a two-part study previously mentioned, Wise (2006) examined if item characteristics could explain the observed variation in RTF. He found controlling for other characteristics, the proportion of examinees engaging in solution behavior on an item was significantly predicted by the square of an item's length (in characters), an item's position on the test (i.e., an item's position relative to other items on the test), and the presence of additional ancillary reading material (i.e., graphs or figures shown on previous items; $\beta = .47$, $\beta = -.54$, $\beta = -.20$, respectively). Controlling for other predictors, the non-linear relationship with item length indicated as the length of an item increased, the proportion of examinees answering the item with solution behavior decreased. Similarly, controlling for other predictors, as an item's position increased, the proportion of examinees answering an item with solution behavior also decreased.

In the second part of his study, Wise (2006) found upperclass college students exhibited solution behavior on 94% of the total item responses, on average, on a reduced 60-item version of the same low-stakes information literacy test (RTF = .94). Thus, in addition to finding upperclass examinees exhibited rapid-guessing behavior more frequently than entering freshmen, he also found the square of item length and an item's position significantly predicted RTF ($\beta = 1.00$, $\beta = -.44$, respectively). However, in

contrast to findings from the first part of the study, additional ancillary reading material did not significantly predict RTF as it had in the freshmen sample (Wise, 2006).

In another study, Setzer and colleagues (2013) found item length, item position, and the presence of ancillary reading material (such as a figure or graph) significantly predicted examinees exhibiting solution behavior on an item (β = -.44, β = -.41, β = -.24, respectively). Similarly, Wise et al. (2009) found item length and item position were negatively and moderately related to RTF (r = -.58, r = -.64, respectively). However, in contrast to previous findings demonstrating a negative relationship between ancillary reading material and RTF, Wise et al. (2009) found the presence of an item graphic (i.e., ancillary reading material) was positively related to RTF (r = .18).

In the studies just reviewed, item characteristics were related to aggregate scores of examinee effort on an item. That is, RTF scores – which reflect the proportion of examinees engaging in solution behavior on an item – were used as criterion variables in the first two of the three studies reviewed (Setzer et al., 2013; Wise, 2006). To date, only two studies have used the SB index as a dependent variable when examining the predictive relationship between effort and item and examinee characteristics.

In the first study, Wise et al. (2009) used hierarchical generalized linear modeling to examine item and examinee characteristics that were predictive of an examinee engaging in solution behavior on a typical item on a low-stakes cognitive test assessing quantitative and scientific reasoning skills. Results indicated several item characteristics including item length, item position, the presence of item graphics, number of response options, and an item position-by-item graphic interaction were significantly related to the log-odds of engagement in solution behavior on a typical item. For instance, the second

strongest predictor of engagement in solution behavior was item position. Specifically, controlling for other item characteristics, as an item's position increased on the test (i.e., appeared later), the log-odds of engaging in solution behavior decreased (β = -.279). In other words, examinees were more likely to exhibit rapid-guessing behavior to items appearing later on a test than on items appearing sooner. Interestingly, controlling for other examinee characteristics, academic ability (as measured by an aggregated SAT Verbal and Quantitative score) was the only examinee characteristic significantly related to the log-odds of engaging in solution behavior (β = .552; Wise et al., 2009). In addition, academic ability was also the strongest predictor of engaging in solution behavior. This finding is contrary to other studies that have found a small or nil relationship between RTE and independent measures of academic ability (e.g., Wise & DeMars, 2010).

In the second study, Setzer et al. (2013) used a three-level hierarchical generalized linear model to examine how much variability in the log-odds of exhibiting solution behavior on an item could be attributed to examinees and how much variability could be attributed to the examinees' institutions. Analyzing data collected from 10,004 college students attending 114 institutions, Setzer and his colleagues (2013) found 41.7% of the variability in the log-odds of exhibiting solution behavior on a low-stakes test was due to variation across examinees, whereas 14.6% of the variability in the log-odds of exhibiting solution behavior was due to variability across college institutions. The proportion of variation in the log-odds of an examinee exhibiting solution behavior due to items was not estimated due to the dichotomous nature of the criterion variable. In summary, studying the relationship between rapid-guessing behavior and item and examinee characteristics is helpful for test developers and practitioners who want to modify a test in

order to reduce rapid-guessing behavior and are interested in explaining why examinees vary in their test-taking motivation.

Solution behavior patterns across items. The third use of the SB index has been to study patterns of examinee test-taking behavior (Pastor et al., 2015; Strickman, Pastor, & Ong, 2015; Wise & Kingsbury, 2015). Evaluating how examinees' response behaviors change during a test provides valuable information for researchers who wish to use a measurement model that assumes examinees' response behaviors follow a specific pattern. For example, the Threshold Guessing model is a modified IRT model developed by Cao and Stokes (2008) that assumes (a) all examinees begin the test motivated, (b) at some point some of the examinees suddenly lose motivation and start exhibiting rapid-guessing behavior, and (c) once these examinees lose motivation, it cannot be recovered. In other words, these examinees will abruptly switch from exhibiting solution behavior to exhibiting rapid-guessing behavior on an item and continue to do so for the remaining items on the test. Wise and Kingsbury (2015) recently examined the test-taking behavior of examinees in primary school (grades 2 – 12) completing a low-stakes test of math and writing proficiency. The researchers found the examinees' behavior response patterns did not follow the pattern assumed by the Threshold Guessing model, thus indicating it would be inappropriate for practitioners to use the Threshold Guessing model with that sample.

Given the SB index reflects examinee behavior at the item level, the index can also be used to see if distinct patterns of solution behavior emerge from a group of examinees. Based on the research previously reviewed, we know examinees vary in the degree to which they exhibit solution behavior. But how much do they vary from one

another? It is not unreasonable to suspect that one group of examinees would exhibit solution behavior on the beginning of the test but then generally decrease in motivation over time and begin to display rapid-guessing behavior. Similarly, a second group of examinees may exist who are motivated and exhibit solution behavior towards all of the test items and a third group of examinees may exist who are unmotivated and exhibit rapid-guessing behavior on all of the test items.

Recently, Pastor et al. (2015) used latent class analyses to see if classes of examinees displaying similar test-taking effort patterns (yet distinctly different from other classes) could be uncovered from a group of entering freshmen and upperclass college examinees completing a 50-item low-stakes test assessing their "knowledge of environmental stewardship principles, issues, and practices" (Pastor et al., 2015, p. 4). Latent class analysis is a model-based technique that assumes a mixture of distributions, or classes, exist in the observed data and that individuals' class membership is unobservable, or latent (McLachlan & Peel, 2000). To provide validity evidence for the distinctiveness of the classes, the authors also examined if the uncovered classes exhibited differential relationships with theoretically related external variables such as gender, class status, testing-session attendance, and self-reported effort scores.

Based on the results, a two-class solution was championed where one class contained approximately 91% of the examinees and exhibited solution behavior at a consistent rate across the majority of test items. In contrast, the second class contained 9% of examinees and exhibited solution behavior at varying rates across the items on the test, primarily starting high but then progressively decreasing as the test continued. Specifically, the probability examinees in the second class engaged in solution behavior

was greater than 50% for the first 25 items on the test and decreased on the last half of the test items. As observed by the authors, there were a few items towards the end of the test where the probability examinees in the second class engaged in solution behavior spiked. Upon reviewing the items, the authors determined the examinees' probability of engaging in solution behavior on these items may have spiked due to inherent interest in the items' content (Pastor et al., 2015).

External validity evidence indicated that examinees in the first class who exhibited solution behavior consistently toward the majority of the test items were more likely to be female, freshmen, attendees of the originally scheduled testing session, and have higher self-reported effort scores (Pastor et al., 2015). In contrast, examinees in the second class were more likely to be male, upperclass college students, attendees of a make-up testing session, and have lower self-reported effort scores than the other class. In addition, examinees in the second less motivated class were more likely to complete the test in a significantly shorter amount of time and perform significantly worse on the test than examinees in the other class (Pastor et al., 2015).

Strickman et al. (2015) extended the research originally conducted by Pastor et al. (2015) and examined if classes of examinees completing two different tests emerged based on their solution behavior patterns on the tests. Specifically, two mixture models were conducted separately for each test; one test assessed examinees' sociocultural knowledge and the second test assessed examinees' quantitative and scientific reasoning skills (Strickman et al., 2015). The researchers championed a two-class solution for examinees completing the sociocultural and quantitative and scientific reasoning tests, where 95.2% and 91% of examinees, respectively, were classified in the first class and

4.8% and 9% of the remaining examinees, respectively, were classified in the second class. The classes across both tests exhibited patterns similar to the two classes in Pastor et al. (2015). That is, across both tests, the majority of examinees in the first class displayed a high probability of engaging in solution behavior on the tests' items whereas examinees in the second class displayed a much more variable pattern in their engagement in solution behavior. For example, across both tests, the probability examinees in the second class exhibited solution behavior on the tests' items decreased as the tests progressed; this pattern was more pronounced for the sociocultural test than for the quantitative and scientific reasoning test. There were several instances on both tests where the probability examinees in the second class exhibited solution behavior on an item spiked, thus leading the researchers to conclude the spikes suggest examinee motivation could potentially be recovered by placing short and interesting items towards the end of a test (Strickman et al., 2015).

Similar to the findings of Pastor et al. (2015), across both tests the second more unmotivated class of examinees were more likely to be upperclass college students completing the test during a makeup session, have lower self-reported effort and importance scores, complete the test faster and perform significantly worse on the test than examinees in the more motivated class. However, in contrast to previous findings, the two classes of examinees completing the quantitative and scientific reasoning test were not differentiated by gender composition although examinees completing the sociocultural test were.

***Effort-monitored tests*.** The fourth way the SB index has been used has been to monitor examinee effort during a testing session. For example, in an experiment designed

to increase examinee motivation on a low-stakes quantitative and scientific reasoning skills test, Wise et al. (2006) used previously defined SB thresholds to monitor examinees' test-taking behavior. Specifically, upperclass college students attending a makeup testing session were randomly divided into a treatment group and control group. Examinees in the treatment group received warning messages if they started exhibiting rapid-guessing behavior while taking the tests whereas examinees in the control group did not receive warning messages. Two different warning messages were administered to the treatment group when they displayed rapid-guessing behavior. The first warning message stressed the importance of the test to the university whereas the second warning message examinees received (if they displayed rapid-guessing behavior a second time) was more direct and threatened they would be required to attend another make-up testing session if they did not put forth more effort.

Results of the experiment showed examinees in the treatment group (i.e., warning group) had significantly higher RTE scores on the test than did examinees in the control group (d = .32; Wise et al., 2006). In addition, examinees in the treatment group who received a first warning and those who received a second warning had significantly higher RTE scores than did examinees in the control group who deserved either a first or second warning but did not receive them (d = .78, d = .83, respectively). This study highlights the use of the SB index to monitor examinee motivation and has major implications for test users concerned about the impact low-motivation has on the validity of test scores. Specifically, by delivering warning messages to examinees exhibiting rapid-guessing behavior, the experimenters were able to reduce the negative impact low examinee effort has on test scores and their relationships with related criteria.

***Effort-moderated IRT model*.** Finally, the fifth use of the SB index has been to incorporate the index into a measurement model to mitigate the impact noneffortful responding has on test scores (e.g., DeMars & Wise, 2010; Wise & DeMars, 2006; Wise & Kingsbury, 2015). Realizing examinee effort impacts the accuracy of test results, Wise and DeMars (2006) developed the effort-moderated item response theory (IRT) model to account for differences in examinees' response behavior at the item level. Specifically, the probability an examinee correctly answers an item depends on the response strategy employed: if an examinee answers an item with rapid-guessing behavior then the probability the item was answered correctly is independent of examinee proficiency and will be close to what is expected by chance. However, if an examinee exhibits solution behavior when answering an item, then the probability the item was answered correctly will increase as examinee proficiency increases and will be much higher than what is expected by chance (Wise & DeMars, 2006).

The effort-moderated IRT model represents a combination of the two item response functions which reflect the probability of an examinee correctly answering an item depending on what type of behavior they exhibit (DeMars & Wise, 2010). Specifically, the effort-moderated IRT model is specified as

$$P_i(\theta) = \left(SB_{ij}\right)\left(c_i + (1 - c_i)\left(\frac{e^{Da_i\left(\theta_j - b_i\right)}}{1+e^{Da_i\left(\theta_j - b_i\right)}}\right)\right) + \left(1 - SB_{ij}\right)(g_i), \quad (4)$$

where the probability examinee *j* correctly answers item *i* is a function of response behavior, SB_{ij}. If an examinee exhibits solution behavior ($SB_{ij} = 1$) then the model simplifies to a traditional three-parameter logistic (3PL) IRT model where D is a scaling constant, a_i is item *i*'s discrimination parameter, b_i is item *i*'s difficulty parameter, c_i is the

lower asymptote of item *i*, and θ_j is the latent variable (e.g., examinee ability) being measured for individual *j*. However, if an examinee exhibits rapid-guessing behavior (SB_{ij} = 0), then the model simplifies to g_i, which is a constant equal to the reciprocal of the total number of response options for item *i*.

In the first part of a two-part study and using the same data set Wise (2006) used, Wise and DeMars (2006) found the effort-moderated IRT model fit the data better for more examinees than a standard 3PL IRT model, regardless of whether rapid-guessing behavior was present or not. For example, the authors compared the fit of the effort-moderated IRT model to the 3PL IRT model using a likelihood ratio approach and found the fit of the observed response patterns was better for 83% of the examinees who exhibited solution behavior on 100% of the test items (RTE = 1.00) and 69% of examinees who exhibited solution behavior on less than 100% of the items (RTE < 1.00; Wise & DeMars, 2006, p. 26). Wise and DeMars (2006) also found the effort-moderated IRT model yielded less biased and more precise item parameter estimates and "generated proficiency estimates with higher convergent validity" than the standard 3PL IRT model did (p. 29). Specifically, in comparison to the effort-moderated IRT model, the 3PL IRT model overestimated the item difficulty and discrimination parameters and this overestimation occurred particularly for easier and more discriminating items. In other words, an interaction effect appeared such that relative to the effort-moderated IRT model, the standard 3PL model overestimated the item difficulty parameters when the items were relatively easy (i.e., had difficulty parameters less than 0.00) and it overestimated the item discrimination parameters when the items were both easy and more discriminating (i.e., had discrimination parameters > .5). In addition, the test

information function for the 3PL model was much higher than the test information function for the effort-moderated model, thus indicating the presence of rapid-guessing behavior in low-stakes testing may artificially inflate reliability estimates.

In the second half of their study, Wise and DeMars (2006) used simulated data based on the characteristics of the data used in first part of the study to examine the extent to which rapid-guessing behavior distorts the accuracy of the parameter estimates and test information functions under both models. The results indicated the 3PL IRT model yielded more biased and less accurate item parameters than the effort-moderated IRT model (Wise & DeMars, 2006). In addition, as the percent of examinees displaying rapid-guessing behavior increased, the amount of positive bias in the item parameters estimated using a 3PL model increased at a greater rate relative to the small amount of negative bias exhibited in the item parameters estimated using the effort-moderated IRT model. Overall, the absolute magnitude of the bias displayed in the item parameters estimated using the 3PL model was substantially higher than the absolute amount of bias displayed in the item parameters estimated using the effort-moderated IRT model. The results also indicated that the 3PL IRT model overestimated the reliability of the proficiency estimates whereas the effort-moderated IRT model slightly underestimated them; moreover, this discrepancy increased as the proportion of examinees exhibiting rapid-guessing behavior increased (Wise & DeMars, 2006). As indicated by these results, given the effort-moderated IRT model yields less biased, more reliable, and more valid scores than a traditional 3PL IRT model, it should be used by practitioners when low examinee motivation is a concern.

In a separate study, DeMars and Wise (2010) used the effort-moderated IRT model and found evidence that examinees displaying various amounts of effort on a test can lead to items being flagged as displaying differential item functioning (DIF). For example, the first part of a two-part study used simulated data based on the data used by DeMars (2007). Specifically, the same item parameters were used to generate data for two groups, with one group simulated to engage in solution behavior on all items (more-motivated group) and another group simulated to rapidly guess on some items (lower-motivated group). The authors found 18% of items were flagged as displaying DIF which favored the more-motivated group of examinees (all with RTE = 1.00) over the lower-motivated group (with RTE < 1.00, mean RTE = .845), even though the item parameters were simulated to be equal across both groups. Specifically, holding other predictors constant, items with lower RTF values (i.e., examinees displaying more rapid-guessing behavior) were more likely to be flagged as displaying DIF. Similarly, more discriminating items and easier items were also more likely to be flagged as displaying DIF and favoring more-motivated examinees over less-motivated examinees.

In the second part of the two-part study, DeMars and Wise (2010) used simulated data to examine the impact of differential guessing across gender. Although males were simulated to exhibit more rapid-guessing behavior on average than females (mean RTE_{Males} = .80, mean $RTE_{Females}$ = .90, respectively), the gender difference was simulated to vary across items, with gender differences in RTFs ranging from .05 to .16. Compared to the first study, the researchers found that when there was less of a discrepancy in rapid-guessing behavior across groups, fewer items were flagged as displaying DIF across males and females. Specifically, only 8% of items were flagged as displaying DIF

favoring females over males. Similar to the first study, they also found that easier items and items displaying more rapid-guessing differences between groups were more likely to be flagged as displaying DIF. The results of this two-part study indicates instances of DIF occurring on low-stakes tests could be attributable to examinees displaying differential rapid-guessing behavior towards test items.

As reviewed, the SB index has been used in a multitude of ways to study examinee effort at either the test level or at the item level. When used on its own, the SB index has been used for a number of purposes, ranging from identifying item and examinee characteristics related to rapid-guessing behavior to examining the impact differential rapid-guessing has on item functioning. Although the SB index has been extensively used, there has been little study of how to empirically define the time thresholds. That is, the majority of the studies that have used the SB index with low-stakes cognitive tests have focused on studying examinee effort rather than examining the effectiveness of the SB index and the various methods used to calculate the time thresholds. Given the difference between items on cognitive tests and noncognitive measures, it is anticipated in the current study that calculating time thresholds for the SB index when applied to a noncognitive measure may be difficult. Thus, it is important to understand how the time threshold calculation methods have been defined and empirically studied when applied to cognitive tests, as these methods may be useful to apply with noncognitive measures. Therefore, the following section first reviews various calculation methods used to define the solution behavior time thresholds and then reviews research comparing the threshold calculation methods and evaluating their effectiveness when applied to low-stakes cognitive tests.

Methods Used to Define the Solution Behavior Time Thresholds

Although the SB index has been used in a variety of ways to study examinees' behavior while taking low-stakes cognitive tests, there remains the challenge of defining the solution behavior time thresholds for each test item. As described by Wise and Ma (2012), there are two competing principles that must be considered when defining a threshold: "First, it is desirable to identify as many instances of non-effortful item responses as possible. Second, it is important to avoid classifying effort responses as non-effortful." (p. 7). Thus, while it is desirable to identify all noneffortful responses, it is prudent to be conservative and not classify an examinee's response as a rapid-guessing response when in fact it was a solution behavior response. Considering these challenges, multiple methods have been developed to define time thresholds for items and are reviewed in detail below.

Two-class lognormal mixture model. In their initial work on speeded tests, Schnipke and Scrams (1997) applied a two-class lognormal mixture model to items' response time distributions to classify examinees based on their item-response strategy. Schnipke and Scrams (1997) posited that examinees exhibiting solution behavior could be distinguished from examinees exhibiting rapid-guessing behavior by an item's response time distribution. Specifically, the authors hypothesized and found an item's response time distribution would appear bimodal, where the lower mode of the distribution was reflective of those examinees exhibiting rapid-guessing behavior and the upper mode of the distribution reflected those examinees exhibiting solution behavior.

Mathematically, a two-class mixture model can be expressed as,

$$F_{Oi} = \rho_i F_{Gi} + (1 - \rho_i) F_{Si}, \tag{5}$$

where F_{Oi} is the observed response time distribution for item i, F_{Gi} is the rapid-guessing response time distribution for item i, F_{Si} is the solution behavior response time distribution for item i, and ρ_i is the mixing proportion, or the proportion of the population characterized by the rapid-guessing response time distribution for item i (Schnipke & Scrams, 1997). Because Schnipke and Scrams (1997) posited the response time distribution for each class would be positively skewed, each class is assumed to follow a lognormal distribution. Therefore, when paired with the two-class mixture model formula expressed in Equation 5, the two-class lognormal mixture model can be expressed as,

$$F_{Gi} = \frac{1}{\sqrt{t\sigma_{Gi}(2\pi)}} exp\left[\frac{-[ln\left(\frac{t}{m_{Gi}}\right)]^2}{2\sigma_{Gi}^2}\right], F_{Si} = \frac{1}{\sqrt{t\sigma_{Si}(2\pi)}} exp\left[\frac{-[ln\left(\frac{t}{m_{Si}}\right)]^2}{2\sigma_{Si}^2}\right], \quad (6)$$

where t is the response time for item i, m_{Gi} and m_{Si} are the scale parameters for the rapid-guessing behavior class and solution behavior class, respectively, and σ_{Gi} and σ_{Si} are the shape parameters for the rapid-guessing behavior and solution behavior class, respectively (Schnipke & Scrams, 1997; Yang, 2007). The scale parameters, m_{Gi} and m_{Si}, are equal to the median natural log of response time for item i whereas the shape parameters, σ_{Gi} and σ_{Si}, are equal to the standard deviation of the natural log of the response time. Because two classes of examinees are assumed to underlie the observed response time distribution and because mixing proportions are constrained to sum to one across classes, only one mixing proportion representing the proportion of examinees in rapid-guessing class needs to be estimated. Thus, a total of five parameters are estimated for each item by a two-class lognormal mixture model: ρ_i, m_{Gi}, m_{Si}, σ_{Gi}, and σ_{Si}. After assessing and championing the fit of the two-class model, Schnipke and Scrams (1997)

defined the solution behavior time threshold by identifying the point at which the two distributions intersected.

To date, this method has been used in only two empirical studies to identify the solution behavior thresholds when applied to low-stakes cognitive tests (Kong et al., 2007; Pastor et al., 2015). Specifically, Kong et al. (2007) used this method to identify the time thresholds for a 60-item low-stakes test assessing upperclass college students' information literacy skills and knowledge. In addition, Pastor et al. (2015) also used two-class lognormal mixture models (albeit in a slightly differently way).[1] Although useful, one reason why this threshold calculation method has not been used more frequently is because it is complex and difficult to apply (Wise & DeMars, 2006).

Visual inspection of an item's response time distribution. A second method that has often been used to define an item's time threshold is to visually inspect an item's response time distribution. As previously described, when examinees completing a low-stakes unspeeded test exhibit a combination of solution behavior and rapid-guessing behavior on an item, a bimodal distribution representing the two groups should emerge, whereby the lower mode reflects the examinees exhibiting rapid-guessing behavior and the upper mode reflects the examinees exhibiting solution behavior. The time threshold is then commonly determined by visually identifying the time at the upper end of the distribution with the lower mode. For example, consider the item response time distribution presented in Figure 1. In this example, the solution behavior time threshold appears to occur at 8 seconds. Thus, examinees who answered this item in 8 seconds or less would be classified as exhibiting rapid-guessing behavior whereas examinees who took longer to respond would be classified as exhibiting solution behavior.

Given its simplicity, the visual inspection method has been used multiple times (e.g., DeMars, 2007; Pastor et al., 2015; Setzer et al., 2013; Stickman et al., 2015; Swerdzewski et al., 2011; Wise et al., 2006; Wise & DeMars, 2006, 2010; Wise & Kong, 2005; Wise et al., 2009). For example, Wise et al. (2009) visually inspected the response time distributions for each of the 64 items on a low-stakes quantitative and scientific reasoning skills test and identified time thresholds ranging from three to 15 seconds in length; the median threshold was 4.4 seconds (Wise et al., 2009). In another study, Setzer and colleagues (2013) visually inspected 120 response time distributions which yielded thresholds ranging from two seconds to ten seconds (the median threshold was equal to four seconds; Setzer et al., 2013).

Item surface features. A third method commonly used to set thresholds is to take an item's surface features into consideration. Surface features are features of an item that add length to the amount of time required by an examinee to carefully read, comprehend, and answer an item. More specifically, they refer to item characteristics such as length (as measured in the number of characters or words), position (i.e., an item's position on a test relative to other items), difficulty, and the presence of ancillary materials such as a graphic or figure. Intuitively, it makes sense that the longer in length an item is the longer it will take an examinee to read the item and consider a response. Item surface features have been used in multiple studies to set time thresholds (e.g., Kong et al., 2007; Wise & Kong, 2005). For example, Kong et al. (2007) used an item's length (as measured in characters) and the presence of ancillary reading material (e.g., graph or figure) when determining items' time thresholds. The authors set a three-second threshold for items less than 200 characters in length, a five-second threshold for items between 200 and

1000 characters in length, and a ten second threshold for items longer than 1000 characters in length or for items that contained new ancillary reading material.

Common threshold. A fourth method that has been used to set solution behavior time thresholds has been to adopt a common threshold that is applied to all items on a test. Although this method is less realistic as it fails to accommodate for varying item characteristics such as item length, researchers have used it because it is the easiest threshold to implement, it is useful when working with large item pools generated for computer adaptive testing, and it is useful when researchers do not have access to the items themselves (e.g., Kong et al., 2007; Wise, Kingsbury, Thomason, & Kong, 2004).

Normative threshold. More recently, a fifth method known as the normative threshold (NT) model has been developed to identify solution behavior time thresholds for items administered via computer adaptive testing (Wise & Ma, 2012). Computer adaptive testing presents a challenge for researchers trying to establish a time threshold because items are selected from a large item pool that may contain hundreds or thousands of items, thus making the use of other methods such as visual inspection impractical. Moreover, examinees taking computer adaptive tests see different sets of items. Unlike imposing a constant threshold across all items, the NT method is a variable identification method because the thresholds vary across items (Wise & Ma, 2012). Specifically, the NT method defines the time threshold for an item as a percentage of the average amount of time examinees respond to an item. For example, a 10% threshold (NT10) for an item with an average response time of 50 seconds would be 5 seconds. In contrast, a 20% threshold (NT20) for an item with an average response time of 50 seconds would be 10 seconds. Researchers using the NT method have often set a maximum time threshold to

serve as an upper bound. For example, Wise and Kingsbury (2015) used a 10% threshold (NT10) and imposed a maximum time threshold of 10 seconds. Thus, if an item's threshold was calculated to be 12 seconds, the threshold for that item would be adjusted downward to be 10 seconds. The NT model has been used in several studies (e.g., O. L. Liu, Rios, & Borden, 2015; Rios et al., 2014; Wise & Kingsbury, 2015; Wise & Ma, 2012; Wise, 2015). Similar to the common threshold method, the NT method is useful to use when researchers do not have access to the wording of the items.[2]

Do the methods yield similar results? As reviewed, multiple threshold calculation methods have been developed to identify the solution behavior time thresholds used to classify examinees' item-response behavior. However, despite their existence and extensive use in the empirical literature, there has not been much research comparing the threshold calculation methods to one another and their efficacy in classifying respondents with the SB index. To date, only three studies have empirically compared different methods used to identify the solution behavior time thresholds (Kong et al., 2007; Pastor et al., 2015; Wise & Ma, 2012). Specifically, Kong et al. (2007) compared the performance of lognormal mixture models, visual inspection, surface features, and using a common three-second threshold to one another; Wise and Ma (2012) compared three different NT percentage levels (NT10, NT15, and NT20) and a common threshold to one another; and Pastor et al. (2015) compared the performance of the visual inspection method to using two-class lognormal mixture models.

Across the three studies, the researchers took different approaches to comparing the threshold calculation methods; all three studies compared the methods using real data – simulated data were not used in the comparisons. For instance, Kong et al. (2007)

examined the level of agreement among time thresholds across the four methods. The researchers found the time thresholds defined by the mixture modeling and visual inspection methods exhibited the highest level of agreement (37% of the thresholds were in exact agreement, 97% were in agreement within three seconds of one another). In addition, the time thresholds defined by the surface feature method exhibited the second highest level of agreement with the time thresholds defined by the mixture modeling and visual inspection methods (30% of the thresholds were in exact agreement and 87% of the thresholds were in agreement within three seconds of one another).

Kong et al. (2007) also compared differences in the resulting RTE scores calculated from the SB classification indices using criteria previously put forth by Wise and Kong (2005) as validity evidence for RTE scores. Specifically, Wise and Kong (2005) hypothesized (a) RTE should demonstrate adequate reliability, (b) RTE should demonstrate evidence of convergent validity by being related to other measures of test-taking effort, (c) RTE should demonstrate evidence of discriminant validity by not being related to independent measures of academic ability, (d) responses answered using rapid-guessing behavior should have an accuracy rate close to chance, and (e) RTE should exhibit motivation filtering effects (i.e., after removing examinees exhibiting low effort, test performance should improve, score variability should decrease, correlations with related measures should increase, and reliability estimates should decrease).

In regards to the five validity criteria for RTE scores, Kong et al. (2007) found the performance of the three variable methods (i.e., mixture modeling, visual inspection of response time distribution, and surface features) generally performed slightly better than the constant threshold method; however, in general all of the results were very similar.

The researchers concluded that although using the mixture modeling method would "most likely be the most psychometrically rigorous method" (Kong et al., 2007, p. 618), researchers should choose the time threshold calculation method that is most appropriate given the information they have about the items and the response times.

In contrast to Kong et al. (2007), Wise and Ma (2012) compared the three different NT percentage levels and a common three-second threshold by examining their impact on non-credible growth scores (i.e., implausible growth scores such as negative or extreme-positive growth scores) and the accuracy of response rates. The accuracy of response rates refers to the rate responses (i.e., items) were correctly answered: responses classified as solution behavior were expected to have an accuracy rate close to 50% (which was expected given the test was a computer-adaptive test) whereas responses classified as rapid-guesses were expected to have an accuracy rate close to chance, which was equal to either 20% or 25% depending on whether an item had five or four responses, respectively. The researchers found each of the NT methods identified more non-credible growth scores than the constant threshold did, but as the percentage of the NT methods increased, so did the misclassification of responses as indicated by the accuracy rates of the responses. That is, as more responses were classified as rapid-guesses, the accuracy rates for the responses classified as rapid-guesses were greater than what was expected by chance alone, thus indicating effortful responses were being misclassified as rapid guesses. Ultimately, the authors championed using the NT10 method as it "maintained accuracy for solution behaviors and rapid-guessing behaviors at their expected rates" (Wise & Ma, 2012, p. 16).

Finally, Pastor et al. (2015) took a completely different approach than the other two studies and compared the proportion of examinees classified as exhibiting solution behavior by the visual inspection and mixture modeling methods for each item on a 50-item test. Similar to Kong et al. (2007), the researchers found very minor differences between the two methods. For example, the largest difference in the proportion of examinees classified as exhibiting solution behavior by the two threshold calculation methods was .08. Specifically, the visual inspection method classified 94% of examinees as exhibiting solution behavior on an item whereas the mixture modeling method classified only 86% of examinees as exhibiting solution behavior on the same item.

As reviewed, researchers have used a variety of time threshold calculation methods to define the solution behavior time thresholds when applied to low-stakes cognitive tests. Based on these results, it appears using a variable threshold calculation method such as visually inspecting items' response time distributions, mixture modeling, or the NT method is preferable to using a common threshold, and using a more conservative threshold is preferable to using a more liberal one. It is difficult to tell, however, if one threshold calculation method should be used over another given the different approaches researchers have taken to providing validity evidence for the time thresholds and resulting SB classification indices. The majority of research using the SB index on low-stakes cognitive tests has primarily relied on Kong et al. (2007) as providing evidence for the validity of the threshold calculation method used (e.g., Pastor et al., 2015; Setzer et al., 2013; Wise & DeMars, 2010). However, this is the only study to have thoroughly compared multiple threshold calculation methods commonly used (e.g., visual inspection, mixture modeling, and surface features).

When classifying responses to items, it is important to keep in mind that classifying a response as a solution behavior response is more ambiguous than classifying a response as a rapid guess (Kong et al., 2007). That is, classifying an examinee's response to an item as a solution behavior response does not necessarily indicate the examinee answered that item with effort: it only indicates the examinee did not rapidly respond to that item. In other words, "no claim is being made that the SB index identifies all noneffortful responses; rather, it identifies only those of which we are reasonably certain" (Kong et al., 2007, p. 608). Because there is no way to prove an examinee is exhibiting solution behavior when responding to an item, the responsibility of providing validity evidence indicating the thresholds are correctly classifying examinees' response behavior falls on researchers' shoulders (Wise, 2015). Given this, researchers have provided a plethora of evidence for the validity of the SB index, RTE scores, and RTF scores by comparing known-groups' test performance when classified by motivation (e.g., Wise & DeMars, 2010), examining the relationship between motivation and related item and examinee characteristics (e.g., Wise et al., 2009), and comparing the performance of RTE scores to self-reported effort scores (e.g., O. L. Liu et al., 2015; Rios et al., 2014; Swerdzewski et al., 2011; Wise & Kong, 2005). However, despite this research, little research has been performed in regards to using solution behavior to identify responses made without effort on low-stakes noncognitive measures.

Previous Research Using the Solution Behavior Index with Noncognitive Measures

To date, the majority of research using the SB index has primarily focused on low-stakes cognitive tests administered for accountability purposes. To my knowledge, only one study has used the SB index with low-stakes noncognitive measures.

Specifically, Swerdzewski et al. (2011) used the SB index with two cognitive tests and four noncognitive measures and then calculated and compared RTE scores to self-reported test-level effort scores. The researchers defined the solution behavior time thresholds by visually inspecting items' response time distributions and identifying the point at which the two distributions intersected; the thresholds were then cross-validated by comparing the threshold value to the minimum amount of time it took the researchers to read an item and its response options. Because only one self-report measure of effort was administered at the end of the testing session, the self-reported examinee effort score was a test-session level or global measure of effort. That is, it reflected how much effort examinees put forth on all of the tests during the testing session.

In general, the purpose of the researchers' study was not to evaluate the efficacy of applying the SB index to noncognitive measures, but rather to evaluate the correspondence of using RTE scores and a self-report global measure of effort on test scores after filtering out unmotivated examinees. When comparing how the methods classified examinees as either motivated or unmotivated, Swerdzewski and his colleagues (2011) found the self-reported measure of test session effort (effort across all tests) was in agreement with the test-level RTE scores 65% to 69% of the time. When the two methods were not in agreement, RTE was considered more conservative than the self-report measure classifying examinees as displaying low motivation on both cognitive and noncognitive measures. This pattern was also displayed – although slightly – when comparing the performance of filtered and unfiltered examinees on the noncognitive tests' subscales. Using a criteria of $d \geq .10$ to indicate a practical difference, the researchers found when examinees were filtered using the global test session level self-

reported effort score, they displayed a practical difference in their scores from unfiltered examinees on three of the 25 noncognitive subscales (12%). In contrast, when examinees were filtered using RTE, they exhibited a practically different score from unfiltered examinees on only one noncognitive subscale (4%). Although RTE appears to be more conservative than self-reported effort, these results should be cautiously interpreted since this is the only study known of to compare the validity of using self-reported effort scores versus RTE to identify and remove examinees displaying low motivation on low-stakes noncognitive measures.[3] Although this study provides supportive validity evidence for the use of the SB index with noncognitive measures, further research is needed given its intention was not to evaluate the use of the SB index with noncognitive measures, but rather compare its performance to a self-report measure of effort.

Conclusion

In summary, the SB index provides researchers with a practical and discreet way to identify and study examinees responding to items without effort on low-stakes tests. Because the SB index values are calculated using items' response times, they provide an unobtrusive way to assess examinees' effort at the item level, which is a major advantage the SB index has over other measures of effort such as self-report measures. Moreover, the SB index can be used in a multitude of ways including (a) calculating test-level measures of examinee effort, (b) studying item and examinee characteristics related to item-response behavior, (c) examining patterns of item-response behavior across items, (d) monitoring test-taking behavior during the testing process, and (e) using the effort-moderated IRT model to yield more accurate and less biased parameter estimates.

However, despite the advantages of using the SB index to covertly identify items answered without effort and despite its utility in answering a wide array of research questions, the SB index has rarely been used to identify and study noneffortful responses made by students completing low-stakes noncognitive measures administered for accountability purposes. The only study known to use the SB index with noncognitive measures did not thoroughly evaluate the use of the index with the noncognitive measures and only used one method to calculate the time thresholds (Swerdzewski et al., 2011). Given the differences between items on cognitive tests and noncognitive measures, it is unclear whether the SB index can be used with noncognitive measures and if various time threshold calculation methods can be used with noncognitive measures. Specifically, some of the time threshold calculation methods such as visual inspection and lognormal mixture modeling assume an item's response time distribution appears bimodal. Because items on noncognitive measures are typically shorter in length and less complex than items on cognitive tests, the distribution of items' response times on noncognitive measures may be shorter, have less variability, and may not appear bimodal. If an item's response time distribution is not bimodal, then time threshold calculation methods that assume bimodality will fail to yield a defined time threshold. In contrast to calculation methods that may not yield a time threshold, other calculation methods will always yield a time threshold, even if there are not two distinct types of responders responding to an item (i.e., motivated and unmotivated). For example, the NT method will always yield a defined time threshold because it is based on a percentage of an item's overall response time. Because some threshold calculation methods such as the NT method will always identify a time threshold, it is important to gather external

validity evidence to determine if the resulting groups are distinct and meaningful. Given the dearth of research applying the SB index to noncognitive measures and based on these considerations, further research examining the application of the SB index to noncognitive measures and its effectiveness in identifying noneffortful responses is needed.

Purpose of the Study

The purpose of the current study was to examine if the SB index could be used to identify rapid responses to items on a low-stakes noncognitive measure when calculated using various time threshold calculation methods, and if so, if the resulting time thresholds and SB classifications were meaningful. Specifically, the purpose of the current study was threefold.

The first purpose of the study was to examine whether various time threshold calculation methods could be used to define the time thresholds and identify rapid responses to a 53-item noncognitive measure assessing the construct meaningful life. Because the SB index has only been used with low-stakes noncognitive measures once, it was important to determine if different time threshold calculation methods besides the visual inspection method used by Swerdzewski et al. (2011) could be used with noncognitive items. Thus, a total of eight methods were used to calculate the time thresholds: (1) visual inspection of the response time distributions, (2) lognormal mixture modeling, (3) NT10, (4) NT20, (5) NT30, (6) reading speed, (7) visual inspection with information, and (8) mixture modeling with information.

The last two threshold calculation methods were used to examine if including data from a known group of rapid responders would impact these methods' ability to calculate

time thresholds. A preliminary investigation of the response time distributions using data from the Makeup Testing 2015 sample (described in further detail below) revealed that some items did not have a clear bimodal response time distribution. As a result, it was of interest to examine whether including data from a known group of rapid responders would alleviate these problems and affect the calculation of the time thresholds. That is, would adding data from a known group of rapid responders make the existing groups of responders (motivated and unmotivated) appear more distinct? Therefore, the visual inspection and lognormal mixture modeling threshold calculation methods were conducted twice: once using a primary sample of data and again using an expanded data sample that included known rapid responders. To distinguish the results from each other, the two threshold calculation methods using a known group of rapid responders are referred to as visual inspection *with information* and lognormal mixture modeling *with information*. If a time threshold could not be defined by any of the threshold calculation methods, then the time threshold for that item was set to missing.

The second purpose of the study was to (a) examine if the threshold calculation methods were able to successfully define time thresholds and if so, (b) examine if the resulting time thresholds and subsequent solution behavior classification indices varied across threshold calculation methods. Because this is the first study to thoroughly examine the application of the SB index to noncognitive measures, it was considered just as important to see if the time thresholds *could* be calculated and used to identify solution behaviors on noncognitive items as it was important to see if the time threshold calculation methods performed differently from one another. It was anticipated that the visual inspection and lognormal mixture modeling methods might fail to yield defined

time thresholds for some items whereas the NT and reading speed methods would *always* yield a time threshold. Similarly, it was anticipated that the time thresholds and proportion of respondents classified as exhibiting solution behavior using higher NT percentage levels (e.g., NT30) would be greater than the time thresholds and amount of respondents classified using lower percentage levels (e.g., NT10).

The third purpose of the study focused on gathering external validity evidence for the time threshold calculation methods. Specifically, it was of interest to determine if the threshold calculation methods yielded meaningful time thresholds by examining if (a) calculated RTE and RTF scores were related to respondent and item characteristics in theoretically expected ways, respectively, and (b) if the relationships differed across threshold calculation methods. Specifically, eight RTE scores were calculated (one for each threshold calculation method) and related to seven respondent characteristics, and eight RTF scores were calculated and related to two item characteristics. The individual relationships between the scores with each characteristic were then examined to determine if relationships aligned with those based on theory and past research and if using different threshold calculation methods yielded differential relationships with the characteristics being analyzed.

The respondent characteristics examined in relation to the RTE scores were: gender, makeup-testing session attendance status, and two measures of academic ability. In addition, RTE scores were also related to scores from an index commonly used in the survey literature to examine effort known as the individual consistency index, and to the length of an open-ended response question administered to a subsample of participants. Two characteristics of items on the substantive noncognitive measure of interest were

examined in relation to RTF scores: item position and length of an item (in words). A-priori hypotheses about the expected relationship between the scores and the respective respondent and item characteristics were made. Given the exploratory nature of the study, no a-priori hypotheses were made about the differential relationships that might occur among the relationships across calculation methods. The respondent and item characteristics examined and the hypothesized relationships are described below.

Respondent characteristics

Gender. As reviewed, previous research has demonstrated there is a relationship between gender and test-taking motivation whereby males tend to have lower RTE scores than females, on average (e.g., DeMars et al., 2013; Setzer et al., 2013). Thus, it was expected that men would have lower RTE scores than women, on average.

Makeup testing session attendance. Students who fail to attend a scheduled university-wide assessment day conducted for accountability purposes are required to attend makeup testing sessions. Similarly, students who fail to attend the requisite makeup testing session are required to complete the makeup testing session as a walk-in at a computer lab on campus. Previous research has shown students who attend makeup testing sessions tend to display less effort than those who attend the regularly scheduled testing session (e.g., Pastor et al., 2015; Swerdzewski, Harmes, & Finney, 2009). By extension, it is reasonable to suspect students who *fail to attend* the scheduled makeup testing session and complete the tests as a walk-in at a computer lab on campus will exhibit less effort than the students attending the originally scheduled makeup testing session. Therefore, it was expected that RTE scores for students who *failed to attend* the scheduled makeup testing session and completed the tests at a computer lab on campus as

walk-ins would be significantly lower on average than the RTE scores for students attending the original makeup testing session.[4] For clarity, students who did *not* attend the originally scheduled makeup testing session and instead completed the tests at a different time were classified as "walk-ins."

Effort. Self-report measures of effort are often used to gather information about students' motivation towards completing low-stakes tests. One self-report measure commonly used is known as the Student Opinion Scale (SOS; Sundre & Moore, 2002; Thelk et al., 2009), which contains a subscale assessing the amount of effort students put forth on a set of tests. The effort subscale on the SOS is a *global* self-report measure because it assesses the amount of effort students put forth on a *set* of tests as opposed to just one test. Previous research has found global self-reported measures of effort exhibit a positive but moderate correlation with RTE (*r*'s = .38 to .41; Kong et al., 2007). Thus, it was of interest in the current study to examine the relationship between global self-reported effort scores (which reflect how much effort was put forth on all tests during the testing session) and the calculated RTE scores (which reflect how much effort students put forth on only the substantive noncognitive measure of interest). It was expected the RTE scores and self-reported effort scores would exhibit a positive correlation low in magnitude.

Academic ability. Previous research using the SB index with low-stakes cognitive tests has often found effort is unrelated to independent measures of academic ability (e.g., Rios et al., 2014; Wise & DeMars, 2010). Because this was the first study to thoroughly study the SB index applied to noncognitive measures and given researchers have often used this nil relationship between effort and academic ability to justify the removal of

unmotivated students from the data (e.g., Wise & DeMars, 2010), it was of interest in the current study to examine if absence of a relationship between RTE and academic ability would hold for noncognitive measures. Two measures of academic ability – SAT critical reading (SAT-CR) and mathematics (SAT-M) – were used to address this question.

Individual consistency index. Consistency indices are used to evaluate how consistent a respondent is in their responses on a survey. The underlying idea behind consistency indices is if a respondent puts forth effort and truthfully answers a survey, then the responses across items should show a high degree of consistency. In contrast, if the respondent is not putting forth effort and is rapidly responding instead, then the responses will not be consistent (Curran, 2015). Semantic synonyms are a priori pairings of items with similar meanings and are designed to identify respondents who "indicate dissimilar responses to similar items" (DeSimone et al., 2015, p. 173). Within-person correlations are calculated across the item pairs and the magnitude of the correlation is used as the consistency index. Thus, higher scores are desirable and values closer to zero indicate low effort (DeSimone et al., 2015). The RTE scores were hypothesized to be positively related to the individual consistency index.

Length of response to an open-ended question. An open-ended question was included at the end of the substantive noncognitive measure of interest for a sample of participants in the current study. The question asked participants "What are three life experiences that you have had as a JMU student that will help you lead a more meaningful life after graduation?" It was hypothesized students who put forth little effort completing the substantive measure of interest would also display low effort while answering this question. In contrast, it was hypothesized students who displayed effort in

responding to the noncognitive measure of interest would also put forth answering this essay question. Thus, it was hypothesized the RTE scores would be positively related to the length of students' response (as counted by number of words).

Item characteristics

Item position. Item position refers to the serial position of an item relative to its order on a test. Previous research has demonstrated rapid-guessing behavior occurs more frequently on items occurring in later positions on low-stakes cognitive tests (e.g., Bovaird, 2002; Setzer et al., 2013; Wise, 2006; Wise et al., 2009). In addition, previous research has shown respondents put forth less effort towards the end of long measures (e.g., Baer, Ballenger, Berry, & Wetter, 1997; Meade & Craig, 2012). Given this information, it was hypothesized the RTF scores would be negatively related to item position.

Item length. Item length is defined as the total number of words an item stem contains. Research using low-stakes cognitive tests have previously found effortful responding is negatively related to item length (e.g., Wise et al., 2009). However, given noncognitive item stems are typically shorter in length and given the responses options are typically fixed for every item, it was of interest to see if item length was negatively related to the RTF scores on a non-cognitive measure. Because one of the threshold calculation methods is based on the total number of words (RSPEED), the magnitude of this relationship and how it compares to other methods was cautiously interpreted.

CHAPTER THREE

Method

Chapter Overview

As stated in Chapter Two, the purpose of the current study was three-fold: (1) to determine if eight time threshold calculation methods could be used to define the solution behavior time thresholds for items on a low-stakes noncognitive measure and if the addition of a known group of rapid responders affected the results of two of the calculation methods, (2) to compare the defined time thresholds and resulting SB index values at the item level across methods, and (3) to gather and examine external validity evidence for the resulting classifications and determine if one method should be used over another. To address the first purpose of the study, four independent samples of students were used. Specifically, three samples of students attending a university-wide assessment day and makeup-testing sessions during the spring semesters of 2015 and 2016 were combined to create the *Primary sample*. The fourth sample of students was collected primarily from an undergraduate psychology participant pool and is referred to as the *Known Rapid Responders sample*. The noncognitive measure of interest used in the current study was a 53 item measure known as the Meaningful Life Scale (MFLS) and is described in further detail below. All of the students in the Primary sample completed the MFLS under low-stakes conditions for university assessment purposes. The Known Rapid Responders sample also completed the MFLS but under instructions to complete the measure as quickly as possible. Six of the eight threshold calculation methods (visual inspection, lognormal mixture modeling, NT10, NT20, NT30, and reading speed) were calculated using the Primary sample, whereas two of the time threshold calculation

methods (i.e., visual inspection with information and lognormal mixture modeling with information) were calculated using the Known Rapid Responders sample, which was combined with the Primary sample. To address the second purpose of the study, the resulting time thresholds and SB index values from the eight calculation methods were compared. Finally, to address the third purpose of the study, external validity evidence was individually examined for each of the eight methods and then compared across methods to determine if the methods yielded similar or different results and if one method should be used over another.

Procedures and Participants

Data for the current study were collected from four independent samples. Three of the four samples completed the MFLS for university assessment purposes under low-stakes conditions and were combined to create the *Primary sample* (N = 568). The fourth sample of students completed the MFLS specifically for the purpose of the current study and is referred to as the *Known Rapid Responders sample* (N = 181). The MFLS (described in further detail below) was administered to all samples on a computer using a web-based survey program known as Qualtrics. In order to measure response time, the MFLS items were administered with one item per page; response time was defined as the total number of seconds a student spent answering an item prior to moving on to the next page. The following section describes the procedures and participants for the four samples individually: the three samples used to create the Primary sample are described first followed by a description of the Known Rapid Responders sample.

Procedures and participants for the Assessment Day sample. Undergraduate students attending the university at which the current study was conducted are required to

participate in a university-wide Assessment Day twice: once as incoming freshmen in the fall semester prior to beginning classes and again eighteen months later after accumulating 45-70 credit hours and classified as either sophomores or juniors (i.e., upperclass students). The results of the assessments are used for institutional accountability purposes and are thus considered important by administrators. Because the results do not directly affect the students, the tests are considered low-stakes to students. To facilitate attendance, classes are canceled the day of testing and an academic hold is placed on students' records if they fail to attend. Students are randomly assigned to testing rooms based on the last three digits of their student ID number and complete the same battery of cognitive and noncognitive tests during both testing occasions. To ensure compliance as well as to motivate students to respond with effort, the tests are administered under standardized conditions and are monitored by trained proctors. Prior to the beginning of the testing session, students watch a video explaining the purpose and importance of the tests and are thanked in advance by the president of the university for putting forth effort in responding to them. Testing sessions are approximately two hours in length. Most testing sessions are conducted in classrooms and are administered using a paper and pencil format. A subset of testing sessions are conducted and administered on computers in computer labs on campus.

A subset of data used in the current study were collected from 77 upperclass students who participated in the spring 2016 Assessment Day. The series of cognitive tests and noncognitive measures completed by the students, including the MFLS, is presented in Table 1. The series of assessments were administered on computers using Qualtrics and were supervised by trained proctors.

Procedures and participants for the Makeup Testing samples. Although there are valid reasons why students may fail to attend Assessment Day (e.g., illness), given the low-stakes nature of the testing environment students have a tendency to skip or not participate in Assessment Day. Previous research has shown students who skip Assessment Day and attend makeup testing sessions are older, less motivated, have lower GPAs, feel more entitled, and more psychologically reactant than students who attend Assessment Day (Brown & Finney, 2011; Kopp & Finney, 2013; Swerdzewski et al., 2009). Because the results of Assessment Day are used for accountability purposes, students are required to attend. Students who fail to attend Assessment Day have an academic hold placed on their record and are required to attend a scheduled makeup testing session in order for the hold to be removed. Three to four makeup testing sessions are typically held in the evening a few weeks after the originally scheduled Assessment Day and are conducted by trained proctors. If students fail to attend the scheduled makeup testing session, then the academic hold on their record is not removed until they complete the battery of tests at a walk-in computer lab on campus designated for assessment testing. This computer lab is staffed by trained proctors. These students are subsequently described as "walk-ins." All makeup tests – scheduled and walk-ins – are administered on Qualtrics via computers.

Makeup Testing 2015 sample. A total of 336 students attended makeup testing sessions during the spring semester of 2015; 153 of these students (46%) were classified as walk-ins. The sequence of cognitive tests and noncognitive measures completed by this sample is presented in Table 1. Students were given 30 minutes to complete the 53-item MFLS which included an additional open-ended response question at the end, which

was used to gather external validity evidence for the threshold calculation methods. Specifically, the open-ended response question asked students "What are three life experiences that you have had as a JMU student that will help you lead a more meaningful life after graduation?"

Makeup Testing 2016 sample. A total of 158 upperclass students attended makeup testing sessions during the spring of 2016. Due to time constraints of the current study, data from students who completed the assessments as walk-ins were not included in this sample. The battery of tests this sample of students completed and the order they completed them in is presented in Table 1. The version of the MFLS this sample of students completed was slightly different than the version completed by the 2015 Makeup sample – it did not include the open-ended question but did include 33 additional items that were not used in this study.

Procedures and participants for the Known Rapid Responders sample. It was of interest for the current study to have a known group of rapid responders complete the substantive measure of interest. Thus, in contrast to the first three samples, which were instructed to respond truthfully and thoughtfully to the MFLS, participants in the Known Rapid Responders sample were instructed to respond to the MFLS as rapidly as possible.

A little less than half of the participants in the Known Rapid Responders sample were recruited through the undergraduate psychology participant pool and received course credit for participating. These participants completed the measure either by attending a testing session or online. Participants who attended testing sessions completed the MFLS on a computer using Qualtrics. Prior to beginning the testing session, students were read a script containing the purpose of the study, the expected length of time, and a

statement indicating participation in the study was voluntary. After providing informed consent, students were instructed not to read and thoughtfully respond to the items, but instead to respond to the 53 items as fast as possible (see Appendix A for instructions). After completing the measures, students were thanked for their participation and given contact information for any follow-up questions.

Participants from the psychology pool who completed the data online were provided a detailed description of the study including its purpose, anticipated length of time, potential consequences and benefits, contact information for the researcher, and a statement of consent. After signing up for the study, participants were provided a hyperlink to the survey and password. On the first page of the survey, participants were provided with a set of instructions indicating items should *not* be read or thoughtfully responded to; instead, responses should be given as fast as possible (see Appendix A for instructions).

To increase the sample size of the Known Rapid Responders sample, data were also collected from two additional sources using two different methods. It is important to note that despite using different methods to collect the additional data, both groups received the same set of instructions previously described (see Appendix A). The first source of additional data came from students attending a makeup testing session during the spring semester of 2016. Specifically, students rapidly completed the MFLS *after* they had completed the required sequence of assessments used for accountability purposes. The second source of additional data came from friends and acquaintances of the researcher. The researcher sent a bulk email including a description of the study, hyperlink, and password to friends and acquaintances. Records with missing data or with

responses greater than 15 seconds were removed from the data set. In total, the Known Rapid Responders sample consisted of 181 participants: 3 participants were recruited through the undergraduate psychology participant pool and completed the MFLS in person, 70 participants were recruited through the undergraduate psychology participant pool and completed the MFLS online, 51 participants completed the MFLS during the makeup testing 2016 sessions, and 57 participants were recruited via email.

Measures

Meaningful Life Scale (MFLS). The MFLS is a combination of four noncognitive measures that assess the construct meaningful life: Meaning in Life Questionnaire, Sources of Meaning and Meaning in Life Questionnaire, Work and Meaning Inventory-Revised, and Life Regard Index. Specifically, these four noncognitive measures were combined to create the 53-item MFLS. Thus, instead of administering the four measures individually, the four measures were administered together as a *set* and appeared to respondents to be one 53-item measure. The items on the MFLS were administered to the four samples in the same order as they are presented in Appendix B. Participants responded to all 53 items using a Likert rating scale ranging from 1 (*Absolutely untrue*) to 7 (*Absolutely true*); the midpoint of the scale, 4, is a neutral response (*Can't say true or untrue*). Higher scores reflect higher levels of the construct. Given the repetitious nature of the items, it was expected that the amount of noneffortful responding displayed by unmotivated students would increase as the test progressed. The four measures that were used to create the MFLS are described in further detail below. Coefficient alpha of the MFLS for the Primary sample in the current study was 0.958.

Meaning in Life Questionnaire (MLQ). The first meaningful life measure used to create the MFLS was the MLQ (Steger, Frazier, Oishi, & Kaler, 2006). This ten-item measure is comprised of two five-item subscales: Presence, which assesses "the presence of meaning or purpose in a person's life" (Steger et al., 2006, p. 83), and Search, which "measures the drive and orientation toward finding meaning in one's own life" (Steger et al., 2006, p. 85). There is one negatively worded item on the Presence subscale that needs to be reverse scored prior to scoring. Scores for the subscales can range from 5 to 35; higher scores indicate higher levels of the construct. Previous research has supported the scales factor structure and provided evidence of their distinctiveness (Steger et al., 2006).

Sources of Meaning and Meaning in Life Questionnaire (SoME). The SoME (Schnell, 2009) is a 151 item questionnaire assessing meaningfulness, crisis of meaning, and sources of meaning. For institutional accountability purposes, university staff have used a five-item subscale from the SoME measuring meaningfulness instead of using the entire measure. Thus, for the purpose of the current study, the five-item subscale from the SoME measuring meaningfulness was used. According to the scales' author, meaningfulness is defined as "a fundamental sense of meaning, based on an appraisal of one's life as coherent, significant, directed, and belonging" (Schnell, 2009, p. 487). Subscale scores can range from 5 to 35.

Work and Meaning Inventory-Revised (WAMI-R). The WAMI-R (Steger, Dik, & Duffy, 2012) is a 10-item multidimensional scale assessing how meaningful people find their work. Specifically, the WAMI-R contains three subscales: Positive Meaning, Meaning Making through Work, and Greater Good Motivation. The Positive Meaning subscale consists of four items and assesses the degree to which people positively

perceive the meaning of their work. The Meaning Making through Work subscale consists of three items and measures "the broader life context of people's work" by capturing how meaningful work enriches their lives. Finally, the Greater Good Motivation subscale consists of three items assessing the degree to which people find their work to meaningfully have an impact on others for the greater good. One item on the Greater Good Motivation scale is negatively worded and needs to be reverse scored prior to use. Scores on the Positive Meaning subscale can range from 4 to 28 whereas scores on the latter two subscales, Meaning Making through Work and Greater Good Motivation can range from 3 to 21.

Life Regard Index (LRI). The LRI (Battista & Almond, 1973) is a 28 item measure composed of two subscales: Framework and Fulfillment. The 14-item Framework subscale "measures the ability of an individual to see his life within some perspective or context" whereas the 14-item Fulfillment subscale "measures the degree to which an individual see himself as having fulfilled or as being in the process of fulfilling his framework or life-goals" (Battista & Almond, 1973, p. 411). Both subscales consist of seven negatively-worded items that need to be reverse-scored prior to scoring; scores for the subscales can range from 14 to 98.

Respondent and item characteristics used to gather external validity evidence. Data from various sources were collected and used to provide external validity evidence for the threshold calculation methods. The seven respondent characteristics and two item characteristics are described below.

Gender. The gender of students' in the Primary sample was obtained from university records. A dichotomous indicator was created where females were coded one.

Makeup testing session attendance status (walk-in). An indicator variable of how students who were required to attend the Makeup 2015 testing session actually completed the sequence of tests was obtained and used to differentiate students who *did not attend* the required make-up testing session and subsequently completed the tests at a walk-in computer lab (walk-in = 1) from those who *did attend* the required makeup testing session (walk-in = 0).

Student Opinion Scale (SOS). The SOS (Sundre & Moore, 2002) is a ten item noncognitive measure containing two five-item subscales, Effort and Importance, which assess the amount of effort students put forth in completing a series of tests as well as the perceived importance of the tests, respectively. An example of an item from the Effort subscale is, "I engaged in good effort throughout these tests." An example of an item from the Importance subscale is "These were important tests to me." The items were answered using a five-point Likert response scale ranging from 1 (*Strongly disagree*) to 5 (*Strongly agree*), thus scores on the subscale range from 5 to 25 with higher scores reflecting higher levels of effort and importance. Previous research has supported the dimensionality of the scale and has provided convergent validity evidence of the scale (e.g., Thelk et al., 2009).[5] For the purpose of the current study, only scores from the self-reported Effort subscale were used. It is important to note that the SOS assesses the amount of effort respondents put forth on *all* of the tests completed during a testing session and not just on one particular test. Thus, in contrast to the SB index and RTE which assess how much effort was put forth on the MFLS, the SOS Effort score is considered a global measure of effort because it refers to how much effort students put

forth on each test completed during the entire testing session. Coefficient alpha of effort in the current study was 0.812.

Academic ability. The SAT Critical Reading (SAT-CR) and SAT Math (SAT-M) scores of students within the Primary sample were obtained from university records and used as a respondent characteristic. SAT-CR and SAT-M scores were available for 452 students in the Primary sample.

Individual Consistency Index. The individual consistency index values were calculated by identifying semantic synonyms – pairs of items on the MFLS that were similar in meaning. Seven pairs of items were identified (see Table 3). Within-person correlations were calculated and the magnitude of the correlation was used as the index value.

Open-ended item length. The length of an open-ended response completed by students participating in the Makeup 2015 Testing sample was calculated by counting the total number of words in the response. SAS 9.4 was use to count the length of the responses.

Item position. The serial position of an item on the MFLS was determined and used as an item characteristic in the current study.

Item length. The length of an item on the MFLS was calculated using SAS 9.4 as the total number of words within an item.

Data Analysis

The current study was conducted in three phases. Phase One focused on determining whether various methods could be used to calculate time thresholds for items on a 53 item noncognitive measure, the MFLS. Data from the Primary sample (N = 568)

was used to calculate six of the eight threshold calculation methods (visual inspection, lognormal mixture modeling, NT10, NT20, NT30, and reading speed; see Table 2). To determine if the inclusion of an independent group of known rapid responders would affect the calculation of the thresholds, data from the Known Rapid Responders sample (N = 181) were combined with the Primary sample and used with the two remaining calculation methods (visual inspection with information and mixture modeling with information). Phase Two of the study calculated exact and approximate rates of agreement for the time thresholds and examined if the resulting time thresholds and the SB index values varied across methods. Finally, Phase Three of the study gathered validity evidence for the various time threshold calculation methods and examined if one method should be used over another, especially when considered in conjunction with results from Phase Two. Ultimately, the results of the three phases were used to determine (a) if various calculation methods could be used to define the time thresholds and if including additional information affected the calculation of the thresholds, (b) if the defined thresholds or solution behavior classification values differed across methods at the item level, and (c) if the external validity evidence indicated the results were meaningful and if the evidence supported the use of one method over the others. The three phases are described in further detail below.

Phase one: Defining the SB time thresholds. In order to address the first purpose of the study, a total of eight methods were used to calculate the solution behavior time thresholds: visually inspecting items' response time distributions, lognormal mixture modeling, NT10, NT20, NT30, reading speed, visual inspection with information, and lognormal mixture modeling with information. Thus, a total of eight methods were used

to calculate the time thresholds which would potentially yield 424 time thresholds (53 items X 8 methods) and subsequently 424 SB index values for each respondent. The time threshold calculation methods are described in detail below.

Visual inspection of response time distribution (INSPECT). To define the time thresholds using the visual inspection method, the response time distribution for each item was visually examined. The visual inspection method is based on the assumption an item's response time distribution will capture two types of responding behavior and will appear to be bimodal (as exhibited below in Figure 1). In other words, two modes (i.e., frequency spikes) should appear upon examination: one smaller mode should appear on the left end of the distribution reflecting rapid-responding behavior whereas a second larger mode should occur further along the continuum reflecting students exhibiting solution behavior. If an item's response time appeared bimodal, then the time threshold was defined as the time occurring at the upper end of the distribution with the shorter mode (Kong et al., 2007). Two raters independently reviewed the response time distribution for each item on the MFLS and visually identified the time thresholds. Agreement between the two raters was examined by calculating the difference between the time thresholds. If the difference between raters was less than two seconds, then the average of the two time thresholds (rounded to the nearest tenth of a second) was used. However, if the time threshold difference between the two raters was larger than two seconds, then the response time distribution for that item was reviewed and discussed. If the raters failed to reach agreement on a common time threshold for an item, then it was concluded that a time threshold could not be confidently set for that item and the time threshold value for the item was set equal to missing. Similarly, if an item's response

time distribution did not appear bimodal and the raters were unable to define a time threshold then the time threshold for that item was defined as missing.

Visual inspection of response time distribution with information (INSPECT2). The visual inspection method of setting time thresholds was conducted twice: once as previously described using the Primary sample and a second time using data from the Known Rapid Responders sample combined with the Primary sample. Specifically, the response time distribution for each item answered by the Primary sample was examined in conjunction with the response time distribution for each item answered by the Known Rapid Responders sample (see Figure 2). The same processes described earlier for the INSPECT condition were used to set the thresholds.

Lognormal mixture modeling (MIXTURE). Lognormal mixture modeling was the second primary method used to define the time thresholds. One- and two-class lognormal mixture models were fit to the untransformed response time distribution for each item and were estimated using maximum likelihood estimation via PROC FMM in SAS 9.4. As reviewed in Chapter Two, the two-class lognormal mixture model can be mathematically expressed as

$$F_{Oi} = \rho_i F_{Gi} + (1 - \rho_i) F_{Si}, \tag{7}$$

where F_{Oi} is the observed response time distribution for item *i*, F_{Gi} is the rapid-guessing response time distribution for item *i*, F_{Si} is the solution behavior response time distribution for item *i*, and ρ_i is the proportion of population respondents in the rapid-responding class on item *i* (Schnipke & Scrams, 1997). More specifically, the rapid-guessing and solution behavior response time distributions can be expressed as,

$$F_{Gi} = \frac{1}{\sqrt{t\sigma_{Gi}(2\pi)}} exp\left[\frac{-[ln\left(\frac{t}{m_{Gi}}\right)]^2}{2\sigma_{Gi}^2}\right], F_{Si} = \frac{1}{\sqrt{t\sigma_{Si}(2\pi)}} exp\left[\frac{-[ln\left(\frac{t}{m_{Si}}\right)]^2}{2\sigma_{Si}^2}\right], \tag{8}$$

where t is the response time for item i, m_{Gi} and m_{Si} are the scale parameters for the rapid-guessing behavior class and solution behavior class, respectively, and σ_{Gi} and σ_{Si} are the shape parameters for the rapid-responding behavior and solution behavior class, respectively (Schnipke & Scrams, 1997; Yang, 2007). A total of five parameters are freely estimated by the two-class model: ρ_i, the mixing proportion representing the number of examinees in the population's rapid-responding class; m_{Gi} and σ_{Gi}, the scale and shape of the natural log of the response time for the rapid-guessing class; and m_{Si}, and σ_{Si}, the scale and shape of the natural log of the response time for the solution behavior class. The mixing proportion reflects the number of examinees in the population's rapid-responding class. Mixing proportions are constrained to be positive and sum to one across classes. This constraint results in only one mixing proportion estimated for a two-class model because the mixing proportion for the solution behavior class is calculated by subtracting the mixing proportion of the rapid-guessing class from one. In contrast to the two-class lognormal mixture model, only two parameters are estimated for the one-class lognormal mixture model: m_i and σ_i, the scale and shape of the natural log of response time.

To ensure the solutions did not converge to a local maximum, the models were estimated using starting values for the scale and shape parameters, which were obtained by kernel smoothing (Kong et al., 2007). Because the scale and shape parameters from the lognormal mixture models reflects the natural log of an item's response time, a kernel density plot of the log-transformed response time distributions was estimated for each item. The starting values for the one-class lognormal mixture models were determined by (a) identifying one mode for each item and (b) averaging the modes across items. The

average mode was then used as the starting value for the scale (i.e., mean) parameter for every item and a value of 1 was used as the starting value for the shape parameter for every item. The starting values for the two-class lognormal mixture model were identified using the same process, except two modes representing the two classes were identified for every item and averaged across items. These values were then used as the starting values for the scale parameter for the first and second class in the two-class model; a value of 1 was used as the starting value for the shape parameter of both classes.

Model fit was evaluated using the Akaike Information Criterion (AIC; Akaike, 1987), the Bayesian Information Criterion (BIC; Schwartz, 1978), and the Sample Size Adjusted BIC (SSABIC; Sclove, 1987). Smaller values were indicative of better fit. Given previous research has shown the SSABIC identifies the correct number of classes more often than the other indices (e.g., Enders & Tofighi, 2008; Tofighi & Enders, 2007), it was given more weight than the other criteria.[6] If the two-class model fit the data better than the one-class model then the mixture densities for the two classes were used to determine the number of respondents expected at each value of response time within each class. The time threshold for the two-class model was defined by the author as the response time at which there were more respondents in the second class than in the first class (i.e., at the point in which the two mixture densities intersected; Schnipke & Scrams, 1997). SAS 9.4 was used to calculate this point of intersection. However, if the two-class model did not fit the data better than the one-class model, than the one-class model was championed and the item did not have a defined time threshold (i.e., the item's time threshold was set to missing). Similarly, if the model failed to converge to a

solution or the two distributions did not intersect, then the time threshold was set to missing.

Lognormal mixture modeling with information (MIXTURE2). The lognormal mixture modeling method was used twice: once as previously described using the Primary sample and a second time using data from the Known Rapid Responders sample combined with the Primary sample. Including the Known Rapid Responders sample may facilitate the estimation process and make the two classes appear more distinct. The same processes described above for the MIXTURE condition was used to estimate the one-class and two-class lognormal mixture models and set the thresholds.

Normative Threshold (NT10, NT20, NT30). The third primary method used to define the time thresholds was the NT method. Specifically, three different NT levels were calculated: 10% (NT10), 20% (NT20), and 30% (NT30). The time thresholds were calculated by taking a percentage of the average response time for each item. For example, the NT10 method defines an item's time threshold as the value equal to 10% of the item's average response time. Thus, if the average response time for an item was 25 seconds, then the NT10 time threshold would equal 2.5 seconds. By design, smaller NT percentage levels yield smaller time thresholds and subsequently classify more responses as solution behavior responses whereas larger NT percentage levels will yield larger time thresholds and classify less responses as solution behavior. As previously mentioned, the NT method will define a time threshold for every item, even if all of the responses on an item were made without effort. Although previous research has supported using NT10 (Lee & Jia, 2014; O. L. Liu et al., 2015; Rios et al., 2014; Wise & Ma, 2012; Wise, 2015), because this is the first time the NT method was used to define thresholds for

items on a noncognitive measure, and given response times on noncognitive items are usually faster than they are on cognitive items, multiple percentage levels larger in magnitude to those previously used with cognitive tests were chosen to be used and compared.

Reading speed (RSPEED). The final method that was used to define time thresholds was based on reading speed. Zhang and Conrad (2013) used a measure of reading speed based on a speed that is considered typical for college students (200 milliseconds per word; Carver, 1992) to identify rapid responders completing a web-based survey. Specifically, the authors used a slightly slower reading speed of 300 milliseconds per word to calculate the time threshold for items by multiplying the total number of words in an item by 300 milliseconds. This method was replicated and used in the current study. The total number of words each for item on the MFLS was calculated using SAS 9.4.

Phase two: Comparing the resulting thresholds. In order to address the second purpose of the study, it was considered just as important to determine which calculation methods could not be used to define the time thresholds as it was to determine which calculation methods could be used. Thus, the proportion of items for which a time threshold could not be defined was calculated for each method. In particular, it was expected that some time thresholds would be missing for the visual inspection, visual inspection with information, lognormal mixture modeling, and lognormal mixture modeling with information methods. In addition to examining the degree to which methods failed to yield defined time thresholds, defined time thresholds were examined individually for each method. The exact rate of time threshold agreement across methods

was calculated as the percentage of items whose time thresholds were in exact agreement. In addition, similar to Kong et al. (2007), an approximate rate of time threshold agreement was also calculated as the percentage of items that were in agreement within two seconds of one another.

To ascertain if there were significant differences across threshold calculation methods in the proportion of respondents classified as engaging in solution behavior, a generalized estimating equation (GEE) was estimated for each item. GEEs were used because they can accommodate (a) the dichotomous nature of the dependent variable (i.e., the SB index) through a logit link function and (b) the within subjects nature of the independent variable (i.e., the calculation method). That is, given multiple threshold calculation methods were used for every respondent, a model was needed that could account for the within subjects correlation introduced by the repeated measures. Conceptually, although the data are nested (i.e., repeated measures nested within respondents), GEEs are considered single-level models because the effect of the repeated measures is not explicitly modeled, but rather treated as a nuisance (Burton, Gurrin, & Sly, 1998; McNeish & Stapleton, in press; Snijders & Bosker, 2012). GEEs only specify and estimate the fixed effects of the regression model. Using a logit link function, the model was specified as,

$$\log\frac{P(SB=1)}{P(SB=0)} = \beta_0 + \beta_1 INSPECT + \beta_2 INSPECT2 + \beta_3 MIXTURE + \beta_4 MIXTURE2 + \beta_5 NT10 + \beta_6 NT20 + \beta_7 NT30, \quad (9)$$

where the log odds of a respondent exhibiting solution behavior on an item was predicted by the threshold calculation method. Because threshold calculation method is categorical, seven dummy-coded variables were used to represent the threshold calculation method in

the model and RSPEED was used as the reference category. Of particular interest were the results of the omnibus test of the calculation method effect, which tests whether there are significant differences among the threshold calculation methods (i.e., H_0: all $\beta_k = 0$). If the omnibus test was significant, the coefficients were used in pairwise comparisons (e.g., H_0: $\beta_7 = 0$ compares NT30 to RSPEED and H_0: $\beta_6 - \beta_7 = 0$ compares NT20 to NT30) with coefficient differences greater than 0.05 considered practically significant.

When applying GEEs, a working correlation matrix needs to be specified by the researcher to account for the within subject correlations (Ballinger, 2004). A correlation matrix with a compound symmetric form was specified in the current analysis because there was not a logical ordering to the calculation methods (Ballinger, 2004). Robust standard errors for the regressions coefficients which take into account the within subject correlations were calculated using a sandwich estimator (Burton et al., 1998). A conservative criterion of $\alpha = .01$ was used given the model was estimated for every item. The data were analyzed in SAS 9.4 using PROC GLIMMIX with an empirical estimator (SAS Institute, 2015).

Phase three: Examining external validity evidence. To address the third purpose of the study, external validity evidence for the threshold calculation methods was gathered by calculating eight RTE and RTF scores based on the eight threshold calculation methods and examining: (a) if the RTE and RTF scores were related to external variables in theoretically expected ways and (b) if these relationships were dependent on the threshold calculation method used to create the scores. Models in Phase 3 used either RTE or RTF scores as the dependent variable along with three predictors: the external variable, threshold calculation method, and the interaction between the

external variable and threshold calculation method. Separate models were estimated for each external variable. In the sections below, the models using RTE as the dependent variable are described first and are followed by the models using RTF as the dependent variable.

RTE and respondent characteristics. A series of GEEs were used to examine the relationship between RTE and the respondent characteristics and to determine if the relationship was dependent on threshold calculation method. GEEs were used because they can accommodate the skewed proportional nature of RTE through a logit link function. That is, RTE scores are skewed proportions, bounded between the values of zero and one with many values near one. The logit-link function transforms the predicted values of RTE, so rather than predicting the RTE score for a respondent, the logit of RTE is predicted. Another reason GEEs were used for these analyses was because they can accommodate the within subject correlation introduced by the repeated measures nature of the RTE scores (Snijders & Bosker, 2012). To clarify the nature of the data, a screen shot of the data set analyzed is presented in Figure 3. Using a logit-link function, the GEE representing the logit of RTE was specified as

$$\log\left(\frac{RTE_i}{1-RTE_i}\right) = \beta_0 + \beta_1(INSPECT_i) + \beta_2(INSPECT2_i) + \beta_3(MIXTURE_i) +$$
$$\beta_4(MIXTURE2_i) + \beta_5(NT10_i) + \beta_6(NT20_i) + \beta_7(NT30_i) +$$
$$\beta_8(Characteristic) + \beta_9(Characteristic)(INSPECT_i) +$$
$$\beta_{10}(Characteristic)(INSPECT2_i) +$$
$$\beta_{11}(Characteristic\ )(MIXTURE_i) +$$
$$\beta_{12}(Characteristic\ )(MIXTURE2_i) + \beta_{13}(Characteristic)(NT10_i) +$$
$$\beta_{14}(Characteristic)(NT20_i) + \beta_{15}(Characteristic)(NT30_i) \qquad (10)$$

where the threshold calculation method main effect was captured by β_1 through β_7, the main effect for the respondent characteristic was captured by β_8, and the interaction between the respondent characteristic and threshold calculation method was captured by the remaining coefficients, β_9 through β_{15}. Because the threshold calculation method is categorical, seven dummy-coded variables were used to represent the threshold calculation method in the model and RSPEED was used as a reference variable. To evaluate if the relationship between the logit of RTE and the respondent characteristic was dependent on threshold calculation method, the significance of an omnibus test evaluating the interactions between the threshold calculation method and respondent characteristic was evaluated (i.e., H_0: $\beta_k = 0$ for $k = 9$ to 15). If the interaction was significant, pairwise comparisons were conducted to identify which calculation methods differed from one another (e.g., H_0: $\beta_{15} = 0$ compares the relationship calculated using NT30 to the relationship calculated using RSPEED and H_0: $\beta_{15} - \beta_{14} = 0$ compares the relationship calculated using NT30 to the relationship calculated using NT20).

To help interpret the results and assess practical significance of the relationships across calculation methods, two different graphs were created and considered in conjunction with the results of the pairwise comparisons. Specifically, one graph reflected the model-implied relationships between the respondent characteristic and the *logit of RTE* and another graph reflected the model-implied relationships between the respondent characteristic and *predicted RTE*. This model was estimated seven times – once for each of the following respondent characteristics: gender, walk-in status, effort, SAT-CR, SAT-M, individual consistency index, and length of an open-ended response question. A correlation matrix with a compound symmetric form was specified because

there was not a logical ordering to the calculation methods (Ballinger, 2004). Robust standard errors for the regressions coefficients which take into account the within subject correlations were calculated using a sandwich estimator (Burton et al., 1998). A conservative criterion of α = .01 was used. The data were analyzed in SAS 9.4 using PROC GLIMMIX with an empirical estimator (SAS Institute, 2015).

RTF and item characteristics. Unlike RTE scores, which are calculated for every respondent across items, RTF scores are calculated for every item, across people. Thus, in order to examine the relationship between RTF scores and two item characteristics, the data set used to analyze the RTE scores was restructured. Specifically, the data set used to examine the relationship between RTE scores and respondent characteristics was structured so each respondent had eight records (one for each calculated RTE score). In order to examine the relationship between RTF scores and item characteristics, the data set was structured so each item had eight records (one for each calculated RTF score). To aid in clarification, a screen shot of the data set is provided in Figure 4.

The relationship between the RTF scores and two item characteristics was analyzed using the same GEE model conducted with the RTE scores. Specifically, two GEEs with a logit link function were used to examine the relationship between (a) RTF and serial item position and (b) RTF and item length. Specifically, the GEE model representing the logit of RTF was specified as

$$\log\left(\frac{RTF_i}{1-RTF_i}\right) = \beta_0 + \beta_1(INSPECT_i) + \beta_2(INSPECT2_i) + \beta_3(MIXTURE_i) +$$
$$\beta_4(MIXTURE2_i) + \beta_5(NT10_i) + \beta_6(NT20_i) + \beta_7(NT30_i) +$$
$$\beta_8(Characteristic) + \beta_9(Characteristic)(INSPECT_i) +$$
$$\beta_{10}(Characteristic)(INSPECT2_i) +$$

$$\beta_{11}(Characteristic\ )(MIXTURE_i) +$$

$$\beta_{12}(Characteristic\ )(MIXTURE2_i) + \beta_{13}(Characteristic)(NT10_i) +$$

$$\beta_{14}(Characteristic)(NT20_i) + \beta_{15}(Characteristic)(NT30_i) \qquad (10)$$

where the threshold calculation method main effect was captured by β_1 through β_7, the main effect for the respondent characteristic was captured by β_8, and the interaction between the item characteristic and threshold calculation method was captured by the remaining coefficients, β_9 through β_{15}. Because the threshold calculation method is categorical, seven dummy-coded variables were used to represent the threshold calculation method in the model and RSPEED was used as a reference variable. To evaluate if the relationship between RTF and the respondent characteristic was dependent on threshold calculation method, the significance of an omnibus test evaluating the interactions between the threshold calculation method and respondent characteristic was evaluated (i.e., H_0: $\beta_k = 0$ for $k = 9$ to 15). If the interaction was significant pairwise comparisons were conducted to identify which calculation methods differed from one another (e.g., H_0: $\beta_{15} = 0$ compares the relationship calculated using NT30 to the relationship calculated using RSPEED and H_0: $\beta_{15} - \beta_{14} = 0$ compares the relationship calculated using NT30 to the relationship calculated using NT20).

To help interpret the results and assess the practical significance of the relationships across calculation methods, two different graphs were created and considered in conjunction with the results of the pairwise comparisons. Specifically, one graph reflected the model-implied relationships between the item characteristic and the logit of RTF and another graph reflected the model-implied relationships between the item characteristic and predicted RTF. A correlation matrix with a compound symmetric

form was specified because there was not a logical ordering to the threshold calculation methods (Ballinger, 2009). Robust standard errors for the regressions coefficients that take into account the within subject correlations were calculated using a sandwich estimator (Burton et al., 1998). The item characteristics were grand-mean centered to aid in interpretation. Robust standard errors for the regressions coefficients that take into account the within subject correlations were calculated using a sandwich estimator (Burton et al., 1998). A conservative criterion of $\alpha = .01$ was used. The data were analyzed in SAS 9.4 using PROC GLIMMIX with an empirical estimator (SAS Institute, 2015).

CHAPTER FOUR

Results

Data Cleaning

Primary sample. The MFLS was completed by 336 students during the Makeup Testing 2015 session, 77 students during Assessment Day 2016, and 158 students during the Makeup Testing 2016 session. Of the 570 students in the Primary sample, two respondents with extreme response times (i.e., response times to items were greater than two minutes) and one respondent with missing data on the SOS measure were deleted. The final sample size of the Primary sample was $N = 568$. Descriptive statistics of the Primary sample's response times are presented below in Table 4. The items' response time distributions were positively skewed. Inspection of the median response times indicates that the majority of the items were answered in about 3 seconds, with response times to items ranging from 1.6 to 7.0 seconds. Demographic information about respondents in the Primary sample is presented in Table 5. The average age of respondents was 20.62 ($SD = 1.77$).

Known Rapid Responders sample. The MFLS was completed by 246 participants under instructions to rapidly respond to the items as fast as possible. Of those that completed the MFLS under the rapid-response instructions, 39 records with missing data and 26 records with item response times 15 seconds or longer were removed.[7] The final sample size of the Known Rapid Responders sample was $N = 181$. Descriptive statistics of the Known Rapid Responders' response times to the MFLS items are presented below in Table 6. Inspection of the median response times indicates that the

majority of the items were answered in about 1.4 seconds, with response times to items ranging from 0.9 to 2.6 seconds.

To examine the impact including a known group of rapid responders had on the calculation of thresholds using the visual inspection with information and mixture modeling with information methods, the Known Rapid Responders sample was combined with the Primary sample, resulting in a combined sample size of $N = 749$. Descriptive statistics of the combined sample response times to items on the MFLS are presented below in Table 7. Inspection of the median response times of the combined sample revealed the majority of items were answered in 4.08 seconds with response times ranging from 2.6 to 7.1 seconds.

Phase One: Calculating Time Thresholds

The purpose of Phase One was to determine if time thresholds for items on the MFLS could be calculated using eight threshold calculation methods: INSPECT, INSPECT2, MIXTURE, MIXTURE2, NT10, NT20, NT30, and RSPEED. The resulting time thresholds calculated by the eight methods are presented in Table 8 and the descriptive statistics of the time thresholds are presented in Table 9. The results for each calculation method are individually discussed below.

Visual inspection (INSPECT). The response time distributions for each MFLS item completed by the Primary sample was graphed and visually inspected by two raters. Contrary to the hypothesis that some of the items response time distributions would not appear bimodal, all of the items appeared to have a bimodal response time distribution, thus indicating a time threshold could be defined for every item. Two raters independently defined the time threshold for each item and the difference between the

raters' thresholds were calculated. All of the differences between the time thresholds defined by raters were less than two seconds, thus, the time threshold was calculated as the average of the two rater-defined thresholds. Overall, the average time threshold for items on the MFLS using the visual inspection method was 2.08 seconds (*SD* = 0.33) and ranged from 1.30 to 2.90 seconds.

Visual inspection with information (INSPECT2). The response time distributions for each item completed by the Primary and Known Rapid Responders samples were graphed and visually inspected by two raters (see Figure 2 for an example). The response time distributions for every item appeared bimodal, thus enabling the raters to define a time threshold for every item. The difference between the rater defined time thresholds were less than two seconds across all items. Thus, the visual inspection with information time thresholds were calculated as the average of the two rater-defined thresholds. The average time threshold defined for items on the MFLS using the visual inspection with information method was 2.34 seconds (*SD* = 0.32) and ranged from 1.70 to 3.00 seconds.

Lognormal mixture modeling (MIXTURE). The starting value used for the scale parameter of the one-class model was .3 and the starting values used for the scale parameters for the two-class model were .3 and 1.5 for the rapid-responder and solution behavior classes, respectively. After determining the starting values, a series of one- and two-class lognormal mixture models were fit to the Primary sample data for every item. All of the models converged to a solution. Model fit indices for both the one-class and two-class models for each item are presented in Table 9. The log-likelihood, AIC and SSABIC values indicated that the two-class model fit better than the one-class model

across all 53 items. However, the BIC index indicated the two-class model did not fit better than the one-class model for items 8 and 15 (see Table 10). Given the SSABIC index indicated otherwise, the two-class model was championed for all 53 items.

A time threshold for item 13 was not calculated because the two mixture distributions did not intersect. Specifically, one of the distributions was completely subsumed within the other distribution. Examination of item 13's mixing proportions for the two-class solution revealed only .7% of the respondents were classified in the rapidly responding class.[8] The average time threshold calculated for the MFLS items using lognormal mixture models was 2.29 seconds (*SD* = 0.50) and ranged from 1.45 to 3.75 seconds.

Lognormal mixture modeling with information (MIXTURE2). The starting value used for the scale parameter of the one-class model was 1.3 and the starting values for the scale parameters for the two-class model were .45 and 1.6 for the rapid-responder and solution behavior classes, respectively. One- and two-class lognormal mixture models with information were fit to the combined Primary and Known Rapid Responders samples. All of the models converged to a solution. Model fit indices for both the one- and two-class models for each item on the MFLS are presented in Table 11. Examination of the log-likelihood, AIC and SSABIC values indicated the two-class model fit better than the one-class model across all 53 items. In contrast, the BIC index indicated the two-class model did not fit better than the one-class model for item 1. However, given the SSABIC model indicated otherwise, the two-class model was championed for all 53 items.

Time thresholds for items 1, 8, and 14 were not calculated because the two mixture distributions did not intersect. Specifically, for each item, one of the mixture distributions was completely subsumed within the other distribution. For example, as shown in Figure 5 the distribution of Class Two for item 8 is subsumed within the distribution of Class One. The average time threshold defined by the MIXTURE2 method was 3.19 seconds (*SD* = 0.49); the thresholds ranged from 1.65 to 4.60 seconds (see Table 8).

NT10. The average time threshold calculated by NT10 was 0.51 seconds (*SD* = .11) and ranged from .33 to 1 seconds.

NT20. The average NT20 time threshold calculated across the MFLS items was 1.02 seconds (*SD* = .22) and ranged from .67 to 2 seconds.

NT30. The average NT30 time threshold calculated across the MFLS items was 1.54 seconds (*SD* = .33) and ranged from 1 to 3 seconds.

Reading speed (RSPEED). The average time threshold calculated using the reading speed calculation method was 3.34 seconds (*SD* = 1.18) and ranged from 1.20 to 6.30 seconds.

Phase Two: Comparing the Time Thresholds

As shown in Table 8, the majority of the threshold calculation methods were able to define time thresholds for items on a noncognitive measure administered in a low-stakes setting. Out of the 424 time thresholds that could have been defined (8 threshold calculation methods X 53 items), only 4 time thresholds were not defined. Specifically, the MIXTURE and MIXTURE2 methods were unable to define time thresholds for four items because the estimated mixture distributions did not intersect (i.e., one distribution

was completely subsumed within another distribution). Overall, the time thresholds defined by the threshold calculation methods were small and ranged from .51 seconds to 3.34 seconds, on average (see Table 9). The variability of the time thresholds across calculation methods was also small, ranging from *SD* = .11 to *SD* = 1.18 seconds. As shown in Table 9, the time thresholds defined by RSPEED displayed the most variability. Intuitively, this makes sense given the time thresholds defined by RSPEED were based on the predicted reading speed of an item (which was calculated based on the length of an item) whereas the time thresholds defined by the other calculation methods were based on the actual response times.

The average time thresholds appeared in the following rank order (from smallest to largest): NT10, NT20, NT30, INSPECT, MIXTURE, INSPECT2, MIXTURE2, RSPEED (see Table 9). This rank-ordered pattern is readily seen in a graph of the time thresholds displayed in Figure 6. Specifically, in comparison to the other threshold calculation methods, on average the NT10, NT20, and NT30 methods yielded the smallest or most conservative time thresholds across items (i.e., guarded against classifying a response as a rapid response). The time thresholds defined by the INSPECT, INSPECT2, MIXTURE, and MIXTURE2 calculation methods, which will be referred to as the "distributional" methods, were slightly larger in magnitude whereby the time thresholds calculated using the Known Group of Rapid Responders, INSPECT2 and MIXTURE2, yielded large time thresholds on average than the time thresholds calculated without the rapid responders, INSPECT and MIXTURE, respectively. In comparison to the other three distributional methods, MIXTURE2 yielded the largest time thresholds,

on average. Finally, the time thresholds defined by RSPEED were the largest or most liberal time thresholds across the eight methods and the most variable.

Time threshold agreement rates. The time threshold agreement rates between items using the eight threshold calculation methods (INSPECT, INSPECT2, MIXTURE, MIXTURE2, NT10, NT20, NT30, and RSPEED) are presented in Table 12. The proportion of time thresholds that were in exact agreement are presented below the diagonal and the proportion of time thresholds that were within two seconds of agreement are presented above the diagonal. Overall, the majority of time thresholds were not in exact agreement; the rates of exact agreement between time thresholds across calculation method ranged from 0% to 17%. The largest proportion of time thresholds that were in exact agreement (17%) was displayed between the time thresholds defined using the INSPECT2 and MIXTURE methods, which was observed in Figure 6.

Larger proportions of agreement were observed between threshold calculation methods when the approximate rates of agreement were examined. For example, time thresholds defined by the three NT methods displayed a perfect rate of approximate agreement (1.00; see Table 12). Similarly, time thresholds defined by the distributional methods (i.e., INSPECT, INSPECT2, MIXTURE, and MIXTURE2) displayed a high rate of approximate agreement with each other, ranging from 0.98 to 1.00. Agreement between the NT methods and the distributional methods was high for NT20, NT30, INSPECT, INSPECT2, and MIXTURE. In contrast, the lowest rates of approximate agreement were observed between the time thresholds calculated with MIXTURE2 and the time thresholds calculated using NT10 and NT20 (0.06 and 0.20, respectively). These lower rates of agreement displayed between the time thresholds defined by MIXTURE2

with the thresholds defined by NT10 and NT20 were unexpected given the time thresholds defined by RSPEED were higher on average than the time thresholds defined by MIXTURE2. However, the time thresholds defined by RSPEED had more variability than did the time thresholds defined by MIXTURE2. Indeed, RSPEED had relatively lower agreement with the NT methods (ranging from .30 to .58) than with the distributional methods (ranging from 0.85-0.89). Based on these results and given the narrow range of the average calculated time thresholds across methods (.51 seconds to 3.34 seconds, on average; see Table 8), there is evidence to suggest using two seconds as a constant to examine the approximate rate of agreement among time thresholds may have been too large of a value to detect meaningful differences between calculation methods.

SB classification indices. The proportion of respondents classified as exhibiting solution behavior on the MFLS items calculated using the different threshold calculation methods are presented in Table 13 and are plotted in Figure 7.[9] Overall, a large proportion of respondents were classified as exhibiting solution behavior on items due to the low time thresholds defined by the calculation methods. Examination of the proportions of classified respondents in Table 13 revealed the magnitude of the proportions followed the reverse of the rank-ordered pattern observed earlier with the thresholds (see Figure 7). For example, NT10, which yielded the lowest or most conservative time thresholds on average, classified the largest proportion of respondents as exhibiting solution behavior across items. Specifically, 100% of respondents on 33 items were classified by NT10 as exhibiting solution behavior; the smallest proportion of respondents classified by NT10 as exhibiting solution behavior on an item was 99.3%

(item 41). In contrast, RSPEED, which yielded the largest or most liberal time threshold on average across methods, classified substantially fewer respondents as exhibiting solution behavior on the items. For example, 80% of respondents or less were classified as exhibiting solution behavior on 29 of the 53 items. With the exception of RSPEED and MIXTURE2, the majority of the threshold calculation methods classified at least 90% of the respondents as exhibiting solution behavior across items (see Figure 7).

Generalized estimating equations (GEEs) were used to assess if the log-odds of the proportion of respondents classified as exhibiting solution behavior on an item differed across threshold calculation methods. Of the 53 GEEs that were estimated (one for each item), only 16 models converged to an admissible solution (items 2, 5, 7, 17, 19, 21, 28, 29, 30, 32, 39, 40, 41, 42, 50, 51). To help diagnose convergence problems, descriptive statistics and correlations of the resulting classification indices were calculated and examined for each item, which revealed two problems. One of the reasons the models may have failed to converge was due to the lack of variability displayed by the SB indices calculated using NT10, and to a lesser extent NT20 (Hox, 2010). The second reason the models may have failed to converge to a solution was due to multicollinearity. Specifically, some of the SB classification indices exhibited a perfect correlation across calculation methods for several items. For example, the SB classification index calculated for item 1 by INSPECT2 exhibited a correlation of r = 1.00 with the SB classification index calculated by MIXTURE. This relationship between the INSPECT2 and MIXTURE SB classification indices occurred for 9 items (items 1, 12, 18, 22, 42, 46, 48, 50, and 51). Similar relationships between the SB classification

indices calculated using INSPECT and INSPECT2, INSPECT and MIXTURE, NT10 and NT20, and INSPECT and NT30 also occurred.

To address these issues, a series of modified GEEs with one or more the previously described threshold calculation methods excluded from the model were estimated. For example, the following GEE predicting the log-odds of a respondent exhibiting solution behavior was modified from the original equation and used for item 3,

$$\log\frac{P(SB=1)}{P(SB=0)} = \beta_0 + \beta_1 INSPECT + \beta_2 INSPECT2 + \beta_3 MIXTURE + \beta_4 MIXTURE2 + \beta_6 NT20 + \beta_7 NT30, \quad (12)$$

where the coefficient modeling the effect of NT10, β_5, was excluded from the model. All of the modified GEEs successfully converged to a solution. The resulting omnibus tests are presented in Table 14.

The results of the analyses indicated the threshold calculation methods differentially classified respondents as exhibiting solution behavior on all of the items except for item 1. Thus, pairwise comparisons were conducted for the latter 52 items. The total number of pairwise comparisons that were conducted was dependent on the total number of threshold calculation methods retained in the model; a maximum of 28 pairwise comparisons could be conducted for each of the 52 items. Due to the large number of tests conducted, the results presented in Table 15 are summarized according to the calculation methods compared by using a series of dichotomous indicators to indicate if a comparison was statistically significant (= 1) or not (= 0) and superscript letters (e.g., [a]) to indicate which calculation method(s) were excluded from the model for an item. A similar method was used to present the practical significance results in Table 16.

Overall, a large majority of the pairwise comparisons that were not conducted were because the NT10 threshold calculation method was excluded from the model. Specifically, the NT10 threshold calculation method was excluded from 36 of the GEEs, NT20 was excluded from six, MIXTURE2 was excluded from two, and NT30 and MIXTURE were both excluded from one. Models that excluded the MIXTURE and MIXTURE2 threshold calculation methods were for items where the two calculation methods failed to define a time threshold. Statistical significance was evaluated using a criterion of $\alpha = .01$. Given the results were reported in log-odds, to assess practical significance, differences between the *proportion* of respondents classified on an item were examined. Proportional differences that were 0.05 or greater were considered practically significant. A summary of the pairwise comparison results are presented in the Table 17. Because Table 17 summarizes a great deal of information and might be confusing at first glance, the results in the first row are described here. The first row summarizes comparisons between the proportion of respondents classified as engaging in solution behavior using the INSPECT and INSPECT2 methods. These comparisons were conducted for 52 of the 53 items, with 27 of those 52 comparisons (52%) being statistically significant and 4 of the 52 comparisons (8%) being both statistically and practically significant.

Several of the pairwise comparisons indicated statistically significant differences across calculation methods in the number of respondents classified as exhibiting solution behavior on an item. However, a substantially smaller amount of these comparisons indicated the differences were statistically *and* practically significant. For example, 94% of the pairwise comparisons conducted across items indicated the log-odds of

respondents being classified using INSPECT were statistically different from the log-odds of respondents being classified using NT10; however, only 31% of these comparisons indicated the differences were also practically different. Differences in the proportion of respondents classified as exhibiting solution behavior across methods that were considered both statistically *and* practically significant tended to occur when threshold calculation methods at either end of the rank-ordered spectrum were compared (e.g., comparing respondents classified by RSPEED to those classified by NT10). Intuitively this finding makes sense given the narrow range displayed by the time thresholds. Recall, the average time thresholds calculated by the eight methods ranged from 0.51 seconds to 3.34 seconds. The small range for the average time thresholds indicates differences between the proportions of classified respondents will also be small across calculation methods.

Examination of the results comparing the SB classifications revealed a similar pattern previously displayed by the time threshold comparisons. Specifically, across the majority of the pairwise comparisons conducted, the proportion of respondents classified using the NT methods did not significantly or practically differ from one another. That is, none of the pairwise comparisons indicated the proportion of respondents classified as exhibiting solution behavior by NT10 differed from those classified by NT20, and none of the proportions classified by NT20 differed from those classified by NT30. Only one comparison indicated the proportion of respondents classified by NT10 significantly and practically differed from the proportion classified by NT30.

The results also indicated there were significant and practical differences between the four distributional calculation methods (INSPECT, INSPECT2, MIXTURE, and

MIXTURE2). Specifically, the proportion of respondents classified using the first three distributional methods (i.e., INSPECT, INSPECT2, and MIXTURE) were similar in magnitude across items and only displayed statistical and practical differences on 8% to 16% of the comparisons conducted. In contrast, the proportion of respondents classified by the same three distributional methods statistically and practically differed from the proportion of respondents classified using MIXTURE2 on 94% to 96% of the pairwise comparisons conducted (see Table 17).

Interestingly, the inclusion of the Known Rapid Responders sample had a larger statistical and practical effect on the proportion of respondents classified using MIXTURE2 than it did on the proportion of respondents classified using INSPECT2. That is, 94% of the pairwise comparisons conducted indicated the proportion of respondents classified using MIXTURE2 were significantly and practically different from the proportion of respondents classified using MIXTURE. In contrast, only 16% of the pairwise comparisons conducted indicated the proportion of respondents classified using INSPECT2 significantly and practically differed from those classified using INSPECT.

Given the NT methods did not differ from one another and the distributional methods (with the exception of MIXTURE2) did not differ from one another, it was of interest to know whether the NT methods differed from the distributional methods in their classification of respondents. NT30 was fairly similar to INSPECT, INSPECT2, and MIXTURE in classifications (only 4% to 27% of pairwise comparisons were statistically and practically significant). More differences were found between NT20 when compared to INSPECT, INSPECT2, and MIXTURE (26% to 54% of pairwise comparisons were

statistically and practically significant). The few tests that could be conducted to compare NT10 to INSPECT, INSPECT2, and MIXTURE, indicated even more differences between NT10 and the three distributional methods (31% to 63% of pairwise comparisons were statistically and practically significant). Just as MIXTURE2 was different in classifying respondents when compared to the other distributional methods, almost all tests (96% to 100%) indicated statistically and practically significant differences between MIXTURE2 and the NT threshold calculation methods.

Similar to MIXTURE2, the proportion of respondents classified as exhibiting solution behavior across items using RSPEED also significantly and practically differed from the proportion of respondents classified using the other threshold calculation methods. For example, the proportion of respondents classified by RSPEED significantly and practically differed from the proportion of respondents classified using NT10, NT20, and NT30 on 100%, 93%, and 88% of the comparisons conducted, respectively. The majority of tests (76-82%) also revealed significant differences in classification between RSPEED and the distributional methods. Overall, in comparison to the other threshold calculation methods, MIXTURE2 and RSPEED classified significantly and practically smaller proportions of respondents as exhibiting solution behavior across items then did the other threshold calculation methods. Thus, in contrast to the NT10 calculation method which was very liberal in classifying respondents as exhibiting solution behavior across items, RSPEED and MIXTURE2 were very conservative classifying respondents as exhibiting solution behavior.

In summary, the results from Phase Two indicated time thresholds can be calculated for items on a noncognitive measure using a variety of threshold calculation

methods. As anticipated, the time thresholds defined for the noncognitive items were smaller in comparison to time thresholds calculated for cognitive items reported in the literature. Consequently, there was less variability among the time thresholds and subsequent proportions of respondents classified as exhibiting solution behavior on items across threshold calculation methods. However, some differences among the threshold calculation methods did emerge.

Across the eight threshold calculation methods, NT10, NT20, and NT30 yielded the smallest time thresholds, on average, which in turn classified the largest proportions of respondents exhibiting solution behavior. Out of the three NT methods, NT10 consistently classified at least 99% of the respondents across items as exhibiting solution behavior. The lack of variability displayed by the classification indices calculated using NT10, and to some extent using NT20, created problems when differences between the classification indices across methods were examined. In contrast to the NT methods, three of the distributional methods (INSPECT, INSPECT2, MIXTURE) defined time thresholds similar in magnitude that were slightly larger than the time thresholds defined by the NT methods and therefore classified a smaller proportion of respondents as exhibiting solution behavior. The fourth distributional method, MIXTURE2, defined the largest time threshold across the four distributional methods, on average, and the second largest time threshold across all eight calculation methods, on average. Finally, RSPEED calculated the largest time thresholds compared to other methods, on average, which subsequently classified the smallest proportion of respondents as exhibiting solution behavior across items. In addition, the time thresholds defined by RSPEED also displayed a lot more variability than the time thresholds defined by the other calculation

methods, as seen in Figure 6. In conclusion, the results of Phase Two indicated that although the SB index can be used with low-stakes noncognitive measures, the resulting classifications will differ depending on which threshold calculation methods are used to define the time thresholds.

Phase Three: External Validity Evidence

RTE and respondent characteristics. RTE scores for each of the threshold calculation methods were calculated using Equation 2 specified in Chapter 2. Recall, RTE scores reflect the proportion of items on the MFLS on which respondents exhibited solution behavior. Thus, respondents with higher RTE scores put forth more effort on average on the MFLS than respondents with lower RTE scores. Descriptive statistics, reliability estimates, and correlations of the eight RTE scores calculated using each of the different threshold calculation methods are presented in Table 18. As expected, the distributions of the RTE scores were negatively skewed. In addition, the mean RTE scores followed the same rank-ordered pattern previously displayed in Phase Two, whereby $RTE_{NT10,}$ RTE_{NT20}, and RTE_{NT30} exhibited the highest mean scores (.97 to 1.00), three of the distributional methods (i.e., $RTE_{INSPECT}$, $RTE_{INSPECT2}$, and $RTE_{MIXTURE}$) displayed similar yet slightly smaller means (.94 to .95), and $RTE_{MIXTURE2}$ and RTE_{RSPEED} yielded the lowest mean scores (.78 to .81). With the exception of the coefficient alpha estimate for NT10, all of the internal consistency estimates were greater than .91 (see Table 18). Coefficient alpha for the RTE scores calculated using NT10 was .447, which can be explained by the lack of variability displayed by the SB indices. The magnitude of the correlations between RTE scores ranged from r = .14 to r = .99, which indicated some

of the RTE scores were distinctly different from one another while others were not very distinct at all.

Generalized estimating equations (GEEs) were used to examine the relationship between the respondent characteristics and the logit of RTE and to determine if the relationship was dependent on the threshold calculation method. The continuous respondent characteristics were grand-mean centered to aid in interpretation. The results of the omnibus tests assessing the significance of the interaction between each respondent characteristic and calculation method are presented in Table 19. The results of each analysis are individually reviewed below.

Gender. Descriptive statistics of the RTE scores by gender are presented in Table 20 and point-biserial correlations between RTE scores and gender are presented in Table 21. The RTE scores for males and females were comparable in magnitude; the largest difference between genders displayed across threshold calculation methods was 0.03. With the exception of one correlation, all of the correlations between gender and RTE were positive, thus indicating females had higher RTE scores than males. Although the direction of these correlations was in the hypothesized direction, their magnitude was small – the only statistically significant correlation was with RTE_{NT30} (r = .11).

The results of the omnibus test assessing the interaction between gender and threshold calculation method indicated the interaction was not statistically significant (see Table 19); thus, the relationship between gender and the logit of RTE was not dependent on threshold calculation method. Given the nonsignificant interaction, the relationship between gender and the logit of RTE controlling for threshold calculation method was assessed by dropping the nonsignificant interaction and re-estimating the model. After

controlling for threshold calculation method, gender was not significantly related to the logit of RTE, $F(1,566) = 0.00$, $p = 0.963$. Although these results did not support the initial hypothesis that RTE would be significantly related to gender, the finding was not entirely unexpected given the similarity of the average RTE scores between male and females across threshold calculation method (see Table 20).

Makeup testing session attendance. Descriptive statistics of the RTE scores by makeup testing session attendance status are presented in Table 22. Examination of the means indicated respondents who completed the makeup tests as walk-ins at a computer lab had lower RTE scores, on average, than did students who attended the originally scheduled makeup testing session. As hypothesized, all of the correlations between makeup attendance status and RTE were negative which indicates respondents who completed the makeup tests as walk-ins had lower RTE scores than respondents who completed the tests during the originally scheduled makeup session. However, the only correlations that were statistically significant were $RTE_{INSPECT}$, $RTE_{INSPECT2}$, and $RTE_{MIXTURE}$.

Results of the GEE examining the relationship between makeup testing session attendance status and the logit of RTE revealed the relationship was dependent on threshold calculation method (see Table 19). The unstandardized slope coefficients and robust standard errors of the model, as well as the simple slopes examining the relationship between makeup testing attendance status and the logit of RTE by threshold calculation method are presented in the bottom of Table 24. All of the simple slopes examining the relationship between makeup testing session attendance and the logit of RTE were negative, indicating respondents who completed the makeup tests as walk-ins

had lower RTE scores than students who completed the makeup tests during the originally scheduled time. In addition, based on four of the threshold calculation methods, the differences between the groups were statistically different from zero (see Table 24). Specifically, makeup testing session attendance status was negatively related to the logit of RTE when calculated using INSPECT, INSPECT2, MIXTURE, and NT20. The threshold calculation methods NT10 and NT30 also yielded negative slopes similar in magnitude to other methods, but were not significantly different from zero. This finding was not surprising given the standard errors for the simple slopes were large. Given the focus of the current analysis was on examining the differential relationships between attendance status and the logit of RTE across threshold calculation methods, the remainder of the discussion will focus on the pairwise comparisons.

Pairwise comparisons examining the differential relationships between attendance status and the logit of RTE across threshold calculation methods are presented in Table 25. To help interpret these results and assess practical significance, graphs displaying the model-implied relationships between makeup testing attendance status and the logit of RTE and the model-implied relationships between makeup testing attendance status and predicted RTE for each threshold calculation method are presented in Figure 8. Results of the pairwise comparisons in Table 25 indicated the slopes calculated using INSPECT, INSPECT2, MIXTURE, NT10, NT20, and NT30 were not significantly different from one another; however, with the exception of NT10 and NT30, all were significantly different from the slopes calculated using MIXTURE2 and RSPEED. Interestingly, the overlapping slopes presented in the top graph of Figure 8 indicates the relationship between makeup testing session attendance status and the logit of RTE appears to be

identical for both the INSPECT2 and MIXTURE threshold calculation methods. Although it is subjective, the top graph of Figure 8 can be used to ascertain the practical significance of the differences between simple slopes. Indeed, the simple slopes appear steeper for INSPECT, INSPECT2, MIXTURE, NT20, and NT30 relative to the slopes for NT10, MIXTURE2, and RSPEED. The differences among the slopes, however, do not appear extreme.

The bottom graph of Figure 8 can also be used to ascertain the practical significance of the results. Unlike the logit scale in the top graph, which is unbounded, the RTE scale in the bottom graph is bounded between zero and one. Because the transformation from the logit scale to the RTE scale is nonlinear and because many of the RTE values are near one, relationships that were significant on the logit scale may not seem practically significant on the probability scale. For instance, the difference between respondents who completed the tests as walk-ins and those who completed the tests during the original makeup testing session on the logit RTE scale for NT20 was -1.11, which was the largest difference across all calculation methods. After the nonlinear transformation to the RTE 0/1 scale, however, the difference between the two groups of respondents was only 0.01, which was one of the smallest differences observed across all calculation methods.

Despite the possible discrepancies between the same relationship graphed two different ways (as in the top and bottom of graphs in Figure 8), the bottom graph of Figure 8 was also used to assess practical significance because it portrays the findings on the more interpretable RTE scale. It is important to note that the predicted RTE values according to the model as shown in the bottom graph of Figure 8 are the same as the RTE

values provided in Table 22. Although the slopes on the logit RTE for INSPECT, INSPECT2, MIXTURE, NT20, and NT30 were not significantly different from one another, some practical differences did appear. As displayed in the bottom graph of Figure 8, the relationship between makeup testing attendance status and RTE calculated using NT20 appears to be practically different from the relationships calculated using INSPECT, INPSECT2, NT30, and MIXTURE. There was essentially no practical difference in the RTE_{NT20} scores for respondents who completed the tests as walk-ins from respondents who completed the tests during the makeup sessions (difference = .01, Table 22). In contrast, there were larger differences between the groups when RTE was calculated using INSPECT, INSPECT2, NT30, and MIXTURE (0.04, 0.06, 0.03, and 0.05, respectively). Even though the threshold calculation methods differed somewhat in their relationships between makeup testing session attendance status and RTE, the RTE differences between walk-in and make-up respondents could be characterized as either negligible or small for all threshold calculation methods. For instance, differences of 0.00 (NT10) or 0.01 (NT20) between the two groups in RTE could be considered negligible, whereas differences of 0.05 or larger (MIXTURE, INSPECT2) could be considered small, but not negligible.

Effort. Descriptive statistics of the effort scores are presented in Table 23. Correlations between RTE and effort scores are presented in Table 21. The magnitude of the correlations ranged from $r = 0.10$ to $r = 0.27$ and all of the correlations except for the correlation between effort and RTE_{NT10} were significant at a .01 alpha level. The significant correlations were in the hypothesized direction and magnitude, indicating there was a positive relationship between self-reported effort and RTE.

The initial GEE used to examine the relationship between effort and the logit of RTE failed to converge to a solution. To help diagnose convergence problems, descriptive statistics and correlations between effort and the RTE scores were reexamined. Because RTE_{NT10} was not significantly related to effort, and given previous problems associated with this threshold calculation method, RTE_{NT10} was dropped from the model and a modified GEE predicting the logit of RTE (and still using RSPEED as a reference variable) was specified as

$$\log\left(\frac{RTE_i}{1-RTE_i}\right) = \beta_0 + \beta_1(INSPECT_i) + \beta_2(INSPECT2_i) + \beta_3(MIXTURE_i) + \beta_4(MIXTURE2_i) + \beta_5(NT20_i) + \beta_6(NT30_i) + \beta_7(Characteristic) + \beta_8(Characteristic)(INSPECT_i) + \beta_9(Characteristic)(INSPECT2_i) + \beta_{10}(Characteristic)(MIXTURE_i) + \beta_{11}(Characteristic)(MIXTURE2_i) + \beta_{12}(Characteristic)(NT20_i) + \beta_{13}(Characteristic)(NT30_i) \quad (13)$$

The modified GEE successfully converged to a solution.

The result of the omnibus test between effort and threshold calculation method was significant, thus indicating the relationship between effort and logit of RTE was dependent on threshold calculation method (see Table 19). The unstandardized slope coefficients, robust standard errors, and simple slopes examining the relationship between effort and the logit of RTE across threshold calculation methods are presented in Table 26. All of the simple slopes for the threshold calculation methods were positive and significant. Pairwise comparisons examining the differential relationships across threshold calculation methods are presented in Table 27. To help interpret the results and assess practical significance, graphs displaying the model-implied relationships between

effort and the logit of RTE and the model-implied relationships between effort and predicted RTE for each threshold calculation method are presented in Figure 9.

Examination of the pairwise comparison presented in Table 27 revealed several of the slopes exhibited significant differential relationships. In particular, the majority of the significant differential relationships occurred with RSPEED and MIXTURE2, which can easily be seen in the top graph of Figure 9. For example, the slope calculated using RSPEED significantly differed from five of the other slopes. The smaller magnitude of these slopes in comparison to the other relationships suggests that the logit RTE computed using the MIXTURE2 and RSPEED calculation methods is not as strongly related to self-reported effort from those who are rapidly responding as well as some of the other threshold calculation methods. Interestingly, as previously exhibited, the slopes calculated using INSPECT2 and MIXTURE perfectly overlapped each other indicating there was no difference between the methods. The simple slope for INSPECT2 and MIXTURE significantly differed from the slightly larger simple slopes of INSPECT and NT30, which were not different from one another.

Although the results indicated the relationship between self-reported effort and the logit of RTE significantly differed between INSPECT2/MIXTURE and INSPECT/NT30, practically there was not a large difference between the methods as indicated by lack of large differences in the slopes for these methods in the top graph in Figure 9. Overall, examination of the bottom graph in Figure 9 revealed the INSPECT, INSPECT2, MIXTURE, and NT30 calculation methods displayed similar relationships with each other and differed most from the relationships displayed by NT20, MIXTURE2, and RSPEED. With respect to which methods yielded effort/RTE

relationships of practical importance, the bottom graph of Figure 9 indicates a negligible relationship between the two variables using the NT20 method and small, non-negligible relationships for the remaining methods.

Academic ability. Descriptive statistics of the SAT-Mathematics (SAT-M) and SAT-Critical Reading (SAT-CR) scores are presented in Table 23. Correlations between RTE scores with the two independent measures of academic ability, SAT-M and SAT-CR are presented in Table 21. As hypothesized, SAT-M scores were not significantly related to any of the RTE scores. However, SAT-CR did exhibit a significant negative relationship with $RTE_{MIXTURE2}$ and RTE_{RSPEED} ($p < .01$). For clarity, the current section will present the results of the model used to examine the relationship between SAT-M and the logit of RTE first, and will then present the results of the model used to examine the relationship between SAT-CR and the logit of RTE.

The GEE used to examine the relationship between SAT-M and the logit of RTE failed to converge to a solution. Given the difficulties encountered comparing the proportion of classified respondents in Phase Two with thresholds calculated using NT10 and due to the absence of a significant correlation between SAT-M and RTE_{NT10}, a modified GEE without the NT10 calculation method was estimated (see Equation 13). The results of the modified GEE examining the relationship between SAT-M and the logit of RTE successfully converged to a solution. The result of the omnibus test assessing the interaction between SAT-M and threshold calculation methods was not significant, which indicated the relationship between SAT-M and the logit of RTE was not dependent on threshold calculation method (see Table 19). The nonsignificant interaction was dropped and the model was re-estimated to examine the significance of

the main effect of SAT-M. The results indicated the main effect of SAT-M was not significantly related to the logit of RTE, thus supporting the hypothesis that an independent measure of academic ability is not related to the logit of RTE, $F(1,450) = 0.16$, $p = 0.69$.

The GEE used to examine the relationship between SAT-CR and the logit of RTE also failed to converge to a solution. A modified GEE examining the relationship between SAT-CR and the logit of RTE without the NT10 threshold calculation method was estimated using Equation 13, but also failed to converge to a solution. Given the problems encountered with NT20 in Phase Two, a second modified GEE was estimated whereby both NT10 and NT20 threshold calculation methods were removed. This modified model successfully converged to a solution. The results of the omnibus test assessing the interaction between SAT-CR and threshold calculation methods was not significant, indicating the relationship between SAT-CR and the logit of RTE was not dependent on threshold calculation method (see Table 19). The nonsignificant interaction was dropped and the model was re-estimated to examine the significance of the main effect of SAT-CR. Interestingly, the results indicated the main effect of SAT-CR was statistically significant, $F(1,450) = 6.76$, $p = 0.0096$; thus indicating SAT-CR was negatively related to the logit of RTE when controlling for threshold calculation method, $b = -0.0024$. This result is counter to the hypothesis that RTE would not be significantly related to an independent measure of academic ability. Although small in magnitude, the negative relationship between SAT-CR and logit of RTE suggests that respondents with high critical reading scores might read faster than respondents with lower critical reading scores and thus have lower RTE scores. However, given the largest correlations between

RTE and SAT-CR scores in Table 21 were with MIXTURE2 and RSPEED, this suggests that using a more liberal time threshold calculation method may misclassify respondents who are fast readers as not putting forth effort in responding.

Individual consistency index. Descriptive statistics of the individual consistency index values are presented in Table 23. Correlations between the individual consistency index values and RTE scores are presented in Table 21. All of the correlations were positive thus indicating respondents with higher scores on the individual consistency index had higher RTE scores. Six of the eight RTE scores were significantly related to the index ($p < .01$); neither of the RTE scores calculated using the NT10 and NT20 methods were significantly related. The GEE used to examine the relationship between the index and the logit of RTE failed to converge to a solution. Given the index was not related to RTE_{NT10} and given the previous trouble associated with NT10, this effect was dropped from the model and a modified GEE was estimated (see Equation 13). Results from the modified GEE indicated the relationship between the individual consistency index and the logit of RTE was dependent on threshold calculation method (see Table 19).

The unstandardized slope coefficients and robust standard errors are presented in Table 28 along with simple slopes examining the relationship between the logit of RTE and the individual consistency index across threshold calculation methods. Examination of the simple slopes revealed all of the slopes were statistically significant with the exception of NT20 and RSPEED. Pairwise comparisons examining the differential relationships across threshold calculation methods are presented in Table 29. To help interpret the results and assess practical significance, graphs displaying the model-

implied relationships between the individual consistency index and the logit of RTE and the model-implied relationships between the individual consistency index and predicted RTE for each threshold calculation method are presented in Figure 10.

The pairwise comparisons of simple slopes indicated the simple slopes for RSPEED and MIXTURE2 were significantly different from one another and from all other threshold calculation methods, with the exception of NT20. Interestingly, the simple slope for NT20 did not differ from any other simple slope (perhaps due to the low variance associated with RTE_{NT20} and correspondingly high standard error). The remaining four calculation methods (INSPECT2, MIXTURE, INSPECT, and NT30) had relatively higher simple slopes that did not significantly differ from one another. Inspection of the top graph in Figure 10 indicated minor differences among all threshold calculation methods; the slopes for INSPECT, INSPECT2, MIXTURE, and NT30 appeared slightly steeper than the slopes for RSPEED and MIXTURE2. As previously found, the slopes defined by INSPECT2 and MIXTURE are overlapping in Figure 10, thus indicating there was no difference in these relationships. Examination of the slopes in the bottom graph of Figure 10 indicates there no practical differences between the relationships calculated using the various threshold calculation methods. Moreover, the graph also reveals that the relationship displayed between the individual consistency index and RTE was relatively weak for all of the threshold calculation methods.

Length of open-ended response option. Descriptive statistics of the response length to an open-ended question provided by respondents included in the Makeup Testing 2015 sample are presented in Table 23. The average length of response provided by respondents was 116.2 words (*SD* = 94.7). As seen in Table 21, response length was

significantly and positively related to INSPECT, INSPECT2, MIXTURE, and MIXTURE2, thus supporting the hypothesis that lengthier responses would be positively related to higher levels of effort.

The result of the GEE failed to converge to a solution. Given the absence of a significant relationship between length of response and RTE_{NT10}, and given previous issues encountered with this threshold calculation method, a modified GEE was estimated whereby the effect of RTE_{NT10} was removed from the model (see Equation 13). Unfortunately, the modified GEE also failed to converge to a solution. A series of modified GEE were subsequently estimated whereby problematic methods encountered in Phase Two (e.g., NT20, MIXTURE2) were dropped from the models. Ultimately, none of the modified models successfully converged to a solution. The failure of the models to converge to a solution suggests the GEE is not an appropriate model. Thus, a different approach to estimating the model was taken. Specifically, a modified model was estimated in which all of the original RTE scores were included in the model but the repeated measures nature of the data was not taken into account. It is important to note that failure to account for the within subject correlations introduced by the repeated measures will underestimate the standard errors which in turn increases the rate of Type I errors (Burton et al., 1998). Thus, these results should be cautiously interpreted. The modified model including all RTE scores but not accounting for the within subjects nature of RTE successfully converged to a solution (see Table 19). The omnibus test assessing the interaction between response length and threshold calculation method was not significant which indicates the relationship between response length and RTE does not depend on the threshold calculation method. Given the nonsignificant interaction, the

relationship between response length and the logit of RTE controlling for threshold calculation method was assessed by dropping the nonsignificant interaction and re-estimating the model. After controlling for threshold calculation method, response length was significantly related to the logit of RTE, $F(1,2655) = 42.07$, $p < 0.0001$. These results support the hypothesis the length of a response to an open-ended question is positively related to RTE.

RTF and item characteristics. RTF scores for each of the threshold calculation methods were calculated using Equation 3 specified in Chapter 2. Recall, RTF scores reflect the proportion of respondents exhibiting solution behavior on an item. RTF scores were not calculated for one item defined using the MIXTURE calculation method and for three items defined using the MIXTURE2 calculation method. Descriptive statistics of the RTF scores indicated higher proportion of respondents were classified as exhibiting solution behavior on items by the NT methods whereas smaller proportions of respondents were classified as exhibiting solution behavior across items when calculated using RSPEED and MIXTURE2 methods (see Table 30). Descriptive statistics indicated the RTF_{NT10} scores had no variability ($SD = 0$). Thus, it was anticipated that estimation problems similar to those encountered with RTE_{NT10} would be encountered. The magnitude of the correlations between the RTF scores ranged from -0.32 to 0.86 (see Table 30). Unexpectedly, RTF_{NT10} and $RTF_{MIXTURE2}$ both displayed negative relationships with other RTF scores. The lack of variability displayed by the RTF_{NT10} scores can account for its negative relationship with other RTF scores. However, the negative relationships between $RTF_{MIXTURE2}$ and other scores, including RTF_{RSPEED} is noteworthy.

Generalized estimating equations (GEEs) were used to examine the relationship between the logit of RTF and two characteristics of MFLS items – item length and item position. Both models successfully converged to a solution. The results of the omnibus tests assessing the significance of the interaction between each item characteristic and calculation method as well as the main effects are presented in Table 31. The results for each item characteristic examined are reviewed below.

Item position. Descriptive statistics of the serial position of the MFLS items are presented in Table 32 and correlations between item position and the RTF scores are presented in Table 33. Item position exhibited a statistically significant negative relationship with RTF scores calculated using the NT20 and NT30 threshold calculation methods. Results of the GEE revealed the omnibus test assessing the interaction between item position and threshold calculation methods was statistically significant, thus the relationship between item position and the logit of RTF was dependent on the threshold calculation method used (see Table 31).

The unstandardized coefficients and robust standard errors are presented in Table 34 along with simple slopes examining the relationship between item position and the logit of RTF by threshold calculation method. Statistical tests of the simple slopes for INSPECT2 and NT10 presented in Table 34 were not calculated by SAS because the models failed to converge. As a result, these simple slopes were calculated by hand; the value of both simple slopes were equal to zero, indicating item position was not related to the logit of RTF when RTF was calculated using INSPECT2 or NT10. Overall, the magnitude of the simple slopes was small; the only statistically significant slopes were calculated using NT20, NT30, and RSPEED (see Table 34). The direction of the

significant relationships was negative, thus supporting the hypothesis that RTF was negatively related to item position for these methods. Interestingly, the simple slope for MIXTURE2 was positive, which is counter to the typical negative relationship seen between item position and effort in low-stakes cognitive assessment. The lack of significance among the majority of the threshold calculation methods is contrary to what was hypothesized, although not surprising given the small magnitude of the coefficients. Given the other threshold calculation methods were not significantly related to item position, the remainder of the discussion of these results will focus on the three threshold calculation methods that were significantly related.

Pairwise comparisons examining differential relationships between item position and the logit of RTF across threshold calculation methods are presented in Table 35. To help interpret the results and assess practical significance, graphs displaying the model-implied relationships between item position with the logit of RTF and predicted RTF for each threshold calculation method are presented in Figure 11. The comparisons revealed the majority of the significant differential relationships that occurred between the methods that yielded negative slopes (i.e., INSPECT, NT20, NT30, and RSPEED) and the methods that yielded slopes with a magnitude of zero (see Table 35). Pairwise comparisons between the slopes that were negative (i.e., INSPECT, NT20, NT30, and RSPEED) were not significantly different from one another with the exception of the slope calculated using NT20, which was significantly steeper than the slopes calculated using INSPECT and NT30.

Examination of the top graph in Figure 11 shows differences in the slopes calculated using INSPECT, NT20, NT30, and RSPEED; however, the bottom graph in

Figure 11 revealed the slope calculated using RSPEED was practically different and much steeper from the slopes calculated using INSPECT, NT20, and NT30. This relationship, as well as the relationships with NT methods, should be cautiously interpreted because they are confounded by item length. That is, RSPEED calculates time thresholds based on the number of words in an item: longer items yield larger time thresholds. Similarly, NT methods calculate time thresholds based on an items mean response time: longer items yield larger time thresholds. In addition to these considerations, item position exhibited a positive relationship with item length in the current study (r = 0.45). As a result, given how RSPEED and the NT methods calculate time thresholds, and given the relationship between item position and item length, the relationships between item position and RTF calculated using RSPEED and the NT methods are conflated with the relationship between item length and RTF scores calculated using these methods. Although the relationships between item position and RTF scores calculated using the NT methods are also confounded, the effects were not as pronounced in the current analyses. Overall, the bottom graph of Figure 11 conveys there is essentially no relationship between item position and RTF, with the exception of RSPEED and MIXTURE2.

Item length. Descriptive statistics of the length of items on the MFLS are presented in Table 32 and correlations between the RTF scores and item length are presented in Table 33. The average length of items on the MFLS was 11.13 words (*SD* = 3.95). Item length exhibited a significant negative relationship with RTF when calculated using INSPECT, NT20, NT30, and RSPEED. The negative relationship between item length and RTF was anticipated for the NT methods and RSPEED based on how they

calculate time thresholds. That is, time thresholds calculated by the NT methods are a function of the total response time to an item. Given item length was positively related to mean item response time in the current study (r = .43), longer items will yield larger response times and larger time thresholds and in turn, lower RTFs. Similarly, as previously discussed, time thresholds calculated by RSPEED are a function of an item's length (i.e., the total number of words). Thus, given lengthier items occurred towards the end of the MFLS, longer items will have larger time thresholds and accordingly, lower RTFs.

The GEE examining the relationship between item length and the logit of RTF successfully converged to a solution. Results indicated the omnibus test of the interaction between item length and threshold calculation methods was statistically significant, thus, the relationship between item length and the logit of RTF was dependent on threshold calculation method (see Table 31). The unstandardized slopes coefficients and robust standard errors of the model are presented in Table 36 along with the simple slopes examining the relationship between item length and the logit of RTF for each threshold calculation method. Four simple slopes associated with the INSPECT, NT20, NT30, and RSPEED methods exhibited a significant negative relationship with item length. Interestingly, the slopes for MIXTURE2, and to a lesser extent NT10, were positive (but not significant), which is the opposite direction of the relationship typically observed in the low-stakes cognitive assessment literature.

Pairwise comparisons examining the differential relationships between item length and the logit of RTF across threshold calculation methods are presented in Table 37. To help interpret the results and assess practical significance, graphs displaying the

model-implied relationships between item length and the logit of RTF and the model-implied relationships between item length and predicted RTF for each threshold calculation method are presented in Figure 12. Pairwise comparisons presented in Table 37 revealed a great deal of overlap among the simple slopes, with the exception of MIXTURE2, which had a positive slope that was significantly different from almost all other calculation methods, and RSPEED, which had the largest negative slope and was significantly different from all other calculation methods. The uniqueness of the slopes for these two threshold calculation methods is readily apparent in both the top and bottom graphs of Figure 12. Considering the two graphs together, there does not seem to be incredibly large differences between the NT methods, INSPECT, INSPECT2, and MIXTURE in the nature of the item length/RTF relationship. In fact, the bottom graph in Figure 12 indicates the relationships between item length and RTF for these calculation methods are negligible. In contrast, the relationship between item length and RTF is very large when calculated using RSPEED. As previously mentioned, this relationship was expected and is an artifact of the time thresholds being calculated as a function of item length for the RSPEED method. In contrast, the relationship between item length and RTF for MIXTURE2 is meaningful in magnitude and positive, which is counter to the typical negative relationship seen between item length and effort in low-stakes cognitive assessment.

In summary, the results from the Phase Three analysis found evidence that some of the relationships between the external variables with either RTE or RTF were dependent on threshold calculation method. Specifically, three respondent characteristics displayed a significant interaction with threshold calculation method (makeup testing

session attendance status, effort, and the individual consistency index) and both of the item characteristics (item position and item length) exhibited a significant interaction with threshold calculation method. The simple slopes for these five models are presented in Table 38. Some similarities within the respective analyses (i.e., RTE and RTF) emerged. For example, within the RTE analyses, the NT30, INSPECT, INSPECT2, and MIXTURE threshold calculation methods exhibited meaningful relationships in the hypothesized direction with all three of the respondent characteristics. Similarly, within the RTF analyses, NT20, NT30, and RSPEED were significantly related and in the hypothesized direction to both item position and item length. Although several of the relationships were statistically significant, few appeared to be practically significant. For example, item length was statistically related with RTE scores calculated using: INSPECT, NT20, NT30, and RSPEED. However, examination of these slopes presented in the bottom graph of Figure 12 revealed the only relationship that appeared practically meaningfully was RSPEED.

Overall, the results from Phase Three provided supportive validity evidence for the NT30, INSPECT, INSPECT2, and MIXTURE threshold calculation methods. The relationships displayed between INSPECT2 with the respondent and item characteristics were almost identical to the relationships observed with MIXTURE. This suggests that researchers can use either method and will achieve similar results. When it was included in the model, none of the respondent or item characteristics were significantly related to the NT10 method, which was not surprising due to its lack of variability and very low reliability. Another pattern observed across the analyses was the tendency for the relationships calculated using NT20 to be less distinct across characteristics. That is,

although the slopes were significantly different from zero, often the slopes were not practically different, thus indicating NT20 might be too conservative in calculating the time thresholds to classify respondents into distinct groups. Similarly, the majority of significant differential relationships that occurred between threshold calculation methods were relationships that were not of substantive interest or were relationships pertaining to the RSPEED method, the MIXTURE2 method, or both. The differential relationships displayed with RSPEED and MIXTURE2 suggests these threshold calculation methods may be tapping into something other than effort and are misclassifying respondents. These findings, how they integrate with the results from Phases One and Two, and their implications are discussed in the next chapter.

CHAPTER FIVE

Discussion

Given noncognitive measures are increasingly used for accountability purposes, and the negative impact responding without effort has on the validity of results, there is a need to discreetly identify respondents displaying low effort on low-stakes noncognitive measures. One method based on response time that can discreetly assess student effort at the item level is the solution behavior (SB) index (Kong et al., 2007). A challenging task in using the SB index, however, is the identification of an appropriate time threshold that meaningfully distinguishes solution behavior responses (i.e., responses made with effort) from responses occurring so quickly (i.e., rapid responding) the responses are essentially meaningless (DeMars, 2007; Swerdzewski et al., 2011). Although the SB index has been extensively used with low-stakes cognitive tests (e.g., Kong et al., 2007; Wise, 2006), it has only been used once with low-stakes noncognitive measures (Swerdzewski et al., 2011). Fundamentally, items on cognitive tests differ from items on noncognitive measures: items on cognitive tests tend be longer in length and more complex than items on noncognitive measures, suggesting the response times for items on cognitive tests are longer in length and more variable than response times to items on noncognitive measures.

Given the dearth of research examining the application of the SB index to noncognitive measures and given the differences between items on cognitive and noncognitive assessments, the purpose of the current study was to examine if the SB index could be used with low-stakes noncognitive measures to distinguish responses – and ultimately respondents – exhibiting solution behavior from responses made without

any effort. In particular, it was of interest to determine: (a) if time thresholds for items on a 53-item noncognitive measure assessing the construct meaningful life could be calculated using eight different threshold calculation methods, (b) if the defined time thresholds and resulting SB classification indices differed across the threshold calculation methods, and (c) if the resulting SB classifications were meaningfully related to external criteria in theoretically expected ways. In addition, it was also of interest to examine if including data from a known group of rapid responders would have an effect on the defined time thresholds and subsequent SB classification indices. To that end, eight threshold calculation methods were used in the current study to define the SB time thresholds: visual inspection (INSPECT), visual inspection with information (INSPECT2), lognormal mixture modeling (MIXTURE), lognormal mixture modeling with information (MIXTURE2), 10% normative threshold (NT10), 20% normative threshold (NT20), 30% normative threshold (NT30), and a constant based on predicted reading speed (RSPEED). The INSPECT and MIXTURE threshold calculation methods were applied twice: once using response time data from the primary sample of respondents (i.e., INSPECT and MIXTURE) and a second time using response time data from the primary sample of respondents combined with data from a known group of rapid responders (i.e., INSPECT2 and MIXTURE2). The results of the current study, practical implications, and limitations are discussed in detail below.

Phase One Results

The results of Phase One indicated that the SB time thresholds could be defined for items on a low-stakes noncognitive measure assessing the construct meaningful life. Across threshold calculation methods, the defined time thresholds were small, ranging

from 0.5 to 3.3 seconds, on average. In comparison to time thresholds defined for items on cognitive tests (e.g., Pastor et al., 2015; Strickman et al., 2015), the range of time thresholds for items on the MFLS were much shorter on average. For example, time thresholds calculated for items on a cognitive test reported by Wise and Kong (2005) ranged from 3 to 10 seconds whereas time thresholds for a different cognitive test reported by Pastor et al. (2015) ranged from 3 to 22 seconds. Similarly, time thresholds reported by Strickman et al. (2015) ranged from 3 to 40 seconds. This finding was anticipated given items on noncognitive measures are typically shorter and less complex than items on cognitive tests. Future researchers should examine if this pattern holds with other noncognitive measures that measure different constructs besides meaningful life.

Results addressing the research question of whether time thresholds could be set for the items indicated only minor issues limited to the MIXTURE and MIXTURE2 threshold calculation methods. Recall, time thresholds can always be set using the NT methods and RSPEED. The question of interest was whether issues would be encountered using the distributional methods (i.e., INSPECT, INSPECT2, MIXTURE, and MIXTURE2), which work the best when the response time distributions of rapid responders are distinct from non-rapid responders. In particular, the MIXTURE method was unable to define a time threshold for one item and the MIXTURE2 method was unable to define a time threshold for three different items. Both of these methods were unable to define time thresholds for these items because the mixture distributions of the two classes were not distinguishable, which suggests the models may have been misspecified. In other words, perhaps the one-class model should have been championed for these items instead of the two-class model. For instance, the BIC index calculated

using MIXTURE indicated the one-class model for item 8 fit better than the two-class model; however, given preference was given to the SSABIC index, a two-class model for item 8 was championed. Although this pattern was not observed for item 8 when MIXTURE2 was used, the two mixture distributions for item 8 calculated using MIXTURE2 were not distinguishable from one another (see Figure 5), thus suggesting the model may have been misspecified.

Previous research applying the SB index to cognitive tests has also found evidence of model misspecification. For example, in their seminal study developing and applying the SB index to high-stakes cognitive speeded tests, Schnipke and Scrams (1997) found the two-class lognormal mixture did not fit for several of the items at the beginning of the test because "there were not enough rapid guesses to obtain reliable parameter estimates of the guessing distribution" (p. 231). Given the majority of the time thresholds were calculated for the 53 items on the MFLS, the defined time thresholds and SB classification indices were compared across threshold calculation methods in Phase Two, which is described below.

Phase Two Results

Results of Phase Two indicated the magnitude of the time thresholds and SB classification indices differed across the threshold calculation methods and the calculation methods could be rank-ordered based on the magnitude of the results. Specifically, the time thresholds defined by the NT methods were smaller, or more conservative in comparison to the time thresholds defined by the other methods. In particular, the time thresholds and SB classification indices defined by NT10 displayed very little variability, which subsequently created problems with other analyses. In

contrast, the time thresholds defined by MIXTURE2 and RSPEED were larger, or more liberal in comparison to the time thresholds defined by other methods and exhibited low rates of exact and approximate agreement across calculation methods.

The time thresholds defined by INSPECT, INSPECT2, and MIXTURE were moderate in magnitude and displayed the highest rates of exact agreement across methods. Similarly, the resulting SB classification indices calculated by INSPECT, INSPECT2, and MIXTURE exhibited the least differences across threshold calculation methods. This last set of results are similar to findings by Kong et al. (2007) and Pastor et al. (2015) who also found little differences between the INSPECT and MIXTURE methods.

Interestingly, the inclusion of the Known Rapid Responders sample appeared to have a larger impact on the time thresholds defined using MIXTURE2 than it did on the time thresholds that were defined using INSPECT2. Specifically, the inclusion of a known rapid responding group made the time thresholds more distinct for MIXTURE vs. MIXTURE2 in comparison to INSPECT vs. INSPECT2. Defining an appropriate time threshold to distinguish solution behavior responses from rapid responses is a challenging task. To help understand these results and assess if the resulting SB classifications were meaningful, results of the external validity evidence gathered in Phase Three are reviewed below.

Phase Three Results

External validity evidence for the resulting time thresholds and SB classification indices was collected by examining the relationships between aggregated measures of effort across items (i.e., RTE) and respondents (i.e., RTF) with respondent and item

characteristics, respectively. Specifically, analyses were conducted to examine: (a) the relationships between the logit of RTE and gender, makeup testing session attendance status, self-reported effort scores, academic ability, the individual consistency index, and the response length to an open-ended question; and (b) the relationships between the logit of RTF and position and length of the MFLS items. A-priori hypotheses about the relationships between the logit of RTE scores with the respondent characteristics and between the logit of RTF scores with the item characteristics were made. Given the exploratory nature of the study, no a-priori hypotheses were made about the differential relationships that might occur between the threshold calculation methods and respondent and item characteristics. Overall, the results of Phase Three found supporting validity evidence for some, but not all of the threshold calculation methods.

Relationships with respondent characteristics. For clarity, the results in the current section will be discussed in the following order. First, the results of the four models examining the relationship between logit of RTE and respondent characteristics that *were not* moderated by the threshold calculation method will be discussed. Then, the results of the three models examining the relationship between logit of RTE and respondent characteristics that *were* moderated by the threshold calculation method will be discussed.

***Gender*.** The hypothesis gender would be significantly related to logit of RTE was not supported, although the majority of the observed correlations between gender and logit of RTE were in the direction hypothesized. One reason the relationship between gender and logit of RTE were not significant may be because the RTE scores are not measuring effort, but are instead measuring something else such as response time.

Similarly, another reason why the relationship was not significant may be because some of the threshold calculation methods that yielded larger time thresholds (e.g., MIXTURE2, RSPEED) are misclassifying motivated respondents as rapid-responders, which is in turn driving the nonsignificant relationship. A third reason gender failed to display a significant relationship with logit of RTE may be due to the absence of gender differences in rapid responding on noncognitive measures. That is, although the majority of previous research examining motivation on low-stakes cognitive tests has found females typically exhibit higher amounts of effort on low-stakes tests than do males (e.g., DeMars et al., 2013; Setzer et al., 2013; Wise et al., 2009), this may not be the case on noncognitive measures or at least on noncognitive measures of the construct meaningful life. Future research should examine whether gender differences in effort (as measured by either the SB index or RTE) occurs using different noncognitive measures. A fourth reason the relationship between gender and the logit RTE was not significantly related may be due to the sample used. That is, the majority of respondents in the Primary sample were attendants of a makeup testing session – only 77 respondents attended the originally scheduled testing session on Assessment Day. Thus, it could be that male and female students attending makeup testing sessions do not display differential amounts of effort when attending makeup testing sessions. A final reason the relationship was nonsignificant may have been because females were not as interested in the construct being assessed.

Academic ability. Similar to the relationship between response length and logit of RTE, the relationship between SAT-M and logit of RTE also failed to be dependent on threshold calculation method. However, this result was not entirely surprising given it

was hypothesized RTE would not be related to SAT-M scores. The magnitude of the observed correlations displayed between the RTE scores and SAT-M scores (i.e., -.03 to -.08) in the current study were similar in magnitude and direction to correlations between RTE and SAT-M scores reported in other studies (e.g., Kong et al., 2007; Wise & Kong, 2005).

In contrast to SAT-M, the hypothesis that SAT-CR scores would not be related to logit of RTE was not supported. Specifically, after controlling for threshold calculation method, SAT-CR scores were negatively related to logit of RTE, indicating respondents with higher SAT-CR scores were less likely to exhibit solution behavior on the MFLS. One reason this relationship may have been significant is because RTE may not be measuring effort, but instead measuring something different, such as reading skills. However, a small amount of previous research has also found this result. For example, Wise and Kong (2005) found RTE scores based on a low-stakes cognitive test were negatively related to SAT verbal scores (r = -.08), albeit the relationship was not statistically significant. Another reason the relationship between logit of RTE and SAT-CR scores may have been significant could have been due to the threshold calculation methods misclassifying respondents. In particular, the observed correlations between the MIXTURE2 and RSPEED threshold calculation methods were the largest in magnitude across methods. Given these two calculation methods yielded the largest time thresholds and subsequently classified the largest proportion of respondents as rapidly responding, it is possible that motivated respondents were misclassified as rapidly responding and it is this misclassification that is driving the significant main effect.

***Length of open-ended response*.** The relationship between the length of the open-ended response and logit of RTE did not depend on the threshold calculation method, although the relationship was in the hypothesized direction. This significant main effect of response length should be interpreted with caution, however, given the GEE used to analyze this model ignored the within subject correlations of the data introduced by the repeated measures of the RTE scores (Burton et al., 1998). One possible reason the relationship between response length and logit of RTE was not moderated by threshold calculation method could have been due to the non-normality of the RTE scores and the length of the open-ended responses (Goodwin & Leech, 2006). Although a GEE was used to take the non-normality of the RTE scores into account (which were negatively skewed), the non-normality of the length of the open-ended responses (which were positively skewed) may have driven the relationship downward. However, a second possible reason the relationship was not significant was because only a subsample of the Primary sample was used in this analysis. Recall, only participants attending the Makeup 2015 Testing Session were administered the open-ended question after completing the MFLS. Thus, this subsample of students could have differed from the larger sample in some way. Finally, a third possible reason that the relationship between the length of the open-ended response item and logit of RTE was not moderated by threshold calculation method was simply because there was not an effect.

***Makeup testing session, effort, and individual consistency index*.** The results indicated the relationships between the logit of RTE with makeup testing session attendance status, self-reported effort, and the individual consistency index were dependent on threshold calculation method. Overall, the three respondent characteristics

exhibited meaningful relationships in the hypothesized direction with RTE when calculated using NT30, INSPECT, INSPECT2, and MIXTURE. For example, as hypothesized, makeup testing session attendance status was negatively related to RTE when calculated using NT30, INSPECT, INSPECT2, and MIXTURE. Similarly, self-reported effort and the individual consistency index were both positively related to RTE when calculated using NT30, INSPECT, INSPECT2, and MIXTURE methods, as hypothesized.

The results also indicated that across respondent characteristics, the relationships calculated using INSPECT2 and MIXTURE were nearly identical and some differential relationships did occur across respondent characteristics. For example, when related to makeup testing session attendance status, differential relationships small in magnitude occurred between the relationships calculated using INSPECT, INSPECT2, MIXTURE, and NT30. In contrast, when related to self-reported effort scores, there were no practical differences in the relationships calculated using INSPECT, INSPECT2, MIXTURE, and NT30 methods; these relationships were similar in magnitude and differed most from each other when reported effort was low and differed most overall from the relationships calculated using NT20, MIXTURE2, and RSPEED (NT10 was not included in the model). When related to the individual consistency index, the relationships calculated using NT30, INSPECT, INSPECT2, and MIXTURE methods were relatively similar to one another and displayed stronger positive relationships with the individual consistency index than did the RTE scores calculated using RSPEED and MIXTURE2. However, the magnitude of the relationships between the individual consistency index and RTE across

calculation methods was relatively small and overall there were no practical differences in the relationships.

Across the three analysis, RTE calculated using NT20 displayed negligible relationships with the respondent characteristics, which also occurred when RTE calculated using NT10 was related to makeup session attendance status. Given the RTE values of NT20 were so close to one, the relationship between the respondent characteristic and the logit of RTE appeared practically significant, however, when the relationship was converted from the logit of RTE back to the RTE scale, the relationship appeared negligible. In summary, the results of the differential relationships displayed across threshold calculation methods with the respondent characteristics indicates support for the NT30, INPSECT, INSPECT2, and MIXTURE methods. Across respondent characteristics, RTE calculated using NT20 appeared to be too high to display a relationship of significant or practical magnitude.

Relationships with item characteristics.

***Item position*.** Examination of the relationship between item position and logit of RTF scores indicated the relationship was dependent on threshold calculation method. Interestingly, the relationship between item position and logit of RTF was positive for MIXTURE2. The direction of this relationship is opposite of the hypothesized direction and suggests the method may be misclassifying motivated respondents as rapid responders. In regards to the other threshold calculation methods, item position was negatively related to the logit of RTF when calculated using INSPECT, NT20, NT30, and RSPEED. Although the relationships were in the hypothesized direction, the magnitude

of these relationships were small and RSPEED appeared to be the only method that was practically different from the other methods, with the exception of MIXTURE2.

The relationships calculated using NT20, NT30, and RSPEED should be cautiously interpreted, given they are confounded with item length. That is, item position and item length were moderately correlated (r = 0.45), indicating as the serial position of an item increased (i.e., as respondents progressed through the test), the length of the items also increased. Examination of the four noncognitive measures combined to create the MFLS revealed the last measure, the Life Regard Index, had lengthier items then the other three noncognitive measures. Based on the how time thresholds are defined by RSPEED, longer items yield larger time thresholds. Similarly, because the time thresholds defined by the NT methods are a function of response time, longer items yield larger response times, which yields larger time thresholds and subsequently lower RTF scores. Future researchers who are interested in studying the relationship between item position and logit of RTF using the RSPEED and NT calculation methods should use (a) noncognitive measures with items of similar length or (b) noncognitive measures with items of various lengths dispersed evenly throughout the measure. Another option for future researchers is to use one of the distributional threshold calculation methods (with the exception of MIXTURE2) instead, as these methods calculate time thresholds differently than the NT methods and RSPEED.

***Item length*.** Results indicated the relationship between item length and logit of RTF was dependent on threshold calculation method. Interestingly, as previously observed with item position, when RTF was calculated using MIXTURE2, the results revealed a large positive relationship with item length. Given this relationship was in the

opposite of the hypothesized direction, there is evidence to suggest MIXTURE2 may be measuring something else or is misclassifying motivated respondents as rapidly responding due to the large time thresholds the method defined. In regards to the other methods, when calculated using NT20, NT30, INSPECT, MIXTURE, and RSPEED, item length was negatively related to logit of RTF as hypothesized. In addition, the only relationship that appeared to be practically different was using the RSPEED method; the differences between the other relationships calculated using NT20, NT30, INSPECT, and MIXTURE did not exhibit large practical differences. The large differential relationship displayed by RSPEED should be cautiously interpreted given the relationship is confounded by how the time thresholds are defined. Similarly, the relationships displayed by the NT methods should also be cautiously interpreted given the confounding relationship between item length and response time. Based on these considerations, future researchers interested in gathering validity evidence for the RSPEED or NT calculation methods should use item characteristics other than item position and item length.

In summary, when considered in conjunction with the results from Phase Two, the results of Phase Three provide external validity evidence for some of the threshold calculation methods. In particular, the results of the differential relationships between the respondent characteristics and the logit of RTE indicates support for using the NT30, INSPECT, INSPECT2, and MIXTURE threshold calculation methods. In contrast, differential relationships between the item characteristics and the logit of RTF scores indicated the NT methods and RSPEED should not be used when the relationships between item characteristics and RTF are of interest due to their confounding relationships with the item characteristics. Interestingly, both of the item characteristics

displayed positive relationships with logit of RTF when calculated using MIXTURE2. Because these findings were not in the direction hypothesized, they provide evidence indicating the time thresholds defined by the MIXTURE2 method might be misclassifying respondents. However, given this is the first study to use this threshold calculation method, further research is needed.

Integration of Results from Phases One, Two, and Three

Several patterns emerged when the results from the three phases were examined in conjunction. First, the time thresholds defined in the current study were shorter on average for items on the noncognitive measure used in the current study than typically seen with cognitive tests (e.g., Pastor et al., 2015; Wise et al., 2006; Wise et al., 2009). The majority of threshold calculation methods (with the exception of MIXTURE2 and RSPEED) yielded mean RTE scores that ranged from .94 to 1.00, on average. The magnitude of these RTE scores are similar to and slightly higher than RTE scores previously reported for low-stakes cognitive tests. This finding is in line with previous studies that have found students put forth (slightly) more effort on less cognitively demanding tests (i.e., noncognitive measures) than on cognitive tests (e.g., Barry, Horst, Finney, Brown, & Kopp, 2010; Barry & Finney, 2016). For example, when applied to low-stakes cognitive tests, the average RTE score reported by Wise et al. (2009) was .90, whereas the average RTE scores calculated using four different methods and reported by Kong et al. (2007) ranged from 0.93 to 0.95. Similarly, Wise and DeMars (2007) reported the average RTE scores for incoming freshmen students was 0.996 and for upperclassmen was 0.943.

Second, the time thresholds defined by the NT10 and NT20 calculation methods were very small relative to time thresholds defined by other calculation methods and subsequently classified the largest proportions of respondents as exhibiting solution behavior across items. On average, respondents exhibited solution behavior on 100% of the items when calculated using NT10 (*SD* = 0.01) and 99% of the items when calculated using NT20 (*SD* = 0.04). The high proportions of respondents classified by these methods is noteworthy given respondents' motivation in the current study was expected to be low, which was anticipated for two reasons. First, it was anticipated respondents would put forth low effort given the majority of students from the Primary sample completed the battery of low-stakes assessments for accountability purposes during makeup testing sessions. Previous research has shown students completing low-stakes tests during makeup testing sessions tend to put forth less effort than do students attending the regularly scheduled testing session (e.g., Swerdzewski et al., 2009). Second, it was also anticipated respondents possessing low levels of the construct meaningful life would not exhibit high amounts of effort in responding to items on the MFLS. That is, asking respondents with low levels of meaningful life to answer multiple items such as "I really don't have much purpose for living, even for myself" and "I just don't know what I really want to do with my life" may result in lower motivation levels due to the depressing nature of the questions (see Appendix B for items).

Based on these reasons, it is unlikely respondents displayed enough effort to be classified as exhibiting solution behavior on the majority of items although the descriptive statistics of the RTE scores calculated using these methods suggest otherwise. This observation is supported by examining the time thresholds defined by NT10 in

relation to the response times for the Known Rapid Responders sample. In particular, 52 of the time thresholds defined by NT10 were less than the minimum response time to items completed by the Known Rapid Responders sample, which indicates the time thresholds are too small. Although the NT10 and NT20 calculation methods have successfully been used with items on low-stakes cognitive tests (e.g., Wise & Ma, 2012), these threshold calculation methods might not be useful in distinguishing solution behavior responses from rapid responses on a low-stakes noncognitive measure when motivation is anticipated to be low. It is important to remember that the NT methods will always define a time threshold for an item, regardless of whether respondents are exhibiting rapid-responding behavior or not. Based on these considerations and given this is the only study known of to apply the NT methods to a low-stakes noncognitive measure, further research is needed.

Third, the time thresholds defined by the MIXTURE2 and RSPEED threshold calculation methods were larger on average than the other time thresholds and subsequently classified the smallest proportions of respondents as exhibiting solution behavior across items. Thus, in contrast to the NT methods that likely classified too many respondents as exhibiting solution behavior, the MIXTURE2 and RSPEED methods may have classified too *few* respondents as exhibiting solution behavior. When related to respondent characteristics, the relationships displayed between logit of RTE calculated using RSPEED and MIXTURE2 were weaker and practically lower than the other relationships, which suggests the two groups of respondents were not as distinct from one another. Interestingly, when related to the item characteristics, MIXTURE2 displayed positive relationships with item position and length, which is contrary to what was

hypothesized. Although the majority of the data for the Primary sample did come from students completing the tests during makeup testing sessions, the makeup testing sessions were conducted in controlled, proctored environments, which has been shown by previous research to increase motivation (e.g., Lau, Swerdzewski, Jones, Anderson, & Markle, 2009). Given the purpose of the SB index is to distinguish respondents who could be putting forth effort in responding from those who are most assuredly not putting forth any effort in responding (i.e., the worst of the worst respondents in regards to motivation), practitioners have been recommend to err on the side of caution to prevent misclassifying a respondent who was a fast reader as one who rapidly responded. Based on these considerations, and given this is the first time these methods have been applied to noncognitive measures, further research is needed.

Fourth, the results of the current study found support for using NT30, INSPECT, INSPECT2, and MIXTURE threshold calculation methods. The proportion of respondents classified by the time thresholds defined by these methods did not practically differ from one another on the majority of items nor did they practically differ from one another in their relationships to respondent and item characteristics. In particular, INSPECT2 and MIXTURE displayed nearly-equivalent relationships with the external characteristics, which suggests researchers can use either method to calculate the thresholds and end up with identical results. The equivalency between the two methods is beneficial for those practitioners who either (a) do not have advanced statistical skills or (b) do not have the ability to collect an additional group of known rapid responders, either due to financial or resources constraints.

However, given the similarities of the results using NT30, practitioners without advanced statistical skills or the time or resources to collect an additional group of known rapid responders can calculate time thresholds using only the raw response time data. Based on the results of the current study, it is unclear if collecting a known group of rapid responders is beneficial, especially if similar results can be obtained using a different and less complicated method. However, having a known group of rapid responders might be beneficial since the data can be used in a multitude of ways. For example, having a known group of rapid responders would provide practitioners with a baseline rate of how fast respondents can respond to the substantive measure of interest. In turn, this information can be useful in gauging the validity of threshold calculation methods. Although the time thresholds defined by NT30, INSPECT, INSPECT2, and MIXTURE appear to be the most promising methods for identifying solution behavior respondents on low-stakes noncognitive measures, further research examining these calculation methods is needed.

Fifth, the current study found mixed evidence for including a known group of rapid responders during the calculation of the time thresholds. In particular, the inclusion of the known group of rapid responders appeared to have a different impact on the time thresholds defined using INSPECT2 than it did on the time thresholds defined using MIXTURE2. The average time threshold defined using INSPECT2 was slightly higher (1/3 of a second) than the average time threshold defined using INSPECT and both methods were related to the respondent and item characteristics in similar ways. In contrast, the average time threshold defined by MIXTURE2 was almost a full second higher than the average time threshold defined by MIXTURE and both methods were not

related to the respondent and item characteristics in similar ways. Although the inclusion of the Known Rapid Responders appeared to make the bimodal distribution of the items' response time distributions more distinct when reviewed using INSPECT2, in hindsight, the additional group may have made the two groups appear less distinct from a mixture modeling point of view, especially in consideration of the problems encountered collecting the data (discussed below). That is, the inclusion of rapid responders with large response times may have had an impact. Overall the results of the study suggest MIXTURE2 may have erroneously classified motivated respondents as rapidly responding instead of exhibiting solution behavior. However, further research is needed.

Given this was the first study to examine if the SB index could be meaningfully used with a noncognitive measure, more validity studies using different noncognitive measures and samples are needed. Based on the results of the current study, future studies should continue to examine the effectiveness of using the more promising threshold calculation methods found in the current study, which includes NT30, INSPECT, INSPECT2, and MIXTURE. In addition, future research should examine the effect of including a known group of rapid responders and using them to define time thresholds.

Practical Implications

Assuming the NT30, INSPECT, INSPECT2, and MIXTURE threshold calculation methods are measuring respondent effort and validly distinguishing respondents exhibiting solution behavior from respondents exhibiting rapid responding behavior, the results have several practical implications for practitioners. For example, practitioners can use the SB index to examine how much effort respondents put forth in completing a measure and then use the results to develop target interventions to improve

motivation. Related, practitioners can use the SB index with predetermined time thresholds to monitor respondents' effort while completing a sequence of low-stakes noncognitive measures. If respondents start to display rapid-guessing behavior, then a warning message can be issued to respondents reminding them to put forth effort. This approach is based on a model employed by Wise and his colleagues (2006), whereby students' effort was monitored during a low-stakes cognitive testing session.

Practitioners can also use the SB index with noncognitive measures to study how respondent effort changes during the duration of a testing session. An inherent advantage of the SB index is its ability to measure respondent effort at the item level, thereby affording researchers with information about respondents' effort during the entirety of the testing session and at each moment in time (e.g., Pastor et al., 2015; Strickman et al., 2015). Similarly, the SB index would allow practitioners to see what responses are chosen by respondents displaying rapid-responding. This information would be very useful to help distinguish those who are rapidly responding from those who are fast readers, and would help practitioners determine if there is a need to worry about taking into account respondent effort. Finally, another way practitioners can use the SB index with noncognitive measures is to study respondent and item characteristics that are related to effort with the goal of finding characteristics that will aid in increasing motivation.

Although the application of the SB index to noncognitive measures has several practical applications, practitioners are cautioned if the SB index will be used with low-stakes noncognitive measures for the purpose of motivation filtering. Specifically, motivation filtering is the process of identifying and filtering out unmotivated students

displaying effort below a predetermined threshold. Although this practice is commonly conducted using the SB index with low-stakes cognitive tests (e.g., Kong et al., 2007; Wise & Kong, 2005), it has only been conducted once using effort scores based on the SB index and with noncognitive measures (Swerdzewski et al., 2011). An implicit assumption made by practitioners who use motivation filtering with cognitive tests is students' effort is unrelated to independent measures of academic ability. By extension, using motivation filtering with noncognitive measures seems to imply it would only be appropriate if respondents' effort was not related to the construct being measured. Thus, if practitioners plan on using the SB index with noncognitive measures to conduct motivation filtering, they need to be aware if the construct they are measuring is related to effort then filtering out respondents with low motivation might systematically bias the results. Given the wide array of noncognitive constructs, this assumption seems unlikely to hold. Thus, further research examining the application of the SB index to noncognitive measures should be conducted.

Limitations and Future Research

It is important to discuss the limitations of the current study as they inform future research. One of the limitations of the current study is the results, particularly in regards to the relationships between item characteristics and RTF scores for certain methods, are confounded. Given these methods explicitly use response times in the calculation of the time thresholds and given lengthier items will increase reading time, these relationships are confounded. Future researchers that are interested in examining the relationship between item characteristics and RTF scores should plan in advance of how to account for these potential confounding effects.

A second limitation of the current study is related to the noncognitive measure used. Specifically, four separate measures assessing the construct of meaningful life were administered as a set and used as the substantive measure of interest. Given this was the only noncognitive measure used in the current study, future researchers should examine if using a noncognitive measure assessing a different construct or a collection of noncognitive measures assessing different constructs with different response scales has an impact on the time thresholds defined by the threshold calculation methods and subsequent SB classification indices.

A third limitation of the study is in regards to the sample of participants used. Data for the Primary sample came from a homogenous group of college students. Moreover, a large proportion of these students were expected to exhibit low effort on the MFLS because they attended makeup testing sessions. Previous research has shown that respondents attending the makeup-testing sessions put forth less effort than respondents who attend the required campus-wide assessment day (Swerdzewski et al., 2009). As a result, it is unclear if these results will generalize to other populations of students who are more heterogeneous and motivated.

Similarly, the sample of respondents collected for the Known Rapid Responders was different than the respondents used in the Primary sample. Although demographic data were not collected for the Known Rapid Responders sample, 40% of the respondents came from the undergraduate psychology participant pool, 28% came from makeup-testing sessions, and 31% came from friends and acquaintances of the researcher. Given the differences between samples, it is unclear if these results will generalize when other populations are used. Future researchers should examine if the results of the current study

can be replicated when respondents for both the Primary sample and Known Rapid Responders sample come from the same population. In addition, future researchers should also examine if the time threshold calculation methods can be used to calculate SB indices with different samples of respondents and if the findings from the current study generalize to other samples.

A fourth limitation of the current study was due to the problems encountered collecting data for the Known Rapid Responders sample. Recall, due to issues collecting the Known Rapid Responders sample, the majority of the data for the sample was gathered online in uncontrolled conditions. Consequently, not all of the participants adhered to the instructions to rapidly respond, which was indicated by the presence of large response times in the data. Although records with response times greater than fifteen seconds were removed from the data, the response times for the Known Rapid Responders may have been too large on average to truly represent rapid responding. As a result, respondents classified in the rapid-guessing class by MIXTURE2 may have included some respondents who were exhibiting solution behavior and some who were rapidly responding. Moreover, this may have had an impact on the thresholds that were not defined for some of the items. Future studies that examine the impact of including a known group of rapid responders should investigate if collecting the data in a controlled environment or restricting the response times of the known group of responders to a faster response time (e.g., 8 seconds) would impact the resulting time thresholds. Similarly, given the results of the current study essentially found no difference between INSPECT2 and MIXTURE, would using data from a known group of rapid responders

collected under controlled conditions impact the similarity of the results seen using INSPECT2 and MIXTURE?

A fifth limitation of the current study was the software used to estimate the lognormal mixture models, MIXTURE and MIXTURE2. Specifically, in comparison to other software programs, SAS 9.4 provides a limited amount of information about the fit of each model. In particular, although PROC FMM provides the AIC and BIC fit indices (the SSABIC fit indices can be easily calculated by hand), other fit indices such as the Lo-Mendall-Rubin likelihood ratio test (Lo, Mendall, & Rubin, 2001) and measures of classification accuracy such as posterior probabilities and entropy are not provided. Thus, decisions about the fit of the models estimated using MIXTURE and MIXTURE2 were made on limited information. Future researchers should examine the effect of the MIXTURE and MIXTURE2 threshold calculation methods using other software programs that are both capable of estimating lognormal mixture models and that provide more fit information about the resulting models.

A sixth limitation of the study is in regards to the generalizability of the results due to how the items were administered. Specifically, because response time information was collected at the item level, the items were administered individually per page, which is not how items on noncognitive measures are typically administered. Instead, items on noncognitive measures answered using a Likert response scale are typically administered as a set on one page. Administering the items individually per page may have slowed down the respondents and may impact the generalizability of the results.

Conclusions

One of the main purposes of the current study was to determine if threshold calculation methods commonly used to define the SB index on cognitive tests could be successfully used with noncognitive measures. The results of the current study indicate the majority of these threshold calculation methods can be successfully used with noncognitive measures. Out of the eight threshold calculation methods examined, the following four show promise: NT30, INSPECT, INSPECT2, and MIXTURE. Although the current study contributes to the low-stakes testing literature by examining if the SB index can be used with noncognitive measures, further research is needed. Specifically, researchers should replicate the current study to see if the results generalize to other populations, contexts, and noncognitive measures. Ultimately, the goal of using the SB index is to classify respondents according to how much effort they put forth in responding to an item. Classifying a response as a solution behavior response is ambiguous because it only indicates the respondent did not rapidly respond to that item. That is, there is no way to confirm a respondent truly put forth effort in responding to an item or not. Thus, it is imperative that researchers continue to gather external validity evidence for the resulting time thresholds to examine if the respondents were classified into meaningful and distinct groups. Although the SB index is a very versatile tool with several uses, a substantial amount of research needs to be conducted before routinely using it with noncognitive measures.

Footnotes

[1] As previously described, after estimating the two-class lognormal mixture models, Schnipke and Scrams (1997) visually inspected the resulting model-implied distributions to identify the solution behavior thresholds for test items. In contrast, Pastor and her colleagues (2015) used the resulting models' posterior probabilities to classify examinees as exhibiting either solution behavior or rapid-guessing behavior on test items. See Pastor et al. (2015) and Yang (2007) for more details.

[2] Researchers have also used another method in conjunction with the NT method to identify the solution behavior time thresholds for items administered via a computer adaptive test (CAT). Specifically, in addition to considering items' response time, items' response accuracy is also considered. Given these methods are more complex and specific to only CAT items, and given the focus of the current study will not use CAT items, readers interested in more information about this alternative threshold calculation method are referred to Lee and Jia (2014), Ma, Wise, Thum, and Kingsbury (2011), and Wise, Ma, and Theaker (2012) for more details.

[3] Although other studies have compared the validity of using self-reported effort scores to RTE scores for the purpose of identifying and studying low-motivated examinees completing low-stakes cognitive tests (e.g., Rios et al., 2014), Swerdzewski et al. (2011) is the only known to study to compare the methods using cognitive *and* noncognitive tests.

[4] The focus of the current analysis was on comparing the group of students who attended the originally scheduled makeup testing sessions to the group of students who failed to attend the regularly scheduled makeup testing sessions (i.e., walk-ins). This

analysis did not compare the group of students who attended Assessment Day to the group of students who did not attend Assessment Day (i.e., all of the makeups) for two reasons. First, during the planning stages of the study, the size of the Assessment Day sample was expected to be low ($n \sim 60$). Second, previous research has only examined the difference between students who attend Assessment Day and students who did not (i.e., all makeups; e.g., Swerdzewski et al., 2009). Thus, the current study sought to contribute to the literature by comparing this subpopulation of students who attended the regularly scheduled makeup testing session to those students who did not attend the regularly scheduled makeup testing session.

[5] Technically, the SOS2 questionnaire was administered to samples in the current study. Specifically, the SOS2 is a 30-item measure that contains three measures; the first ten items are from the SOS scale and the remaining 20 items on the questionnaire measure text anxiety and expectancy-value-cost theory relative to general education coursework. Data from the latter two scales were not used in the current study.

[6] Other fit indices commonly examined including the Lo-Mendell-Rubin test (Lo, Mendell, Rubin, 2001) and the bootstrap likelihood ratio test (e.g., Nylund, Asparouhov, & Muthén, 2007) were not used in the current study due to limitations of the SAS 9.4 program, which does not calculate them.

[7] Data are missing because respondents went so fast they failed to provide a response and the software feature forcing respondents to provide a response, which was used with the primary sample, was accidentally turned off at the beginning stages of data collection. Long item responses were due to a subset of respondents who failed to follow the directions.

[8] The mixing proportions for the two-class solution using the lognormal mixture threshold calculation method (i.e., MIXTURE) are presented in Appendix C. The mixing proportions for the two-class solution using the lognormal mixture with information threshold calculation method (i.e., MIXTURE2) are presented in Appendix D.

[9] The proportions presented in Table 13 are Response Time Fidelity (RTF) scores, which were calculated using Equation 3.

Appendix A

Instructions for Known Rapid Responders Sample

Please answer the series of questions as quickly as possible **without** actually reading the items. Do **NOT** read the items and do **NOT** think about how you responded to the previous items or how you think you should honestly respond.

The goal is to provide a response as quickly as possible and to move onto the next item.

You may be wondering why we are asking you to do this. We are trying to determine how long it takes students to complete an assessment when they don't even read the items and just rapidly respond. The series of items ask questions about how meaningful your life is.

Appendix B

Meaningful Life Questionnaire

Item	Scale	Subscale
1. I understand my life's meaning.	MLQ	Presence
2. I am looking for something that makes my life feel meaningful.	MLQ	Search
3. I am always looking to find my life's purpose.	MLQ	Search
4. My life has a clear sense of purpose.	MLQ	Presence
5. I have a good sense of what makes my life meaningful.	MLQ	Presence
6. I have discovered a satisfying life purpose.	MLQ	Presence
7. I am always searching for something that makes my life feel significant.	MLQ	Search
8. I am seeking a purpose or mission for my life.	MLQ	Search
9. My life has no clear purpose. (R)	MLQ	Presence
10. I am searching for meaning in my life.	MLQ	Search
11. I lead a fulfilled life.	SoME	Meaningfulness
12. I think that there is meaning in what I do.	SoME	Meaningfulness
13. I have a task in life.	SoME	Meaningfulness
14. I feel part of a bigger whole.	SoME	Meaningfulness
15. I think my life has a deeper meaning.	SoME	Meaningfulness
16. I expect to find a meaningful career.	WAMI-R	Positive meaning
17. I view my future work as contributing to my personal growth.	WAMI-R	Meaning making through work
18. My future work will make no difference in the world. (R)	WAMI-R	Greater good motivation
19. I expect that my future work will contribute to my life's meaning.	WAMI-R	Positive meaning
20. I have a good sense of what will make my future job meaningful.	WAMI-R	Positive meaning
21. I know my future work will make a positive difference in the world.	WAMI-R	Greater good motivation

22. My future work will help me better understand myself.	WAMI-R	Meaning making through work
23. I expect that my work in the future will have a satisfying purpose.	WAMI-R	Positive meaning
24. My future work will help me make sense of the world around me.	WAMI-R	Meaning making through work
25. My future work will serve a greater purpose.	WAMI-R	Greater good motivation
26. I feel like I have found a really significant meaning for leading my life.	LRI	Framework
27. Living is deeply fulfilling.	LRI	Fulfillment
28. I really don't have much purpose for living, even for myself. (R)	LRI	Framework
29. There honestly isn't anything that I totally want to do. (R)	LRI	Framework
30. I really feel good about my life.	LRI	Fulfillment
31. I spend most of my time doing things that really aren't important to me. (R)	LRI	Fulfillment
32. I have really come to terms with what's important for me in my life.	LRI	Framework
33. I need to find something that I can really be committed to. (R)	LRI	Framework
34. I just don't know what I really want to do with my life. (R)	LRI	Framework
35. Other people seem to have a better idea of what they want to do with their lives than I do. (R)	LRI	Framework
36. I have some aims and goals that would personally give me a great deal of satisfaction, if I could accomplish them.	LRI	Framework
37. I don't seem to be able to accomplish those things that are really important to me. (R)	LRI	Fulfillment
38. I really don’t believe very deeply about anything in my life. (R)	LRI	Framework
39. I have a philosophy of life that really gives my living significance.	LRI	Framework
40. Other people seem to feel better about their lives than I do. (R)	LRI	Fulfillment
41. I get completely confused when I try to understand my life. (R)	LRI	Framework
42. Something seems to stop me from doing what I really want to do. (R)	LRI	Fulfillment
43. I have a lot of potential that I don't normally use. (R)	LRI	Fulfillment

44. When I look at my life I feel the satisfaction of really having worked to accomplish something.	LRI	Fulfillment
45. I have real passion in my life.	LRI	Fulfillment
46. I feel that I'm really going to attain what I want in life.	LRI	Fulfillment
47. I don't really value what I'm doing. (R)	LRI	Fulfillment
48. I have a very clear idea of what I'd like to do with my life.	LRI	Framework
49. I get so excited by what I'm doing that I find new stores of energy that I didn't know I had.	LRI	Fulfillment
50. There are things that I devote all my life's energy to.	LRI	Framework
51. Nothing outstanding ever seems to happen to me. (R)	LRI	Fulfillment
52. I feel that I am living fully.	LRI	Fulfillment
53. I have a system or framework that allows me to truly understand being alive.	LRI	Framework

Note. MLQ = Meaning in Life Questionnaire; SoME = Sources of Meaning and Meaning in Life Questionnaire; WAMI-R = Work and Meaning Inventory-Revised; LRI = Life Regard Index; (R) = reverse-scored item.

Appendix C

Table C1. Mixing proportions for the two-class solution calculated using the lognormal mixture model threshold calculation method (MIXTURE)

Item	Class 1	Class 2	Item	Class 1	Class 2
1	0.202	0.798	28	0.178	0.822
2	0.098	0.902	29	0.147	0.853
3	0.149	0.851	30	0.075	0.925
4	0.161	0.839	31	0.190	0.810
5	0.224	0.776	32	0.171	0.829
6	0.261	0.739	33	0.169	0.831
7	0.179	0.821	34	0.206	0.794
8	0.546	0.454	35	0.228	0.772
9	0.075	0.925	36	0.237	0.763
10	0.182	0.818	37	0.247	0.753
11	0.411	0.589	38	0.208	0.792
12	0.161	0.839	39	0.166	0.834
13	---	---	40	0.168	0.832
14	0.375	0.625	41	0.127	0.873
15	0.552	0.448	42	0.141	0.859
16	0.267	0.733	43	0.206	0.794
17	0.174	0.826	44	0.298	0.702
18	0.185	0.815	45	0.188	0.812
19	0.173	0.827	46	0.194	0.806
20	0.412	0.588	47	0.235	0.765
21	0.172	0.828	48	0.192	0.808
22	0.302	0.698	49	0.225	0.775
23	0.311	0.689	50	0.187	0.813
24	0.426	0.574	51	0.152	0.848
25	0.266	0.734	52	0.166	0.834
26	0.295	0.705	53	0.353	0.647
27	0.211	0.789			

Note. Class 1 = Rapid responding class. Class 2 = Solution Behavior class. Dashed lines indicate a time threshold was not defined for that item.

Appendix D

Table D1. Mixing proportions for the two-class solution calculated using the lognormal mixture model with information threshold calculation method (MIXTURE2)

Item	Class 1	Class 2	Item	Class 1	Class 2
1	--	--	28	0.453	0.547
2	0.511	0.489	29	0.440	0.560
3	0.598	0.402	30	0.453	0.547
4	0.561	0.439	31	0.432	0.568
5	0.585	0.415	32	0.501	0.499
6	0.608	0.392	33	0.520	0.480
7	0.464	0.536	34	0.495	0.505
8	--	--	35	0.487	0.513
9	0.136	0.864	36	0.475	0.525
10	0.600	0.400	37	0.519	0.481
11	0.612	0.388	38	0.470	0.530
12	0.488	0.512	39	0.490	0.510
13	0.585	0.415	40	0.459	0.541
14	--	--	41	0.450	0.550
15	0.667	0.333	42	0.374	0.626
16	0.530	0.470	43	0.521	0.479
17	0.547	0.453	44	0.522	0.478
18	0.453	0.547	45	0.703	0.297
19	0.525	0.475	46	0.459	0.541
20	0.595	0.405	47	0.508	0.492
21	0.584	0.416	48	0.486	0.514
22	0.686	0.314	49	0.248	0.752
23	0.694	0.306	50	0.471	0.529
24	0.669	0.331	51	0.477	0.523
25	0.703	0.297	52	0.519	0.481
26	0.549	0.451	53	0.614	0.386
27	0.717	0.283			

Note. Class 1 = Rapid responding class. Class 2 = Solution Behavior class. Dashed lines indicate a time threshold was not defined for that item.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment, 68*, 139–151. http://doi.org/10.1207/s15327752jpa6801_11

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods, 7*(2), 127-150.

Barry, C. L., & Finney, S. J. (2009). Does it matter how data are collected? A comparison of testing conditions and the implications for validity. *Research and Practice in Assessment, 3*(1), 1-15.

Barry, C. L., & Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education, 29*(1), 46-64.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. B., & Kopp, A. B. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*, 342-363.

Battista, J., & Almond, R. (1973). The development of meaning in life. *Psychiatry, 36*(4), 409–427. http://doi.org/10.1177/0040571X7908200403

Bovaird, J. A. (2002). *New applications in testing: Using response time to increase the construct validity of a latent trait estimate* (Unpublished doctoral dissertation). University of Kansas.

Brown, A. R., & Finney, S. J. (2011). Low-stakes testing and psychological reactance: Using the Hong Psychological Reactance scale to better understand compliant and

non-compliant examinees. *International Journal of Testing, 11*(3), 248–270. http://doi.org/10.1080/15305058.2011.570884

Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modeling. *Statistics in Medicine, 17*, 1261-1291.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika, 73*(2), 209–230. http://doi.org/10.1007/s11336-007-9045-9

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*(2), 84–95.

Conway, J. M. (2002). Method variance and method bias in I/O psychology. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial/organizational psychology* (pp. 344–365). Oxford: Blackwell Publishers.

Curran, P. G. (2015). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*. Advance online publication. http://doi.org/10.1016/j.jesp.2015.07.006

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*(1), 23–45. http://doi.org/10.1207/s15326977ea1201_2

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

DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing, 10*(3), 207–229. http://doi.org/10.1080/15305058.2010.496347

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36*, 171–181.

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

Enders, C. K., & Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling, 15*(1), 75–95. http://doi.org/10.1080/10705510701758281

Finn, B. (2015). Measuring motivation in low-stakes assessments. ETS Research Report Series. Retrieved from http://doi.wiley.com/10.1002/ets2.12067

Giffi, C., Dollar, B., Drew, M., McNelly, J., Carrick, G., & Gangula, B. (2015). The skills gap in U.S. manufacturing 2015 and beyond. Retrieved from http://www2.deloitte.com/content/dam/Deloitte/us/Documents/manufacturing/us-pip-the-manufacturing-institute-and-deloitte-skills-gap-in-manufacturing-study.pdf

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *Journal of Experimental Education, 74*(3), 251-266.

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

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*(1), 99–114. http://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*(3), 828–845.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods, 18*, 512–541. http://doi.org/10.1177/1094428115571894

Kirsch, I., Braun, H., Yamamoto, K., & Sum, A. (2007). America's perfect storm: Three forces changing our nation's future. Educational Testing Service. Retrieved from http://www.ets.org/Media/Education_Topics/pdf/AmericasPerfectStorm.pdf

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. http://doi.org/10.1177/0013164406294779

Kopp, J. P., & Finney, S. J. (2013). Linking academic entitlement and student incivility using latent means modeling. *The Journal of Experimental Education, 81*(3), 322. http://doi.org/10.1080/00220973.2012.727887

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014). Knowing what students know and can do: The current state of student learning outcomes assessment in U.S. colleges and universities. Retrieved from http://www.learningoutcomeassessment.org/documents/2013 Abridged Survey Report Final.pdf

Kyllonen, P. C. (2005). The case for noncognitive assessments. ETS: R&D Connections.

Kyllonen, P. C. (2013). Soft skills for the workplace. *Change: The Magazine of Higher Learning*, *45*(6), 16–23. http://doi.org/10.1080/00091383.2013.841516

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *Journal of General Education, 58*(3), 196-217.

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

Liu, M., Bowling, N. A., Huang, J. L., & Kent, T. A. (2013). Insufficient effort responding to surveys as a threat to validity: The perceptions and practices of SIOP members. *The Industrial-Organizational Psychologist, 51*(1), 32–39.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. Educational Testing Service, Princeton, NJ.

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. http://doi.org/10.1080/10627197.2015.1028618

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*, 767-778.

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

MacKenzie, S. B., & Podsakoff, P. M. (2012). Common method bias in marketing: Causes, mechanisms, and procedural remedies. *Journal of Retailing, 88*(4), 542–555. http://doi.org/10.1016/j.jretai.2012.08.001

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83. http://doi.org/10.1016/j.jrp.2013.09.008

Markle, R., Brenneman, M., Jackson, T., Burrus, J., & Robbins, S. (2013). Synthesizing frameworks of higher education student learning outcomes. ETS Research Report Series. Retrieved from http://www.ets.org/Media/Research/pdf/RR-13-22.pdf

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York: Wiley-Interscience. http://doi.org/10.1198/tech.2002.s651

McNeish, D., & Stapleton, L. (in-press). Modeling clustered data with very few clusters. *Multivariate Behavioral Research.*

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. http://doi.org/10.1037/a0028085

Naemi, B., Burrus, J., Kyllonen, P. C., & Roberts, R. D. (2012). Building a case to develop noncognitive assessment products and services targeting workforce readiness at ETS. Retrieved from https://www.ets.org/s/workforce_readiness/pdf/rm_12_23.pdf

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14*(4), 535-569.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. San Francisco, CA: Josey-Bass.

Pastor, D. A., Strickman, S. N., & Ong, T. Q. (2015). *Patterns of solution behavior across items in low-stakes assessments.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Robles, M. M. (2012). Executive perceptions of the top 10 soft skills needed in today's workplace. *Business Communication Quarterly*, *75*(4), 453–465. http://doi.org/10.1177/1080569912460400

SAS Institute [Computer software]. (2002-2015). Cary, NC.

Schnell, T. (2009). The Sources of Meaning and Meaning in Life Questionnaire (SoMe): Relations to demographics and well-being. *The Journal of Positive Psychology*, *4*(6), 483–499. http://doi.org/10.1080/17439760903271074

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.

Schuh, J. H., & Gansemer-Topf, A. M. (2010). The role of student affairs in student learning assessment (Occasional Paper No. 7). *National Institute for Learning Outcomes Assessment*, 1–18. Retrieved from http://www.learningoutcomeassessment.org/documents/studentAffairsrole.pdf

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Sclove, L. S. (1987). Application of model selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–343.

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. http://doi.org/10.1080/08957347.2013.739453

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angelos, CA: Sage.

Society for Human Resource Management. (2015). SHRM research: Workforce readiness and skills shortages. Retrieved from http://www.shrm.org/Research/FutureWorkplaceTrends/Documents/Workforce Readiness and Skills Shortages.pdf

Steger, M. F., Dik, B. J., & Duffy, R. D. (2012). Measuring meaningful work: The Work and Meaning Inventory (WAMI). *Journal of Career Assessment, 20*(3), 322–337. http://doi.org/10.1177/1069072711436160

Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology, 53*(1), 80–93. http://doi.org/10.1037/0022-0167.53.1.80

Strickman, S. N., Pastor, D. A., & Ong, T. Q. (2015). *Evaluating patterns of solution behavior at the item level in low-stakes assessments*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Trumball, CT.

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14*(1), 8–9.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2009). Skipping the test: Using empirical evidence to inform policy related to students who avoid taking low-stakes assessments in college. *The Journal of General Education, 58*(3), 167–195. http://doi.org/10.1353/jge.0.0043

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

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *The Journal of General Education, 58*(3), 129–151. http://doi.org/10.1353/jge.0.0047

Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in growth mixture modeling. In G. R. Hancock (Ed.), *Mixture models in latent variable research* (pp. 317–341). Greenwich, CT.

Torney-Purta, J., Cabrera, J. C., Crotts Roohr, K., Liu, O. L., & Rios, J. A. (2015). Assessing civic competency and engagement in higher education: Research background, frameworks, and directions for next-generation assessment. ETS Research Report Series (Vol. RR-15–34).

U.S. Department of Education. (2006). A test of leadership: Charting the future of U.S. higher education. Washington, DC. Retrieved from http://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/pre-pub-report.pdf

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*(2), 95–114. http://doi.org/DOI 10.1207/s15324818ame1902_2

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237–252. http://doi.org/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. http://doi.org/10.1111/j.1745-3992.2006.00054.x

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

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38. http://doi.org/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 α: A note on Attali's "Reliability of speeded number-right multiple-choice tests." *Applied Psychological Measurement, 33*(6), 488-490.

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

Wise, S. L., & Kingsbury, G. G. (2015). *Modeling student test-taking motivation in the context of an adaptive achievement test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

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

Wise, S. L., Ma, L., & Theaker, R. A. (2012). *Identifying non-effortful student behavior on adaptive tests: Implications for test fraud detection*. Paper presented at the Statistical Detection of Potential Test Fraud Conference, Lawrence, KS.

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. http://doi.org/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. http://doi.org/10.1037/12330-009

Yang, X. (2007). Methods of identifying individual guessers from item response data. *Educational and Psychological Measurement, 67*(5), 745–764. Retrieved from http://epm.sagepub.com/content/67/5/745.short

Zhang, C., & Conrad, F. G. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods, 8*(2), 127–135. http://doi.org/10.18148/srm/2014.v8i2.5453

Table 1

Testing Configurations and Total Amount of Time Allotted, by Sample and Test

Sample	Sample Size	Test 1	Test 2	Test 3	Test 4
Primary sample - Makeup 2015[a]	336	ISNWA1 (50 min)	**MFLS (30 min)**	NONCOG55 (10 min)	SOS2 (5 min)
Primary sample - Assessment Day 2016[a]	77	ERWRA (60 min)	OCP2 (30 min)	**MFLS (15 min)**	SOS2 (5 min)
Primary sample - Makeup 2016[a]	158	INFOCORE (30 min)	**MFLS (20 min)**	SDA7 (30 min)	SOS2 (5 min)
Known Rapid Responders	181	**MFLS (30 min)**			

Note. The substantive scale of interest, the MFLS, is bolded. The MFLS was administered using various total testing times according to how it was administered. That is, in the Makeup 2015 sample, the MFLS was administered with an additional open-ended question added at the end. During the Makeup 2016 testing session, the MFLS was administered with an additional 33 items added (which will not be used as part of the current study).

MFLS = Meaningful Life Scale; ERWRA = Ethical Reasoning Writing Assessment; OCP2 = Oral Communications Pretest 2; SOS2 = Student Opinion Survey 2; ISNWA1 = Stewardship of the Natural World Assessment; INFOCORE = Information Literacy Core; NONCOG55 = Noncognitive Assessment 55; SDA7 = Sociocultural Dimension Assessment 7.

[a] The first three samples (Makeup 2015, Assessment Day, and Makeup 2016) were combined to create the Primary sample.

Table 2

Methods Used to Define Solution Behavior Time Thresholds

Acronym	Description	Samples Used
INSPECT	Visual inspection	Primary Sample[a]
INSPECT2[b]	Visual inspection with information	Primary Sample and Known Rapid Responders Sample
MIXTURE	Lognormal mixture modeling	Primary Sample
MIXTURE2[b]	Lognormal mixture modeling with information	Primary Sample and Known Rapid Responders Sample
NT10	Normative Threshold 10	Primary Sample
NT20	Normative Threshold 20	Primary Sample
NT30	Normative Threshold 30	Primary Sample
RSPEED	Reading speed (300ms/word)	Primary Sample

[a] The Primary sample is a combination of the following three samples: Makeup 2015, Assessment Day 2016, and Makeup 2016.

[b] The Primary sample and Known Rapid Responders samples were combined and used to calculate the thresholds using the visual inspection with information method and the lognormal mixture modeling with information method.

Table 3

Semantic Synonym Item Pairs

Pair	Item	Scale	Subscale
1)	1. I understand my life's meaning.	MLQ	Presence
	5. I have a good sense of what makes my life meaningful.	MLQ	Presence
2)	2. I am looking for something that makes my life feel meaningful.	MLQ	Search
	10. I am searching for meaning in my life.	MLQ	Search
3)	4. My life has a clear sense of purpose.	MLQ	Presence
	9. My life has no clear purpose. (R)	MLQ	Presence
4)	18. My future work will make no difference in the world. (R)	WAMI-R	Greater good motivation
	21. I know my future work will make a positive difference in the world.	WAMI-R	Greater good motivation
5)	27. Living is deeply fulfilling.	LRI	Fulfillment
	52. I feel that I am living fully.	LRI	Fulfillment
6)	33. I need to find something that I can really be committed to. (R)	LRI	Framework
	50. There are things that I devote all my life's energy to.	LRI	Framework
7)	34. I just don't know what I really want to do with my life. (R)	LRI	Framework
	48. I have a very clear idea of what I'd like to do with my life.	LRI	Framework

Note. (R) = reverse-scored item.

Table 4

Descriptive Statistics of the MFLS Items' Response Time Distributions for the Primary Sample (N = 568)

Item	Mean	*SD*	Median	Min.	Max.	Skew	Kurtosis
1	9.99	8.23	7.01	0.54	85.16	4.13	31.43
2	6.12	5.54	3.85	0.56	71.89	9.40	151.11
3	4.64	3.94	2.59	0.59	36.41	4.48	42.95
4	4.28	3.94	1.77	0.75	15.02	1.54	4.69
5	4.59	4.20	2.45	0.40	35.42	4.73	46.81
6	4.56	4.01	2.51	0.75	30.97	3.71	27.24
7	5.08	4.73	2.61	0.22	37.82	4.75	49.33
8	4.77	4.30	2.58	0.59	27.33	2.43	12.64
9	4.86	4.40	2.23	0.61	18.79	1.39	3.75
10	4.56	4.06	2.48	0.59	32.21	3.77	30.45
11	4.03	3.59	2.03	0.83	18.56	2.49	10.33
12	4.80	4.37	2.35	0.34	23.69	2.97	15.75
13	3.35	2.91	2.02	0.94	27.63	5.76	54.22
14	4.05	3.57	2.09	0.56	19.16	2.32	9.66
15	3.79	3.51	1.70	0.96	16.35	2.14	9.03
16	3.74	3.44	1.58	0.74	12.50	1.64	4.75
17	4.77	4.14	3.32	0.19	55.60	7.89	103.44
18	5.41	4.98	2.27	0.77	18.21	1.40	3.87
19	5.06	4.43	2.78	0.24	26.10	3.36	17.44
20	5.03	4.57	2.63	0.69	25.46	2.36	10.69
21	4.62	4.17	2.37	0.31	24.64	2.37	12.72
22	4.50	3.94	3.11	0.78	44.92	6.05	61.38
23	5.06	4.37	3.62	0.56	43.57	5.27	42.91
24	4.71	4.33	2.66	0.52	26.30	2.60	12.69
25	3.73	3.22	2.59	0.60	35.99	6.98	71.47
26	5.32	4.81	3.07	0.78	41.22	4.29	38.92
27	4.54	3.87	4.13	0.60	72.71	10.35	150.15
28	5.59	5.00	2.68	0.44	24.38	1.89	7.73
29	6.42	5.78	3.33	0.42	29.73	2.06	8.93
30	4.37	3.80	4.27	0.09	70.38	11.75	168.08
31	5.96	5.50	2.81	0.37	24.60	1.76	6.69
32	5.59	4.86	4.24	0.53	63.45	8.04	93.84
33	5.58	4.88	3.65	0.54	56.40	6.19	71.12
34	4.74	4.25	2.23	0.77	18.13	1.84	6.75
35	5.29	4.82	2.61	0.60	27.79	2.88	15.93
36	6.83	6.02	3.82	0.92	35.47	2.88	14.23
37	6.08	5.45	3.31	0.70	42.71	3.67	29.40
38	6.00	5.35	3.77	0.86	56.48	6.53	73.51

(continued)

Item	Mean	*SD*	Median	Min.	Max.	Skew	Kurtosis
39	5.50	4.89	3.25	0.55	50.01	5.47	63.30
40	5.30	5.00	2.46	0.43	26.79	2.62	15.10
41	5.24	4.85	3.32	0.24	64.25	10.44	177.61
42	5.85	5.47	2.68	0.15	31.32	2.23	15.60
43	4.77	4.33	2.10	0.81	18.38	1.84	6.68
44	7.49	6.66	4.52	0.91	62.75	4.42	42.90
45	3.63	3.28	1.80	0.73	19.84	3.30	19.65
46	4.85	4.52	2.28	0.67	22.46	2.54	14.28
47	4.37	3.99	2.22	0.73	30.56	4.33	39.65
48	4.70	4.30	2.34	0.73	27.37	3.21	22.59
49	6.84	6.57	3.15	0.83	37.97	2.14	16.82
50	5.01	4.67	2.28	0.39	22.40	1.79	8.16
51	5.38	4.78	2.90	0.36	36.16	4.01	31.26
52	3.83	3.48	2.08	0.57	37.03	7.71	113.53
53	6.12	5.33	3.97	0.69	40.84	3.57	20.86

Table 5

Demographic Information about Respondents in Primary Sample (N = 568)

	Primary Sample	
	%	*N*
Gender		
Female	42.8	243
Male	57.2	523
Ethnicity		
White	86.4	491
Black	5.1	29
Asian	5.3	30
Hispanic	6.5	37
American Indian	0.7	4
Pacific Islander	0.9	5
Not specified	2.5	14
US Citizen	97.0	551

Table 6

Descriptive Statistics of the MFLS Items' Response Time Distributions for the Known Rapid Responders Sample (N = 181)

Item	Mean	*SD*	Median	Min	Max	Skew	Kurtosis
1	4.15	3.23	2.60	0.75	14.28	1.77	3.38
2	2.70	2.11	1.76	0.87	10.97	2.02	4.71
3	2.34	2.00	1.31	0.85	10.26	2.10	7.83
4	2.13	1.80	1.11	0.63	7.48	1.62	3.59
5	2.28	1.87	1.66	0.59	13.99	3.13	15.17
6	2.07	1.75	1.24	0.53	8.91	1.94	5.69
7	2.19	1.70	1.47	0.52	9.63	1.91	4.53
8	2.16	1.68	1.41	0.49	11.36	2.64	11.40
9	2.21	1.77	1.46	0.48	10.04	2.12	6.88
10	2.06	1.76	1.13	0.57	7.02	1.60	3.56
11	2.02	1.67	1.29	0.31	8.95	2.31	7.44
12	2.12	1.61	1.41	0.54	8.52	1.80	3.70
13	1.85	1.64	0.91	0.56	6.63	1.51	4.12
14	2.01	1.68	1.09	0.38	6.97	1.58	2.97
15	1.85	1.59	0.98	0.55	6.96	1.74	4.70
16	1.86	1.60	0.98	0.51	6.62	1.54	3.55
17	1.96	1.58	1.18	0.59	6.88	1.68	2.82
18	2.12	1.67	1.37	0.44	8.55	1.68	2.92
19	2.00	1.58	1.35	0.53	11.21	2.60	11.55
20	2.12	1.55	1.66	0.31	10.91	2.50	7.46
21	2.03	1.58	1.42	0.67	11.20	2.56	10.23
22	1.96	1.66	1.18	0.60	10.11	2.78	13.20
23	1.99	1.60	1.39	0.55	9.41	2.70	9.60
24	1.92	1.63	1.13	0.55	7.21	1.84	4.69
25	1.81	1.49	0.95	0.62	6.24	1.82	4.25
26	2.08	1.54	1.45	0.52	9.05	2.17	5.83
27	1.96	1.59	1.13	0.66	7.24	1.99	4.96
28	2.21	1.60	1.56	0.64	8.91	1.95	4.04
29	2.37	1.60	1.94	0.68	9.91	1.95	3.56
30	2.02	1.67	1.15	0.57	7.72	1.62	3.89
31	2.30	1.55	1.80	0.58	10.54	1.93	4.19
32	2.18	1.58	1.57	0.58	10.33	2.22	6.84
33	2.10	1.56	1.52	0.56	9.88	2.15	6.49
34	2.10	1.58	1.46	0.52	11.17	2.29	8.25
35	2.26	1.62	1.56	0.55	9.05	1.58	2.25
36	2.38	1.56	1.87	0.64	9.66	1.97	3.90
37	2.22	1.61	1.52	0.60	8.49	1.71	2.55
38	2.21	1.52	1.68	0.61	10.64	1.92	4.15

(continued)

Item	Mean	*SD*	Median	Min	Max	Skew	Kurtosis
39	2.11	1.55	1.52	0.56	9.05	2.11	5.33
40	2.02	1.51	1.40	0.59	8.55	1.82	3.55
41	2.13	1.53	1.65	0.48	13.39	2.85	12.72
42	2.08	1.54	1.52	0.42	9.72	1.87	4.00
43	1.96	1.57	1.20	0.54	7.59	1.76	3.31
44	2.17	1.54	1.67	0.74	12.49	2.63	9.24
45	1.88	1.60	0.91	0.68	4.71	1.12	0.64
46	2.04	1.47	1.42	0.49	7.88	1.75	2.89
47	2.03	1.44	1.43	0.65	7.89	2.04	4.65
48	2.05	1.48	1.48	0.52	11.31	2.33	8.56
49	2.30	1.61	1.67	0.70	9.07	1.83	3.17
50	2.20	1.68	1.57	0.54	11.22	2.34	7.35
51	2.11	1.55	1.42	0.54	10.91	2.20	8.05
52	1.88	1.56	1.00	0.58	6.38	1.45	3.11
53	2.03	1.54	1.39	0.57	6.91	1.81	2.69

Table 7

Descriptive Statistics of the Items' Response Time Distributions for the Primary Sample Combined with the Known Rapid Responders Sample (N = 749)

Item	Mean	*SD*	Median	Min	Max	Skew	Kurtosis
1	8.58	6.72	7.10	0.54	85.16	3.96	31.20
2	5.29	3.76	4.91	0.56	71.89	8.05	132.48
3	4.08	2.54	3.57	0.59	36.41	3.97	37.97
4	3.76	1.87	3.53	0.63	15.02	1.25	3.43
5	4.03	2.49	3.71	0.40	35.42	3.83	36.26
6	3.96	2.51	3.59	0.53	30.97	3.21	23.25
7	4.38	2.69	4.14	0.22	37.82	3.65	36.41
8	4.14	2.60	3.75	0.49	27.33	2.18	10.73
9	4.22	2.36	3.90	0.48	18.79	1.19	2.80
10	3.96	2.47	3.53	0.57	32.21	3.26	25.89
11	3.54	2.06	3.25	0.31	18.56	2.15	8.70
12	4.15	2.45	3.97	0.34	23.69	2.31	11.87
13	2.98	1.93	2.64	0.56	27.63	5.42	53.71
14	3.56	2.09	3.27	0.38	19.16	2.12	8.79
15	3.32	1.76	3.08	0.55	16.35	1.78	7.07
16	3.29	1.67	3.07	0.51	12.50	1.35	3.66
17	4.09	3.19	3.73	0.19	55.60	7.19	98.29
18	4.62	2.52	4.51	0.44	18.21	1.00	2.31
19	4.32	2.83	4.03	0.24	26.10	2.80	14.49
20	4.33	2.73	3.99	0.31	25.46	1.95	8.22
21	3.99	2.44	3.68	0.31	24.64	1.99	9.85
22	3.88	2.98	3.44	0.60	44.92	5.66	59.88
23	4.32	3.48	3.86	0.55	43.57	4.91	41.74
24	4.04	2.66	3.72	0.52	26.30	2.35	11.26
25	3.27	2.45	2.90	0.60	35.99	6.67	72.84
26	4.53	3.10	4.23	0.52	41.22	3.57	31.51
27	3.92	3.80	3.44	0.60	72.71	10.44	165.02
28	4.77	2.85	4.46	0.44	24.38	1.49	5.44
29	5.45	3.51	5.10	0.42	29.73	1.64	6.49
30	3.80	3.89	3.39	0.09	70.38	12.11	191.36
31	5.08	3.03	4.94	0.37	24.60	1.32	4.39
32	4.77	4.04	4.35	0.53	63.45	7.43	91.50
33	4.74	3.59	4.26	0.54	56.40	5.34	62.04
34	4.10	2.36	3.77	0.52	18.13	1.49	4.93
35	4.55	2.73	4.41	0.55	27.79	2.21	11.73
36	5.76	3.94	5.38	0.64	35.47	2.35	11.28
37	5.14	3.41	4.85	0.60	42.71	2.92	22.29
38	5.08	3.75	4.76	0.61	56.48	5.40	61.14

(continued)

Item	Mean	*SD*	Median	Min	Max	Skew	Kurtosis
39	4.68	3.27	4.31	0.55	50.01	4.47	50.72
40	4.51	2.65	4.41	0.43	26.79	1.86	9.81
41	4.49	3.28	4.31	0.24	64.25	8.58	147.75
42	4.94	2.93	4.90	0.15	31.32	1.55	9.23
43	4.09	2.26	3.90	0.54	18.38	1.38	4.52
44	6.21	4.62	5.75	0.74	62.75	3.54	32.87
45	3.20	1.79	2.99	0.68	19.84	2.88	17.29
46	4.17	2.42	4.03	0.49	22.46	1.88	9.72
47	3.81	2.29	3.64	0.65	30.56	3.32	29.19
48	4.06	2.44	3.84	0.52	27.37	2.46	16.22
49	5.75	3.46	5.76	0.70	37.97	1.48	9.65
50	4.33	2.44	4.12	0.39	22.40	1.39	5.38
51	4.59	2.97	4.34	0.36	36.16	3.19	24.11
52	3.36	2.06	3.13	0.57	37.03	6.44	95.94
53	5.13	3.94	4.67	0.57	40.84	3.20	19.00

Table 8

Defined Time Thresholds for MFLS Items, by Threshold Calculation Method

Item	INSPECT	INSPECT2	MIXTURE	MIXTURE2	NT10	NT20	NT30	RSPEED
1	2.00	2.50	2.65	--	1.00	2.00	3.00	1.50
2	2.00	2.30	2.20	3.90	0.61	1.22	1.83	3.30
3	2.00	2.30	1.60	3.10	0.46	0.93	1.39	2.70
4	1.80	2.30	1.80	3.10	0.43	0.86	1.28	2.40
5	1.30	2.30	2.10	3.10	0.46	0.92	1.38	3.30
6	1.90	2.00	1.95	3.15	0.46	0.91	1.37	2.10
7	2.50	2.80	2.35	3.20	0.51	1.02	1.52	3.60
8	1.90	2.50	3.00	--	0.48	0.95	1.43	3.00
9	2.00	2.50	1.60	1.65	0.49	0.97	1.46	1.80
10	1.90	2.00	1.70	3.15	0.46	0.91	1.37	2.40
11	1.80	2.00	2.15	2.75	0.40	0.81	1.21	1.50
12	2.00	2.10	2.05	2.95	0.48	0.96	1.44	3.00
13	1.80	1.90	--	2.10	0.33	0.67	1.00	1.80
14	1.60	1.70	1.95	--	0.41	0.81	1.22	2.10
15	1.80	1.90	2.50	3.00	0.38	0.76	1.14	2.40
16	1.80	1.90	1.80	2.45	0.37	0.75	1.12	2.10
17	2.00	2.00	1.95	3.00	0.48	0.95	1.43	3.30
18	2.70	2.60	2.60	3.35	0.54	1.08	1.62	3.00
19	2.40	2.50	2.15	3.10	0.51	1.01	1.52	3.60
20	2.20	2.50	2.90	3.55	0.50	1.01	1.51	3.90
21	2.00	2.50	1.85	3.20	0.46	0.92	1.39	3.90
22	2.00	2.10	2.10	3.35	0.45	0.90	1.35	2.70
23	2.00	1.90	2.30	3.60	0.51	1.01	1.52	3.90
24	2.00	2.20	2.80	3.70	0.47	0.94	1.41	3.90
25	1.50	1.90	1.65	2.75	0.37	0.75	1.12	2.40
26	2.30	2.60	2.70	3.55	0.53	1.06	1.60	4.20
27	1.80	2.20	1.70	3.60	0.45	0.91	1.36	1.20
28	2.40	2.50	2.40	3.35	0.56	1.12	1.68	3.30
29	2.00	2.60	2.40	3.65	0.64	1.28	1.93	3.00
30	1.80	2.00	1.45	2.40	0.44	0.87	1.31	2.10
31	2.40	2.70	2.80	3.50	0.60	1.19	1.79	4.50
32	2.30	2.60	2.20	3.25	0.56	1.12	1.68	4.20
33	2.00	2.50	2.05	3.30	0.56	1.12	1.67	3.90
34	2.00	2.30	2.10	3.00	0.47	0.95	1.42	4.20
35	2.30	2.60	2.65	3.35	0.53	1.06	1.59	6.30
36	2.90	2.90	3.10	3.95	0.68	1.37	2.05	6.30
37	2.60	2.70	2.85	3.80	0.61	1.22	1.82	5.10
38	2.80	2.70	2.60	3.50	0.60	1.20	1.80	3.60
39	2.50	2.50	2.15	3.20	0.55	1.10	1.65	3.60

(continued)

Item	INSPECT	INSPECT2	MIXTURE	MIXTURE2	NT10	NT20	NT30	RSPEED
40	2.10	2.60	2.45	3.30	0.53	1.06	1.59	3.90
41	1.80	2.30	2.25	3.20	0.52	1.05	1.57	3.60
42	2.30	2.60	2.65	3.35	0.59	1.17	1.76	4.20
43	2.00	2.30	2.20	3.10	0.48	0.95	1.43	3.60
44	2.30	2.80	3.60	4.60	0.75	1.50	2.25	5.10
45	1.50	1.70	1.45	3.00	0.36	0.73	1.09	2.10
46	2.10	2.30	2.30	3.05	0.49	0.97	1.46	3.90
47	1.90	2.30	2.10	2.80	0.44	0.87	1.31	2.40
48	2.20	2.10	2.15	2.95	0.47	0.94	1.41	4.50
49	2.50	3.00	3.75	2.85	0.68	1.37	2.05	6.30
50	2.20	2.40	2.35	3.20	0.50	1.00	1.50	3.30
51	2.00	2.30	2.30	3.25	0.54	1.08	1.62	2.70
52	1.90	1.80	1.60	2.45	0.38	0.77	1.15	2.10
53	2.60	2.70	2.95	4.00	0.61	1.22	1.83	4.20

Note. Dashed lines indicate a time threshold was not calculated for that item. INSPECT = visual inspection; INSPECT2 = visual inspection with information; MIXTURE = lognormal mixture modeling; MIXTURE2 = lognormal mixture modeling with information; NT10 = 10% normative threshold; NT20 = 20% normative threshold; NT30 = 30% normative threshold; RSPEED = reading speed.

Table 9

Descriptive Statistics of the Time Thresholds for MLFS Items, by Threshold Calculation Method

Calculation Method	Mean	*SD*	Median	Min	Max	Skew	Kurtosis	N	N Miss
INSPECT	2.08	0.33	2.00	1.30	2.90	0.33	0.22	53	0
INSPECT2	2.34	0.32	2.30	1.70	3.00	-0.17	-0.75	53	0
MIXTURE	2.29	0.50	2.20	1.45	3.75	0.68	0.71	52	1
MIXTURE2	3.19	0.49	3.20	1.65	4.60	-0.33	2.21	50	3
NT10	0.51	0.11	0.49	0.33	1.00	1.83	6.42	53	0
NT20	1.02	0.22	0.97	0.67	2.00	1.83	6.42	53	0
NT30	1.54	0.33	1.46	1.00	3.00	1.83	6.42	53	0
RSPEED	3.34	1.18	3.30	1.20	6.30	0.61	0.45	53	0

Note. INSPECT = visual inspection; INSPECT2 = visual inspection with information; MIXTURE = lognormal mixture modeling; MIXTURE2 = lognormal mixture modeling with information; NT10 = 10% normative threshold; NT20 = 20% normative threshold; NT30 = 30% normative threshold; RSPEED = reading speed. *SD* = standard deviation; Min = minimum; Max = maximum; Skew = skewness; N miss = number of missing thresholds.

Table 10

Model Fit Indices for One- and Two-Class Lognormal Mixture Models

	One-class				Two-class			
Item	LL	AIC	BIC	SSABIC	LL	AIC	BIC	SSABIC
1	3368.90	3372.90	3381.58	3375.23	3337.52	3347.52	3369.23	3343.85
2	2655.48	2659.48	2668.17	2661.82	2583.09	2593.09	2614.80	2589.42
3	2332.52	2336.52	2345.21	2338.86	2302.50	2312.50	2334.21	2308.83
4	2134.68	2138.68	2147.37	2141.02	2112.10	2122.10	2143.82	2118.44
5	2289.15	2293.15	2301.83	2295.48	2221.46	2231.46	2253.17	2227.79
6	2328.80	2332.80	2341.48	2335.13	2291.41	2301.41	2323.12	2297.75
7	2474.42	2478.42	2487.10	2480.76	2374.44	2384.44	2406.15	2380.78
8	2449.54	2453.54	2462.23	2455.88	2431.94	2441.94	2463.65	2438.28
9	2408.75	2412.75	2421.43	2415.08	2378.61	2388.61	2410.32	2384.94
10	2331.55	2335.55	2344.23	2337.88	2296.45	2306.45	2328.16	2302.79
11	2142.53	2146.53	2155.21	2148.86	2107.03	2117.03	2138.74	2113.37
12	2350.36	2354.36	2363.04	2356.69	2258.72	2268.72	2290.43	2265.06
13	1885.91	1889.91	1898.59	1892.24	1844.01	1854.01	1875.72	1850.35
14	2206.58	2210.58	2219.26	2212.91	2187.42	2197.42	2219.13	2193.76
15	2001.88	2005.88	2014.56	2008.21	1985.70	1995.70	2017.41	1992.03
16	1990.34	1994.34	2003.03	1996.68	1960.12	1970.12	1991.83	1966.45
17	2451.81	2455.81	2464.50	2458.15	2315.83	2325.83	2347.54	2322.16
18	2481.38	2485.38	2494.06	2487.71	2393.08	2403.08	2424.79	2399.41
19	2457.91	2461.91	2470.60	2464.25	2339.08	2349.08	2370.79	2345.41
20	2506.86	2510.86	2519.55	2513.20	2460.83	2470.83	2492.54	2467.17
21	2431.17	2435.17	2443.85	2437.50	2370.45	2380.45	2402.16	2376.78
22	2368.01	2372.01	2380.69	2374.34	2298.14	2308.14	2329.85	2304.47
23	2556.65	2560.65	2569.33	2562.98	2472.84	2482.84	2504.55	2479.17
24	2502.65	2506.65	2515.33	2508.98	2437.40	2447.40	2469.12	2443.74
25	2059.34	2063.34	2072.03	2065.68	1985.57	1995.57	2017.28	1991.91
26	2608.96	2612.96	2621.64	2615.30	2534.87	2544.87	2566.58	2541.21
27	2377.14	2381.14	2389.82	2383.47	2315.06	2325.06	2346.77	2321.40
28	2613.38	2617.38	2626.07	2619.72	2540.16	2550.16	2571.87	2546.49
29	2859.69	2863.69	2872.38	2866.03	2775.18	2785.18	2806.89	2781.52
30	2277.56	2281.56	2290.24	2283.89	2113.72	2123.72	2145.43	2120.05
31	2731.27	2735.27	2743.95	2737.60	2592.40	2602.40	2624.11	2598.74
32	2623.03	2627.03	2635.72	2629.37	2522.28	2532.28	2553.99	2528.61
33	2635.00	2639.00	2647.69	2641.34	2562.82	2572.82	2594.54	2569.16
34	2376.45	2380.45	2389.13	2382.78	2332.05	2342.05	2363.76	2338.39
35	2499.64	2503.64	2512.33	2505.98	2392.27	2402.27	2423.99	2398.61
36	2882.96	2886.96	2895.65	2889.30	2794.52	2804.52	2826.23	2800.85
37	2713.15	2717.15	2725.84	2719.49	2630.67	2640.67	2662.38	2637.00
38	2663.15	2667.15	2675.83	2669.48	2581.84	2591.84	2613.55	2588.17

(continued)

	One-class				Two-class			
Item	LL	AIC	BIC	SSABIC	LL	AIC	BIC	SSABIC
39	2618.01	2622.01	2630.69	2624.34	2535.67	2545.67	2567.38	2542.01
40	2515.88	2519.88	2528.57	2522.22	2381.40	2391.40	2413.11	2387.73
41	2533.84	2537.84	2546.52	2540.18	2362.49	2372.49	2394.20	2368.83
42	2749.68	2753.68	2762.37	2756.02	2548.13	2558.13	2579.84	2554.46
43	2314.98	2318.98	2327.66	2321.31	2259.92	2269.92	2291.63	2266.25
44	3045.29	3049.29	3057.98	3051.63	2973.72	2983.72	3005.43	2980.05
45	1962.14	1966.14	1974.83	1968.48	1927.35	1937.35	1959.06	1933.68
46	2404.46	2408.46	2417.14	2410.79	2316.73	2326.73	2348.44	2323.07
47	2227.47	2231.47	2240.15	2233.80	2162.44	2172.44	2194.15	2168.78
48	2372.23	2376.23	2384.91	2378.56	2289.61	2299.61	2321.32	2295.95
49	2909.30	2913.30	2921.98	2915.63	2762.60	2772.60	2794.31	2768.94
50	2482.90	2486.90	2495.58	2489.23	2382.43	2392.43	2414.14	2388.77
51	2566.70	2570.70	2579.38	2573.03	2441.26	2451.26	2472.97	2447.60
52	2024.63	2028.63	2037.31	2030.96	1978.12	1988.12	2009.83	1984.46
53	2801.80	2805.80	2814.48	2808.13	2725.17	2735.17	2756.88	2731.50

Note. LL = log-likelihood; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; SSABIC = Sample Size Adjusted BIC.

Table 11

Model Fit Indices for One- and Two-Class Lognormal Mixture Models with Information

	One-class				Two-class			
Item	LL	AIC	BIC	SSABIC	LL	AIC	BIC	SSABIC
1	4405.94	4409.94	4419.17	4412.82	4388.65	4398.65	4421.75	4395.54
2	3610.33	3614.33	3623.57	3617.22	3528.22	3538.22	3561.32	3535.11
3	3116.64	3120.64	3129.87	3123.52	3093.24	3103.24	3126.33	3100.12
4	2951.55	2955.55	2964.79	2958.44	2902.15	2912.15	2935.25	2909.04
5	3148.94	3152.94	3162.18	3155.83	3073.03	3083.03	3106.12	3079.91
6	3152.72	3156.72	3165.96	3159.61	3103.46	3113.46	3136.56	3110.35
7	3381.69	3385.69	3394.93	3388.58	3244.91	3254.91	3278.00	3251.79
8	3256.26	3260.26	3269.50	3263.15	3223.66	3233.66	3256.76	3230.55
9	3298.68	3302.68	3311.91	3305.56	3215.27	3225.27	3248.36	3222.15
10	3141.79	3145.79	3155.03	3148.68	3100.73	3110.73	3133.82	3107.62
11	2936.35	2940.35	2949.59	2943.24	2889.03	2899.03	2922.13	2895.92
12	3271.20	3275.20	3284.44	3278.09	3139.21	3149.21	3172.30	3146.10
13	2533.55	2537.55	2546.78	2540.43	2507.45	2517.45	2540.55	2514.34
14	2937.00	2941.00	2950.24	2943.89	2914.59	2924.59	2947.68	2921.47
15	2775.90	2779.90	2789.14	2782.79	2734.74	2744.74	2767.83	2741.62
16	2766.58	2770.58	2779.82	2773.47	2709.21	2719.21	2742.31	2716.10
17	3267.84	3271.84	3281.08	3274.73	3156.91	3166.91	3190.00	3163.80
18	3483.10	3487.10	3496.34	3489.99	3313.24	3323.24	3346.34	3320.13
19	3372.14	3376.14	3385.38	3379.03	3234.82	3244.82	3267.91	3241.70
20	3404.89	3408.89	3418.12	3411.77	3317.55	3327.55	3350.64	3324.44
21	3250.28	3254.28	3263.52	3257.17	3182.28	3192.28	3215.37	3189.17
22	3135.66	3139.66	3148.90	3142.55	3080.92	3090.92	3114.01	3087.80
23	3383.96	3387.96	3397.20	3390.84	3308.54	3318.54	3341.64	3315.43
24	3288.45	3292.45	3301.69	3295.34	3215.51	3225.51	3248.61	3222.40
25	2752.58	2756.58	2765.82	2759.47	2705.53	2715.53	2738.62	2712.42
26	3492.69	3496.69	3505.93	3499.58	3387.89	3397.89	3420.98	3394.77
27	3123.26	3127.26	3136.50	3130.14	3082.02	3092.02	3115.11	3088.90
28	3561.79	3565.79	3575.03	3568.68	3443.83	3453.83	3476.93	3450.72
29	3861.78	3865.78	3875.02	3868.67	3720.23	3730.23	3753.32	3727.12
30	3049.82	3053.82	3063.06	3056.71	2935.91	2945.91	2969.01	2942.80
31	3730.03	3734.03	3743.27	3736.92	3522.09	3532.09	3555.19	3528.98
32	3551.67	3555.67	3564.91	3558.56	3429.69	3439.69	3462.78	3436.57
33	3575.04	3579.04	3588.27	3581.92	3467.37	3477.37	3500.46	3474.25
34	3268.74	3272.74	3281.98	3275.63	3178.22	3188.22	3211.32	3185.11
35	3456.51	3460.51	3469.75	3463.39	3283.58	3293.58	3316.67	3290.46
36	3927.77	3931.77	3941.01	3934.66	3768.08	3778.08	3801.17	3774.96
37	3702.94	3706.94	3716.18	3709.83	3562.75	3572.75	3595.84	3569.63
38	3696.93	3700.93	3710.16	3703.81	3532.74	3542.74	3565.83	3539.62

(continued)

	One-class				Two-class			
Item	LL	AIC	BIC	SSABIC	LL	AIC	BIC	SSABIC
39	3552.98	3556.98	3566.22	3559.87	3443.81	3453.81	3476.90	3450.70
40	3493.04	3497.04	3506.28	3499.93	3287.84	3297.84	3320.93	3294.72
41	3463.96	3467.96	3477.20	3470.85	3267.81	3277.81	3300.90	3274.69
42	3721.25	3725.25	3734.49	3728.14	3470.35	3480.35	3503.44	3477.24
43	3240.88	3244.88	3254.11	3247.76	3131.73	3141.73	3164.82	3138.61
44	4090.44	4094.44	4103.68	4097.33	3950.42	3960.42	3983.52	3957.31
45	2667.63	2671.63	2680.87	2674.52	2628.95	2638.95	2662.04	2635.83
46	3330.03	3334.03	3343.26	3336.91	3174.39	3184.39	3207.49	3181.28
47	3105.20	3109.20	3118.44	3112.09	2998.18	3008.18	3031.27	3005.06
48	3273.22	3277.22	3286.46	3280.11	3138.96	3148.96	3172.06	3145.85
49	3947.81	3951.81	3961.04	3954.69	3668.02	3678.02	3701.12	3674.91
50	3373.99	3377.99	3387.22	3380.87	3249.31	3259.31	3282.40	3256.20
51	3498.08	3502.08	3511.31	3504.96	3342.60	3352.60	3375.70	3349.49
52	2805.51	2809.51	2818.75	2812.40	2733.60	2743.60	2766.69	2740.48
53	3732.88	3736.88	3746.12	3739.77	3629.29	3639.29	3662.38	3636.18

Note. LL = log-likelihood; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; SSABIC = Sample Size Adjusted BIC.

Table 12

Time Threshold Agreement Indices for the Eight Threshold Calculation Methods

	INPSECT	INSPECT2	MIXTURE	MIXTURE2	NT10	NT20	NT30	RSPEED
INSPECT	—	1.00	1.00	0.98	0.94	1.00	1.00	0.85
INSPECT2	0.06	—	1.00	1.00	0.77	1.00	1.00	0.89
MIXTURE	0.12	0.17	—	1.00	0.71	0.94	1.00	0.87
MIXTURE2	0.00	0.00	0.00	—	0.06	0.20	0.86	0.92
NT10	0.00	0.00	0.00	0.00	—	1.00	1.00	0.30
NT20	0.02	0.00	0.00	0.00	0.00	—	1.00	0.43
NT30	0.00	0.00	0.00	0.00	0.00	0.00	—	0.58
RSPEED	0.02	0.00	0.02	0.02	0.00	0.00	0.00	—

Note. Values on the lower diagonal represent the proportion of items whose thresholds were in exact agreement. Values on the upper diagonal represent the proportion of items whose thresholds differed by no more than two seconds. INSPECT = visual inspection; INSPECT2 = visual inspection with information; MIXTURE = lognormal mixture modeling; MIXTURE2 = lognormal mixture modeling with information; NT10 = 10% normative threshold; NT20 = 20% normative threshold; NT30 = 30% normative threshold; RSPEED = reading speed.

Table 13

Proportion of Respondents Classified as Exhibiting Solution Behavior, by Item and Threshold Calculation Method

Item	INSPECT	INSPECT2	MIXTURE	MIXTURE2	NT10	NT20	NT30	RSPEED
1	0.993	0.989	0.989	--	0.998	0.993	0.982	0.995
2	0.977	0.972	0.975	0.826	0.996	0.993	0.977	0.905
3	0.972	0.933	0.979	0.759	1.000	0.996	0.984	0.873
4	0.970	0.921	0.970	0.768	1.000	0.996	0.989	0.905
5	0.986	0.942	0.954	0.803	0.998	0.998	0.981	0.731
6	0.963	0.958	0.961	0.741	1.000	0.998	0.977	0.951
7	0.933	0.912	0.933	0.845	0.998	0.995	0.972	0.769
8	0.942	0.871	0.780	--	1.000	0.993	0.974	0.780
9	0.965	0.926	0.972	0.972	1.000	0.993	0.979	0.968
10	0.963	0.952	0.974	0.739	1.000	0.996	0.982	0.905
11	0.958	0.930	0.914	0.782	1.000	1.000	0.988	0.967
12	0.970	0.970	0.970	0.866	0.998	0.996	0.981	0.863
13	0.924	0.898	--	0.836	1.000	1.000	0.998	0.924
14	0.965	0.956	0.928	--	1.000	0.998	0.979	0.900
15	0.954	0.940	0.827	0.641	1.000	1.000	0.995	0.857
16	0.961	0.954	0.961	0.833	1.000	0.998	0.984	0.930
17	0.952	0.952	0.954	0.838	0.996	0.989	0.972	0.761
18	0.951	0.952	0.952	0.891	1.000	0.991	0.974	0.921
19	0.960	0.952	0.963	0.863	0.998	0.991	0.979	0.741
20	0.924	0.896	0.857	0.750	1.000	0.991	0.968	0.662
21	0.937	0.891	0.940	0.754	0.996	0.988	0.972	0.548
22	0.933	0.921	0.921	0.671	1.000	0.991	0.970	0.827
23	0.947	0.949	0.919	0.704	1.000	0.986	0.967	0.620
24	0.914	0.898	0.820	0.637	1.000	0.984	0.960	0.599
25	0.970	0.930	0.961	0.685	1.000	0.996	0.988	0.822
26	0.926	0.898	0.884	0.775	1.000	0.989	0.952	0.644
27	0.949	0.901	0.958	0.574	1.000	0.995	0.974	0.981
28	0.944	0.942	0.944	0.879	0.998	0.988	0.963	0.884
29	0.963	0.942	0.952	0.857	0.995	0.981	0.963	0.919
30	0.968	0.958	0.975	0.910	0.998	0.995	0.981	0.951
31	0.945	0.937	0.931	0.882	0.998	0.974	0.954	0.710
32	0.956	0.935	0.958	0.856	0.998	0.984	0.968	0.669
33	0.963	0.937	0.960	0.852	0.998	0.982	0.975	0.713
34	0.956	0.931	0.947	0.845	1.000	0.995	0.972	0.518
35	0.944	0.933	0.931	0.854	1.000	0.993	0.970	0.217
36	0.940	0.940	0.935	0.866	1.000	0.982	0.960	0.461
37	0.949	0.947	0.935	0.827	1.000	0.988	0.970	0.579
38	0.940	0.940	0.947	0.882	1.000	0.991	0.975	0.868

(continued)

Item	INSPECT	INSPECT2	MIXTURE	MIXTURE2	NT10	NT20	NT30	RSPEED
39	0.954	0.954	0.960	0.856	0.998	0.981	0.965	0.778
40	0.961	0.940	0.945	0.870	0.998	0.984	0.970	0.746
41	0.965	0.949	0.951	0.884	0.993	0.991	0.968	0.799
42	0.940	0.931	0.931	0.891	0.996	0.977	0.949	0.775
43	0.965	0.952	0.960	0.847	1.000	0.996	0.977	0.720
44	0.952	0.937	0.894	0.819	1.000	0.977	0.954	0.754
45	0.977	0.956	0.979	0.613	1.000	1.000	0.989	0.889
46	0.945	0.938	0.938	0.845	1.000	0.989	0.968	0.667
47	0.956	0.933	0.945	0.850	1.000	0.996	0.975	0.924
48	0.951	0.952	0.952	0.850	1.000	0.989	0.967	0.452
49	0.937	0.910	0.875	0.919	1.000	0.979	0.951	0.548
50	0.940	0.933	0.933	0.854	0.998	0.988	0.967	0.835
51	0.961	0.954	0.954	0.882	0.998	0.984	0.967	0.931
52	0.954	0.960	0.972	0.847	1.000	0.998	0.988	0.923
53	0.928	0.921	0.910	0.745	1.000	0.981	0.958	0.702

Note. Dashed lines indicate a time threshold was not calculated for that item. INSPECT = visual inspection; INSPECT2 = visual inspection with information; MIXTURE = lognormal mixture modeling; MIXTURE2 = lognormal mixture modeling with information; NT10 = 10% normative threshold; NT20 = 20% normative threshold; NT30 = 30% normative threshold; RSPEED = reading speed.

Table 14

Omnibus Test Results for the Generalized Estimating Equations Analyzing Differences in Solution Behavior Classification Indices across Threshold Calculation Methods, by Item

Item	df_1	df_2	F	p	Item	df_1	df_2	F	p
1	4	3402	2.19	0.0673	28	6	3969	11.30	<.0001
2	6	3969	16.47	<.0001	29	6	3969	13.58	<.0001
3	6	3402	24.46	<.0001	30	7	3969	7.19	<.0001
4	5	3402	27.62	<.0001	31	6	3402	28.11	<.0001
5	6	3969	26.75	<.0001	32	7	3969	30.09	<.0001
6	6	3402	22.80	<.0001	33	6	3402	28.68	<.0001
7	6	3969	23.31	<.0001	34	6	3402	54.45	<.0001
8	4	2835	32.40	<.0001	35	6	3402	97.93	<.0001
9	5	3402	7.55	<.0001	36	5	3402	72.48	<.0001
10	6	3402	26.19	<.0001	37	6	3402	45.85	<.0001
11	4	2268	28.22	<.0001	38	5	3402	14.43	<.0001
12	4	3402	17.34	<.0001	39	6	3969	21.43	<.0001
13	4	2268	17.12	<.0001	40	7	3969	22.18	<.0001
14	4	2268	11.69	<.0001	41	7	3969	16.70	<.0001
15	5	2835	45.22	<.0001	42	6	3969	22.44	<.0001
16	5	3402	19.08	<.0001	43	6	3402	28.11	<.0001
17	6	3969	23.49	<.0001	44	6	3402	23.69	<.0001
18	5	3402	11.70	<.0001	45	5	2835	48.28	<.0001
19	7	3969	22.55	<.0001	46	5	3402	41.55	<.0001
20	6	3402	36.08	<.0001	47	6	3402	14.06	<.0001
21	7	3969	44.34	<.0001	48	4	2835	90.03	<.0001
22	5	3402	41.50	<.0001	49	6	3402	48.38	<.0001
23	6	3402	39.06	<.0001	50	6	3969	15.97	<.0001
24	6	3402	43.69	<.0001	51	6	3969	11.01	<.0001
25	6	3402	33.23	<.0001	52	6	3402	14.79	<.0001
26	6	3402	37.90	<.0001	53	6	3402	29.24	<.0001
27	6	3402	44.71	<.0001					

Note. Statistical significance was assessed using $\alpha = .01$.

Table 15

Pairwise Comparison Results Examining Differential Solution Behavior Classification Indices across Threshold Calculation Methods, by Item

	1. INSPECT							2. INSPECT2						3. MIXTURE					4. MIXTURE2				5. NT10			6. NT20		7. NT30
Item	2	3	4	5	6	7	8	3	4	5	6	7	8	4	5	6	7	8	5	6	7	8	6	7	8	7	8	8
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	0	0	1	1	1	0	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
3	1	0	1	c	1	1	1	1	1	c	1	1	1	1	c	1	0	1	c	1	1	1	c	c	c	0	1	1
4	1	0	1	c	1	1	1	1	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	0	1	1
5	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1
6	0	0	1	c	1	1	1	0	1	c	1	1	0	1	c	1	1	0	c	1	1	1	c	c	c	1	1	1
7	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
8	1	1	b	c	1	1	1	1	b	c	1	1	1	b	c	1	1	0	b	b	b	b	c	c	c	1	1	1
9	1	0	0	c	1	1	0	1	1	c	1	1	1	0	c	1	0	0	c	1	0	0	c	c	c	1	1	0
10	0	0	1	c	1	1	1	1	1	c	1	1	1	1	c	1	0	1	c	1	1	1	c	c	c	0	1	1
11	1	1	1	c	d	e	0	1	1	c	d	e	1	1	c	d	e	1	c	d	e	1	c	c	c	d	d	e
12	0	0	1	c	1	0	1	0	1	c	1	0	1	1	c	1	0	1	c	1	1	0	c	c	c	1	1	1
13	1	a	1	c	d	1	0	a	1	c	d	1	1	a	a	a	a	a	c	d	1	1	c	c	c	d	d	1
14	0	1	b	c	d	1	1	1	b	c	d	1	1	b	c	d	1	1	b	b	b	b	c	c	c	d	d	1
15	1	1	1	c	d	1	1	1	1	c	d	1	1	1	c	d	1	1	c	d	1	1	c	c	c	d	d	1
16	0	0	1	c	1	1	1	0	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	0	1	1
17	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
18	0	0	1	c	1	1	1	0	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	1	1	1
19	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1
20	1	1	1	c	1	1	1	1	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	1	1	1
21	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
22	1	1	1	c	1	1	1	0	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	1	1	1
23	0	1	1	c	1	1	1	1	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	1	1	1
24	1	1	1	c	1	1	1	1	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	1	1	1
25	1	0	1	c	1	1	1	1	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	0	1	1
26	1	1	1	c	1	1	1	1	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	1	1	1
27	1	0	1	c	1	1	1	1	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	1	1	0
28	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1

(continued)

	1. INSPECT							2. INSPECT2						3. MIXTURE					4. MIXTURE2				5. NT10			6. NT20		7. NT30
Item	2	3	4	5	6	7	8	3	4	5	6	7	8	4	5	6	7	8	5	6	7	8	6	7	8	7	8	8
29	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
30	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1
31	0	1	1	[c]	1	0	1	0	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
32	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
33	1	0	1	[c]	1	1	1	1	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	0	1	1
34	1	0	1	[c]	1	1	1	1	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
35	0	1	1	[c]	1	1	1	0	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
36	0	0	1	[c]	1	1	1	0	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
37	0	1	1	[c]	1	1	1	1	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
38	0	0	1	[c]	1	1	1	0	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
39	0	0	1	1	1	0	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
40	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
41	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
42	0	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
43	1	0	1	[c]	1	1	1	0	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
44	1	1	1	[c]	1	0	1	1	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
45	1	0	1	[c]	[d]	1	1	1	1	[c]	[d]	1	1	1	[c]	[d]	0	1	[c]	[d]	1	1	[c]	[c]	[c]	[d]	[d]	1
46	0	0	1	[c]	1	1	1	0	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
47	1	0	1	[c]	1	1	1	1	1	[c]	1	1	0	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
48	0	0	1	[c]	[d]	1	1	0	1	[c]	[d]	1	1	1	[c]	[d]	1	1	[c]	[d]	1	1	[c]	[c]	[c]	[d]	[d]	1
49	1	1	1	[c]	1	1	1	1	0	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1
50	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
51	0	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
52	0	1	1	[c]	1	1	1	1	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	0	1	1
53	0	1	1	[c]	1	1	1	0	1	[c]	1	1	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	1	1	1

Note. A maximum of 28 pairwise comparisons were conducted for each item. Statistical significance was assessed using $\alpha = .01$. Practical significance was assessed using a difference of .05. A value of 1 indicates the test was statistically significant whereas a value of 0 indicates the test was not statistically significant.

Calculation method 1 = INSPECT; Calculation method 2 = INSPECT2; Calculation method 3 = MIXTURE; Calculation method 4 = MIXTURE 2; Calculation method 5 = NT10; Calculation method 6 = NT20; Calculation method 7 = NT30; Calculation method 8 = RSPEED.

- = indicates the contrast tests were not conducted because the omnibus test was not significant.

[a] = indicates the contrast test was not conducted because the SB index calculated using the MIXTURE method was not included as a main effect.

[b] = indicates the contrast test was not conducted because the SB index calculated using the MIXTURE2 method was not included as a main effect.

[c] = indicates the contrast test was not conducted because the SB index calculated using the NT10 method was not included as a main effect.

[d] = indicates the contrast test was not conducted because the SB index calculated using the NT20 method was not included as a main effect.

[e] = indicates the contrast test was not conducted because the SB index calculated using the NT30 method was not included as a main effect.

Table 16

Practical Significance of the Pairwise Comparisons Examining Differential Solution Behavior Classifications across Threshold Calculation Methods, by Item

	1. INSPECT							2. INSPECT2						3. MIXTURE					4. MIXTURE2				5. NT10			6. NT20		7. NT30
Item	2	3	4	5	6	7	8	3	4	5	6	7	8	4	5	6	7	8	5	6	7	8	6	7	8	7	8	8
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	.	.	1	0	0	.	1	.	1	0	0	.	1	1	0	0	.	1	1	1	1	1	.	0	1	0	1	1
3	0	.	1	c	0	0	1	1	1	c	1	1	1	1	c	0	.	1	c	1	1	1	c	c	c	.	1	1
4	1	.	1	c	0	0	1	1	1	c	1	1	0	1	c	0	0	1	c	1	1	1	c	c	c	.	1	1
5	0	0	1	.	.	.	1	0	1	1	1	0	1	1	0	0	0	1	1	1	1	1	.	0	1	0	1	1
6	.	.	1	c	0	0	0	.	1	c	0	0	.	1	c	0	0	0	c	1	1	1	c	c	c	0	1	0
7	0	.	1	1	1	0	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	.	0	1	0	1	1
8	1	1	b	c	1	0	1	1	b	c	1	1	1	b	c	1	1	.	b	b	b	b	c	c	c	0	1	1
9	0	.	.	c	0	0	.	1	1	c	1	1	0	.	c	0	.	.	c	0	.	.	c	c	c	0	0	0
10	0	0	1	c	0	0	1	0	1	c	0	0	1	1	c	0	.	1	c	1	1	1	c	c	c	0	1	1
11	0	0	1	c	d	e	.	0	1	c	d	e	0	1	c	d	e	1	c	d	e	1	c	c	c	d	d	e
12	.	.	1	c	0	0	1	.	1	c	0	0	1	1	c	0	0	1	c	1	1	.	c	c	c	0	1	1
13	0	a	1	c	d	1	.	a	1	c	d	1	0	a	a	a	a	a	c	d	1	1	c	c	c	d	d	1
14	.	0	b	c	d	0	1	0	b	c	d	0	1	b	c	d	1	0	b	b	b	b	c	c	c	d	d	1
15	0	1	1	c	d	0	1	1	1	c	d	1	1	1	c	d	1	0	c	d	1	1	c	c	c	d	d	1
16	.	.	1	c	0	0	0	.	1	c	0	0	0	1	c	0	0	0	c	1	1	1	c	c	c	.	1	1
17	.	.	1	0	0	0	1	.	1	0	0	0	1	1	0	0	0	1	1	1	1	1	.	0	1	0	1	1
18	.	.	1	c	0	0	0	.	1	c	0	0	0	1	c	0	0	0	c	1	1	0	c	c	c	0	1	1
19	.	.	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	1	1	1	1	.	0	1	0	1	1
20	0	1	1	c	1	0	1	0	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	0	1	1
21	1	.	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	.	0	1	0	1	1
22	0	0	1	c	1	0	1	.	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	0	1	1
23	.	0	1	c	0	0	1	0	1	c	0	0	1	1	c	1	1	1	c	1	1	1	c	c	c	0	1	1
24	0	1	1	c	1	1	1	1	1	c	1	1	1	1	c	1	1	1	c	1	1	0	c	c	c	0	1	1
25	0	.	1	c	0	0	1	0	1	c	1	1	1	1	c	0	0	1	c	1	1	1	c	c	c	.	1	1
26	0	0	1	c	1	0	1	0	1	c	1	1	1	1	c	1	1	1	c	1	1	1	c	c	c	0	1	1
27	1	.	1	c	1	0	0	1	1	c	1	1	1	1	c	0	0	0	c	1	1	1	c	c	c	0	0	.
28	.	.	1	1	0	0	1	.	1	1	1	0	1	1	1	0	0	1	1	1	1	.	.	0	1	0	1	1

(continued)

	1. INSPECT							2. INSPECT2						3. MIXTURE					4. MIXTURE2				5. NT10			6. NT20		7. NT30
Item	2	3	4	5	6	7	8	3	4	5	6	7	8	4	5	6	7	8	5	6	7	8	6	7	8	7	8	8
29	0	0	1	0	0	.	0	0	1	1	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1	0	1	0
30	0	.	1	0	0	0	0	0	1	0	0	0	.	1	0	0	.	0	1	1	1	0	.	0	1	0	0	0
31	.	0	1	[c]	0	.	1	.	1	[c]	0	0	1	1	[c]	0	0	1	[c]	1	1	1	[c]	[c]	[c]	0	1	1
32	0	.	1	0	0	0	1	0	1	1	1	0	1	1	0	0	0	1	1	1	1	1	.	0	1	0	1	1
33	0	.	1	[c]	0	0	1	0	1	[c]	1	0	1	1	[c]	0	0	1	[c]	1	1	1	[c]	[c]	[c]	.	1	1
34	0	.	1	[c]	0	0	1	0	1	[c]	1	0	1	1	[c]	1	0	1	[c]	1	1	1	[c]	[c]	[c]	0	1	1
35	0	0	1	[c]	1	0	1	.	1	[c]	1	0	1	1	[c]	1	0	1	[c]	1	1	1	[c]	[c]	[c]	0	1	1
36	.	.	1	[c]	0	0	1	.	1	[c]	0	0	1	1	[c]	1	0	1	[c]	1	1	1	[c]	[c]	[c]	0	1	1
37	.	0	1	[c]	0	0	1	0	1	[c]	0	0	1	1	[c]	1	0	1	[c]	1	1	1	[c]	[c]	[c]	0	1	1
38	.	.	1	[c]	1	0	1	.	1	[c]	1	0	1	1	[c]	0	0	1	[c]	1	1	0	[c]	[c]	[c]	0	1	1
39	.	.	1	0	0	0	1	.	1	0	0	0	1	1	0	0	.	1	1	1	1	1	0	0	1	0	1	1
40	0	0	1	0	0	.	1	.	1	1	0	0	1	1	1	0	0	1	1	1	1	1	.	0	1	0	1	1
41	0	0	1	0	0	.	1	.	1	0	0	0	1	1	0	0	0	1	1	1	1	1	.	0	1	0	1	1
42	.	.	1	1	0	.	1	.	0	1	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1	0	1	1
43	0	.	1	[c]	0	0	1	.	1	[c]	0	0	1	1	[c]	0	0	1	[c]	1	1	1	[c]	[c]	[c]	0	1	1
44	0	1	1	[c]	0	.	1	0	1	[c]	0	0	1	1	[c]	1	1	1	[c]	1	1	1	[c]	[c]	[c]	0	1	1
45	0	.	1	[c]	[d]	0	1	0	1	[c]	[d]	0	1	1	[c]	[d]	.	1	[c]	[d]	1	1	[c]	[c]	[c]	[d]	[d]	1
46	.	.	1	[c]	0	0	1	.	1	[c]	1	0	1	1	[c]	1	0	1	[c]	1	1	1	[c]	[c]	[c]	0	1	1
47	0	0	1	[c]	0	0	0	0	1	[c]	1	0	.	1	[c]	1	0	0	[c]	1	1	1	[c]	[c]	[c]	0	1	1
48	.	.	1	[c]	[d]	0	1	.	1	[c]	[d]	0	1	1	[c]	[d]	0	1	[c]	[d]	1	1	[c]	[c]	[c]	[d]	[d]	1
49	0	1	0	[c]	0	0	1	0	.	[c]	1	0	1	0	[c]	1	1	1	[c]	1	0	1	[c]	[c]	[c]	0	1	1
50	.	.	1	1	1	0	1	.	1	1	1	0	1	1	1	1	0	1	1	1	1	0	.	0	1	0	1	1
51	.	.	1	0	0	.	0	.	1	0	0	0	0	1	0	0	0	0	1	1	1	1	.	0	1	0	1	0
52	.	0	1	[c]	0	0	0	0	1	[c]	0	0	0	1	[c]	0	0	1	[c]	1	1	1	[c]	[c]	[c]	.	1	1
53	.	0	1	[c]	1	0	1	0	1	[c]	1	0	1	1	[c]	1	1	1	[c]	1	1	0	[c]	[c]	[c]	0	1	1

Note. A maximum of 28 pairwise comparisons were conducted for each item. Statistical significance was assessed using $\alpha = .01$. Practical significance was assessed using a difference of .05. A value of 1 indicates the test was practically significant whereas a value of 0 indicates the test was not practically significant.

Calculation method 1 = INSPECT; Calculation method 2 = INSPECT2; Calculation method 3 = MIXTURE; Calculation method 4 = MIXTURE 2; Calculation method 5 = NT10; Calculation method 6 = NT20; Calculation method 7 = NT30; Calculation method 8 = RSPEED.

- = indicates the contrast tests were not conducted because the omnibus test was not significant.

. = indicates the contrast tests was not statistically significant.

[a] = indicates the contrast test was not conducted because the SB index calculated using the MIXTURE method was not included as a main effect.

[b] = indicates the contrast test was not conducted because the SB index calculated using the MIXTURE2 method was not included as a main effect.

[c] = indicates the contrast test was not conducted because the SB index calculated using the NT10 method was not included as a main effect.

[d] = indicates the contrast test was not conducted because the SB index calculated using the NT20 method was not included as a main effect.

[e] = indicates the contrast test was not conducted because the SB index calculated using the NT30 method was not included as a main effect.

Table 17

Total and Average Proportion of Statistically and Practically Significant Pairwise Comparisons, by Threshold Calculation Method

Threshold Calculation methods		*N*	Statistical significance		Practical significance	
			Count	Proportion	Count	Proportion
1. INSPECT	2	52	27	0.52	4	0.08
	3	51	19	0.37	6	0.12
	4	50	49	0.98	48	0.96
	5	16	15	0.94	5	0.31
	6	46	45	0.98	12	0.26
	7	51	40	0.78	2	0.04
	8	52	49	0.94	40	0.77
2. INSPECT2	3	51	27	0.53	8	0.16
	4	50	49	0.98	48	0.96
	5	16	16	1.00	10	0.63
	6	46	46	1.00	25	0.54
	7	51	48	0.94	14	0.27
	8	52	49	0.94	40	0.77
3. MIXTURE	4	49	48	0.98	46	0.94
	5	16	16	1.00	6	0.38
	6	46	46	1.00	19	0.41
	7	50	40	0.80	11	0.22
	8	51	48	0.94	39	0.76
4. MIXTURE2	5	16	16	1.00	16	1.00
	6	45	45	1.00	44	0.98
	7	49	48	0.98	47	0.96
	8	50	47	0.94	41	0.82
5. NT10	6	16	2	0.13	0	0.00
	7	16	14	0.88	1	0.06
	8	16	16	1.00	16	1.00
6. NT20	7	46	37	0.80	0	0.00
	8	46	46	1.00	43	0.93
7. NT30	8	51	49	0.96	45	0.88

Note. A maximum of 28 pairwise comparisons were conducted for each item. Statistical significance was assessed using $\alpha = .01$. Practical significance was assessed using a difference of .05.
Calculation method 1 = INSPECT; Calculation method 2 = INSPECT2; Calculation method 3 = MIXTURE; Calculation method 4 = MIXTURE 2; Calculation method 5 = NT10; Calculation method 6 = NT20; Calculation method 7 = NT30; Calculation method 8 = RSPEED; *N* = the total number of pairwise comparisons conducted; Count = the total number of comparisons that were either statistically or statistically and practically significant; Proportion = the proportion of comparisons that were either statistically or statistically and practically significant.

Table 18

Descriptive Statistics and Correlations of RTE, by Threshold Calculation Method

RTE[a]	Mean	*SD*	Median	Min	Max	Skew	Kurt	α
$RTE_{INSPECT}$	0.95	0.13	1.00	0.06	1.00	-4.70	24.52	0.963
$RTE_{INSPECT2}$	0.94	0.15	0.98	0.02	1.00	-3.99	17.75	0.967
$RTE_{MIXTURE}$	0.94	0.14	0.98	0.06	1.00	-4.07	18.80	0.962
$RTE_{MIXTURE2}$	0.81	0.21	0.88	0.00	1.00	-1.73	3.00	0.950
RTE_{NT10}	1.00	0.01	1.00	0.94	1.00	-6.79	50.25	0.447
RTE_{NT20}	0.99	0.04	1.00	0.49	1.00	-8.48	87.18	0.911
RTE_{NT30}	0.97	0.10	1.00	0.13	1.00	-6.04	39.92	0.969
RTE_{RSPEED}	0.78	0.19	0.83	0.02	1.00	-1.49	2.45	0.942

	$RTE_{INSPECT}$	$RTE_{INSPECT2}$	$RTE_{MIXTURE}$	$RTE_{MIXTURE2}$	RTE_{NT10}	RTE_{NT20}	RTE_{NT30}	RTE_{RSPEED}
$RTE_{INSPECT}$	1.00							
$RTE_{INSPECT2}$	0.98	1.00						
$RTE_{MIXTURE}$	0.98	0.99	1.00					
$RTE_{MIXTURE2}$	0.76	0.84	0.83	1.00				
RTE_{NT10}	0.21	0.22	0.22	0.17	1.00			
RTE_{NT20}	0.78	0.73	0.73	0.51	0.34	1.00		
RTE_{NT30}	0.95	0.89	0.91	0.64	0.23	0.88	1.00	
RTE_{RSPEED}	0.74	0.80	0.81	0.96	0.14	0.49	0.61	1.00

Note. All correlations were statistically significant ($p < .01$). RTE = Response Time Effort; $RTE_{INSPECT}$ = RTE visual inspection; $RTE_{INSPECT2}$ = visual inspection with information; $RTE_{MIXTURE}$ = lognormal mixture modeling; $RTE_{MIXTURE2}$ = lognormal mixture modeling with information; RTE_{NT10} = 10% normative threshold; RTE_{NT20} = 20% normative threshold; RTE_{NT30} = 30% normative threshold; RTE_{RSPEED} = reading speed; SD = standard deviation; Min = minimum; Max = maximum; Skew = skewness; Kurt = kurtosis; α = coefficient alpha.

[a] $N = 568$.

Table 19

Omnibus Test Results for the Generalized Estimating Equations Analyzing the Relationship between the Logit of RTE and Various Respondent Characteristics

Respondent characteristic	Effect	df_1	df_2	F	p
Gender	Calculation Method	7	3962	84.84	<.0001
	Gender	1	566	4.46	0.035
	Gender*Calculation Method	7	3962	2.58	0.012
Walk-in	Calculation Method	7	2317	77.87	<.0001
	Walkin	1	331	7.66	0.006
	Walkin*Calculation Method	7	2317	3.10	0.003
Effort[a]	Calculation Method	6	3396	140.25	<.0001
	Effort	1	566	23.26	<.0001
	Effort*Calculation Method	6	3396	8.05	<.0001
SAT-M[a]	Calculation Method	6	2700	110.42	<.0001
	SAT-M	1	450	2.26	0.134
	SAT-M*Calculation Method	6	2700	1.86	0.084
SAT-CR[b]	Calculation Method	5	2250	87.74	<.0001
	SAT-CR	1	450	3.92	0.048
	SAT-CR*Calculation Method	5	2250	1.42	0.212
Individual Consistency Index[a]	Calculation Method	6	3276	155.28	<.0001
	Index	1	546	16.34	<.0001
	Index*Calculation Method	6	3276	4.94	<.0001
Open-ended response length[c]	Calculation Method	7	2648	86.36	<.0001
	Length	1	2648	28.34	<.0001
	Length*Calculation Method	7	2648	2.17	0.034

Note. Statistical significance was assessed using $\alpha = .01$.

[a] Indicates RTE calculated using NT10 were excluded from the model.

[b] Indicates RTE calculated using NT10 and NT20 were excluded from the model.

[c] Indicates the estimated model ignored the within subject correlations introduced by the repeated measures nature of the RTE scores

Table 20

Descriptive Statistics of RTE, by Gender

Gender	RTE scores	Mean	*SD*	Median	Min	Max	Skew	Kurtosis
Female[a]								
	$RTE_{INSPECT}$	0.97	0.09	1.00	0.06	1.00	-6.43	52.03
	$RTE_{INSPECT2}$	0.95	0.11	0.98	0.02	1.00	-4.92	31.97
	$RTE_{MIXTURE}$	0.95	0.10	0.98	0.06	1.00	-4.89	32.06
	$RTE_{MIXTURE2}$	0.82	0.18	0.88	0.00	1.00	-1.58	2.93
	RTE_{NT10}	1.00	0.01	1.00	0.94	1.00	-6.85	49.80
	RTE_{NT20}	1.00	0.02	1.00	0.76	1.00	-8.59	97.42
	RTE_{NT30}	0.99	0.06	1.00	0.25	1.00	-9.56	107.61
	RTE_{RSPEED}	0.78	0.17	0.78	0.04	1.00	-1.28	2.08
Male[b]								
	$RTE_{INSPECT}$	0.94	0.15	1.00	0.08	1.00	-3.99	16.88
	$RTE_{INSPECT2}$	0.92	0.17	0.98	0.04	1.00	-3.47	12.64
	$RTE_{MIXTURE}$	0.93	0.16	0.98	0.06	1.00	-3.58	13.66
	$RTE_{MIXTURE2}$	0.81	0.22	0.88	0.00	1.00	-1.72	2.65
	RTE_{NT10}	1.00	0.00	1.00	0.96	1.00	-6.17	39.48
	RTE_{NT20}	0.99	0.05	1.00	0.49	1.00	-7.09	58.72
	RTE_{NT30}	0.96	0.12	1.00	0.13	1.00	-4.96	26.21
	RTE_{RSPEED}	0.78	0.21	0.85	0.02	1.00	-1.54	2.29

Note. RTE = Response Time Effort; $RTE_{INSPECT}$ = RTE visual inspection; $RTE_{INSPECT2}$ = visual inspection with information; $RTE_{MIXTURE}$ = lognormal mixture modeling; $RTE_{MIXTURE2}$ = lognormal mixture modeling with information; RTE_{NT10} = 10% normative threshold; RTE_{NT20} = 20% normative threshold; RTE_{NT30} = 30% normative threshold; RTE_{RSPEED} = reading speed; SD = standard deviation; Min = minimum; Max = maximum; Skew = skewness.

[a] $N = 243$.
[b] $N = 325$.

Table 21

Correlations between RTE and Respondent Characteristics, by Threshold Calculation Method

RTE	Female[a]	Walk-in[b]	Effort[a]	SAT-M[c]	SAT-CR[c]	Individual consistency index[d]	Open-ended item length[b]
$RTE_{INSPECT}$	0.10	-0.14*	0.25*	-0.07	-0.07	0.18*	0.17*
$RTE_{INSPECT2}$	0.09	-0.15*	0.26*	-0.08	-0.10	0.18*	0.19*
$RTE_{MIXTURE}$	0.08	-0.15*	0.27*	-0.07	-0.09	0.19*	0.19*
$RTE_{MIXTURE2}$	0.04	-0.09	0.21*	-0.06	-0.15*	0.17*	0.14*
RTE_{NT10}	-0.02	-0.06	0.10	0.07	0.05	0.05	0.12
RTE_{NT20}	0.09	-0.13	0.16*	-0.06	-0.04	0.10	0.12
RTE_{NT30}	0.11*	-0.13	0.22*	-0.07	-0.05	0.14*	0.14
RTE_{RSPEED}	0.01	-0.08	0.20*	-0.03	-0.15*	0.14*	0.13

Note. RTE = Response Time Effort; INSPECT = visual inspection; INSPECT2 = visual inspection with information; MIXTURE = lognormal mixture modeling; MIXTURE2 = lognormal mixture modeling with information; NT10 = 10% normative threshold; NT20 = 20% normative threshold; NT30 = 30% normative threshold; RSPEED = reading speed; Effort = SOS Effort subscore; SAT-M = SAT – Mathematics; SAT-CR = SAT Critical Reading.

[a] $N = 568$.
[b] $N = 333$.
[c] $N = 452$.
[d] $N = 548$.

*$p < 0.01$.

Table 22

Descriptive Statistics of RTE, by Makeup Testing Attendance Status

Attendance status	RTE	Mean	*SD*	Median	Min	Max	Skew	Kurtosis
Walk-in = 0[a]								
	$RTE_{INSPECT}$	0.96	0.12	1.00	0.08	1.00	-5.54	34.16
	$RTE_{INSPECT2}$	0.95	0.13	0.98	0.04	1.00	-4.85	27.03
	$RTE_{MIXTURE}$	0.94	0.12	0.98	0.08	1.00	-4.95	28.54
	$RTE_{MIXTURE2}$	0.81	0.18	0.86	0.00	1.00	-1.69	3.55
	RTE_{NT10}	1.00	0.00	1.00	0.96	1.00	-6.31	41.93
	RTE_{NT20}	0.99	0.02	1.00	0.79	1.00	-5.44	39.37
	RTE_{NT30}	0.98	0.09	1.00	0.26	1.00	-6.89	51.32
	RTE_{RSPEED}	0.78	0.17	0.81	0.02	1.00	-1.46	3.26
Walk-in = 1[b]								
	$RTE_{INSPECT}$	0.92	0.19	1.00	0.06	1.00	-3.11	9.97
	$RTE_{INSPECT2}$	0.89	0.21	0.98	0.02	1.00	-2.64	6.72
	$RTE_{MIXTURE}$	0.89	0.20	0.98	0.06	1.00	-2.65	6.99
	$RTE_{MIXTURE2}$	0.77	0.26	0.88	0.00	1.00	-1.50	1.43
	RTE_{NT10}	1.00	0.01	1.00	0.94	1.00	-5.18	28.46
	RTE_{NT20}	0.98	0.07	1.00	0.49	1.00	-5.24	29.72
	RTE_{NT30}	0.95	0.15	1.00	0.13	1.00	-4.06	17.06
	RTE_{RSPEED}	0.75	0.24	0.83	0.02	1.00	-1.34	1.13

Note. RTE = Response Time Effort; INSPECT = visual inspection; INSPECT2 = visual inspection with information; MIXTURE = lognormal mixture modeling; MIXTURE2 = lognormal mixture modeling with information; NT10 = 10% normative threshold; NT20 = 20% normative threshold; NT30 = 30% normative threshold; RSPEED = reading speed; Min = minimum; Max = maximum; Skew = skewness.

[a] $N = 151$.
[b] $N = 182$.

Table 23

Descriptive Statistics of the Respondent Characteristics

Respondent Characteristics	Mean	*SD*	Median	Min	Max	Skew	Kurtosis	*N*
Effort	18.59	3.79	19.00	5.00	25.00	-0.48	0.47	568
SAT-CR	572.79	69.79	580.00	350.00	760.00	-0.13	-0.17	452
SAT-M	565.80	67.44	560.00	310.00	740.00	-0.06	0.07	452
Individual Consistency Index	0.30	0.43	0.33	-0.78	1.00	-0.29	-0.93	548
Open-ended length item	116.21	94.69	91.00	0.00	530.00	1.46	2.35	333

Note. Effort = SOS Effort subscore; SAT-M = SAT – Mathematics; SAT-CR = SAT Critical Reading; Min = minimum; Max = maximum; Skew = skewness.

Table 24

Results of the GEE Examining the Relationship between Makeup Testing Attendance Status and the Logit of RTE and Simple Slopes Examining the Relationship between Makeup Testing Attendance Status and the Logit of RTE, by Threshold Calculation Method

Effect	*b*	*SE*	*df*	*t*	*p*
Intercept, β_0	1.26*	0.07	331	17.38	<0.001
INSPECT, β_1	1.91*	0.18	2317	10.61	<0.001
INSPECT2, β_2	1.59*	0.14	2317	11.59	<0.001
MIXTURE, β_3	1.57*	0.12	2317	12.73	<0.001
MIXTURE2, β_4	0.21*	0.03	2317	7.22	<0.001
NT10, β_5	5.82*	0.43	2317	13.56	<0.001
NT20, β_6	3.73*	0.22	2317	17.22	<0.001
NT30, β_7	2.53*	0.26	2317	9.74	<0.001
Walk-in, β_8	-0.17	0.13	331	-1.37	0.173
Walk-in*INSPECT, β_9	-0.61*	0.22	2317	-2.81	0.005
Walk-in*INSPECT2, β_{10}	-0.55*	0.17	2317	-3.27	0.001
Walk-in*MIXTURE, β_{11}	-0.52*	0.15	2317	-3.36	0.001
Walk-in*MIXTURE2, β_{12}	-0.07	0.04	2317	-1.71	0.088
Walk-in*NT10, β_{13}	-0.50	0.57	2317	-0.88	0.379
Walk-in*NT20, β_{14}	-0.93*	0.33	2317	-2.85	0.004
Walk-in*NT30, β_{15}	-0.75	0.32	2317	-2.36	0.018

Threshold calculation method	Simple slopes	95% CI	*SE*	*t*	*p*
INSPECT	-0.79*	[-1.38, -0.20]	0.30	-2.64	0.009
INSPECT2	-0.73*	[-1.23, -0.22]	0.26	-2.83	0.005
MIXTURE	-0.69*	[-1.17, -0.21]	0.24	-2.84	0.005
MIXTURE2	-0.24	[-0.54, 0.05]	0.15	-1.62	0.105
NT10	-0.67	[-1.79, 0.45]	0.57	-1.18	0.238
NT20	-1.11*	[-1.86, -0.35]	0.38	-2.89	0.004
NT30	-0.92	[-1.68, -0.17]	0.38	-2.41	0.017
RSPEED	-0.17	[-0.42, 0.08]	0.13	-1.37	0.173

Note. The threshold calculation method RSPEED was used as a reference variable.

*$p < .01$.

Table 25

Pairwise Comparison Results Examining the Relationship between Makeup Testing Attendance Status and the Logit of RTE, by Threshold Calculation Method

Threshold Calculation methods		Estimate	*SE*	*df*	*t*	*p*
1. INSPECT	2	-0.06	0.06	2317	-0.94	0.345
	3	-0.10	0.08	2317	-1.29	0.198
	4	-0.54*	0.20	2317	-2.67	0.008
	5	-0.12	0.59	2317	-0.19	0.846
	6	0.32	0.22	2317	1.43	0.151
	7	0.14	0.14	2317	0.98	0.328
	8	-0.61*	0.22	2317	-2.81	0.005
2. INSPECT2	3	-0.04	0.04	2317	-1.00	0.315
	4	-0.48*	0.15	2317	-3.17	0.002
	5	-0.05	0.58	2317	-0.09	0.925
	6	0.38	0.24	2317	1.56	0.119
	7	0.20	0.19	2317	1.05	0.295
	8	-0.55*	0.17	2317	-3.27	0.001
3. MIXTURE	4	-0.45*	0.14	2317	-3.24	0.001
	5	-0.02	0.57	2317	-0.03	0.978
	6	0.42	0.24	2317	1.72	0.085
	7	0.23	0.19	2317	1.23	0.220
	8	-0.52*	0.15	2317	-3.36	0.001
4. MIXTURE2	5	0.43	0.57	2317	0.76	0.448
	6	0.86*	0.32	2317	2.71	0.007
	7	0.68	0.31	2317	2.22	0.026
	8	-0.07	0.04	2317	-1.71	0.088
5. NT10	6	0.43	0.57	2317	0.76	0.446
	7	0.25	0.62	2317	0.40	0.687
	8	-0.50	0.57	2317	-0.88	0.379
6. NT20	7	-0.18	0.18	2317	-1.00	0.320
	8	-0.93*	0.33	2317	-2.85	0.004
7. NT30	8	-0.75	0.32	2317	-2.36	0.018

Note. Calculation method 1 = INSPECT; Calculation method 2 = INSPECT2; Calculation method 3 = MIXTURE; Calculation method 4 = MIXTURE 2; Calculation method 5 = NT10; Calculation method 6 = NT20; Calculation method 7 = NT30; Calculation method 8 = RSPEED.

*$p < .01$.

Table 26

Results of the GEE Examining the Relationship between Effort and the Logit of RTE and Simple Slopes Examining the Relationship between Effort and the Logit of RTE, by Threshold Calculation Method

Effect	*b*	*SE*	*df*	*t*	*p*
Intercept, β_0	1.26*	0.05	566	26.75	<0.001
INSPECT, β_1	1.91*	0.09	3396	20.56	<0.001
INSPECT2, β_2	1.54*	0.07	3396	21.58	<0.001
MIXTURE, β_3	1.56*	0.07	3396	23.78	<0.001
MIXTURE2, β_4	0.22*	0.02	3396	12.69	<0.001
NT20, β_5	3.48*	0.18	3396	19.68	<0.001
NT30, β_6	2.53*	0.14	3396	18.30	<0.001
Effort, β_7	0.05*	0.01	566	3.15	0.002
Effort*INSPECT, β_8	0.10*	0.02	3396	5.70	<0.001
Effort*INSPECT2, β_9	0.09*	0.01	3396	6.00	<0.001
Effort*MIXTURE, β_{10}	0.08*	0.01	3396	6.32	<0.001
Effort*MIXTURE2, β_{11}	0.01*	0.00	3396	3.40	0.001
Effort*NT20, β_{12}	0.06	0.03	3396	1.83	0.068
Effort*NT30, β_{13}	0.12*	0.02	3396	5.16	<0.001

Threshold calculation method	Simple slopes	95% CI	*SE*	*t*	*p*
INSPECT	0.15*	[0.09, 0.20]	0.03	5.32	<0.001
INSPECT2	0.13*	[0.08, 0.18]	0.02	5.28	<0.001
MIXTURE	0.13*	[0.08, 0.18]	0.02	5.44	<0.001
MIXTURE2	0.06*	[0.03, 0.09]	0.02	3.60	<0.001
NT20	0.11*	[0.03, 0.18]	0.04	2.67	0.008
NT30	0.16*	[0.10, 0.23]	0.03	5.21	<0.001
RSPEED	0.05*	[0.02, 0.07]	0.01	3.15	0.002

Note. The threshold calculation method RSPEED was used as a reference variable.

*$p < .01$.

Table 27

Pairwise Comparison Results Examining the Relationship between Effort and the Logit of RTE, by Threshold Calculation Method

Threshold Calculation methods		Estimate	*SE*	*df*	*t*	*p*
1. INSPECT	2	0.02*	0.01	3396	2.67	0.008
	3	0.02*	0.01	3396	2.85	0.004
	4	0.09*	0.02	3396	5.26	<0.001
	6	0.04	0.03	3396	1.41	0.159
	7	-0.02	0.01	3396	-1.40	0.163
	8	0.10*	0.02	3396	5.70	<0.001
2. INSPECT2	3	0.00	0.00	3396	0.26	0.791
	4	0.07*	0.01	3396	5.63	<0.001
	6	0.03	0.03	3396	0.79	0.428
	7	-0.03*	0.02	3396	-2.05	0.040
	8	0.09*	0.01	3396	6.00	<0.001
3. MIXTURE	4	0.07*	0.01	3396	5.85	<0.001
	6	0.02	0.03	3396	0.79	0.431
	7	-0.03	0.01	3396	-2.33	0.020
	8	0.08*	0.01	3396	6.32	<0.001
4. MIXTURE2	6	-0.05	0.03	3396	-1.39	0.163
	7	-0.10*	0.02	3396	-4.65	<0.001
	8	0.01*	0.00	3396	3.40	0.001
6. NT20	7	-0.06	0.02	3396	-2.50	0.012
	8	0.06	0.03	3396	1.83	0.068
7. NT30	8	0.12*	0.02	3396	5.16	<0.001

Note. Calculation method 1 = INSPECT; Calculation method 2 = INSPECT2; Calculation method 3 = MIXTURE; Calculation method 4 = MIXTURE 2; Calculation method 6 = NT20; Calculation method 7 = NT30; Calculation method 8 = RSPEED.

*$p < .01$.

Table 28

Results of the GEE Examining the Relationship between the Individual Consistency Index and the Logit of RTE and Simple Slopes Examining the Relationship between the Individual Consistency Index and the Logit of RTE, by Threshold Calculation Method

Effect	*b*	*SE*	*df*	*t*	*p*
Intercept, β_0	1.33*	0.04	546	29.74	<0.001
INSPECT, β_1	2.02*	0.09	3276	23.16	<0.001
INSPECT2, β_2	1.63*	0.07	3276	24.39	<0.001
MIXTURE, β_3	1.63*	0.06	3276	26.78	<0.001
MIXTURE2, β_4	0.24*	0.02	3276	13.37	<0.001
NT20, β_5	3.60*	0.16	3276	22.83	<0.001
NT30, β_6	2.68*	0.13	3276	20.91	<0.001
Index, β_7	0.26	0.11	546	2.45	0.015
Index*INSPECT, β_8	0.73*	0.17	3276	4.41	<0.001
Index*INSPECT2, β_9	0.61*	0.14	3276	4.31	<0.001
Index*MIXTURE, β_{10}	0.59*	0.12	3276	4.74	<0.001
Index*MIXTURE2, β_{11}	0.17*	0.04	3276	4.08	<0.001
Index*NT20, β_{12}	0.42	0.34	3276	1.25	0.210
Index*NT30, β_{13}	0.75*	0.23	3276	3.33	0.001

Threshold calculation method	Simple slopes	95% CI	*SE*	*t*	*p*
INSPECT	0.99*	[0.56, 1.41]	0.22	4.56	<0.001
INSPECT2	0.86*	[0.46, 1.26]	0.20	4.25	<0.001
MIXTURE	0.84*	[0.48, 1.21]	0.19	4.50	<0.001
MIXTURE2	0.43*	[0.18, 0.67]	0.12	3.42	0.001
NT20	0.68	[-0.02, 1.38]	0.36	1.91	0.057
NT30	1.01*	[0.50, 1.52]	0.26	3.92	<0.001
RSPEED	0.26	[0.05, 0.47]	0.11	2.45	0.015

Note. The threshold calculation method RSPEED was used as a reference variable.

*$p < .01$.

Table 29

Pairwise Comparison Results Examining the Relationship between the Individual Consistency Index and the Logit of RTE, by Threshold Calculation Method

Threshold Calculation methods		Estimate	*SE*	*df*	*t*	*p*
1. INSPECT	2	0.13	0.06	3276	2.10	0.036
	3	0.15	0.07	3276	2.17	0.030
	4	0.56*	0.15	3276	3.68	<0.001
	6	0.31	0.28	3276	1.09	0.275
	7	-0.02	0.13	3276	-0.17	0.867
	8	0.73*	0.17	3276	4.41	<0.001
2. INSPECT2	3	0.02	0.05	3276	0.44	0.660
	4	0.44*	0.12	3276	3.51	<0.001
	6	0.18	0.31	3276	0.59	0.558
	7	-0.15	0.16	3276	-0.89	0.373
	8	0.61*	0.14	3276	4.31	<0.001
3. MIXTURE	4	0.42*	0.11	3276	3.82	<0.001
	6	0.16	0.31	3276	0.53	0.597
	7	-0.17	0.16	3276	-1.03	0.304
	8	0.59*	0.12	3276	4.74	<0.001
4. MIXTURE2	6	-0.25	0.34	3276	-0.75	0.451
	7	-0.58*	0.22	3276	-2.65	0.008
	8	0.17*	0.04	3276	4.08	<0.001
6. NT20	7	-0.33	0.20	3276	-1.65	0.098
	8	0.42	0.34	3276	1.25	0.210
7. NT30	8	0.75*	0.23	3276	3.33	0.001

Note. Calculation method 1 = INSPECT; Calculation method 2 = INSPECT2; Calculation method 3 = MIXTURE; Calculation method 4 = MIXTURE 2; Calculation method 6 = NT20; Calculation method 7 = NT30; Calculation method 8 = RSPEED.

*$p < .01$.

Table 30

Descriptive Statistics and Correlations of RTF, by Threshold Calculation Method

RTF scores	Mean	*SD*	Median	Min	Max	Skew	Kurt
$RTF_{INSPECT}$[a]	0.95	0.02	0.95	0.91	0.99	-0.09	0.05
$RTF_{INSPECT2}$[a]	0.94	0.02	0.94	0.87	0.99	-0.64	0.70
$RTF_{MIXTURE}$[b]	0.94	0.04	0.95	0.78	0.99	-1.97	4.28
$RTF_{MIXTURE2}$[c]	0.81	0.08	0.85	0.57	0.97	-1.05	0.74
RTF_{NT10}[a]	1.00	0.00	1.00	0.99	1.00	-1.96	4.43
RTF_{NT20}[a]	0.99	0.01	0.99	0.97	1.00	-0.59	-0.50
RTF_{NT30}[a]	0.97	0.01	0.97	0.95	1.00	-0.07	-0.17
RTF_{RSPEED}[a]	0.78	0.16	0.78	0.22	0.99	-1.11	1.47

	$RTF_{INSPECT}$	$RTF_{INSPECT2}$	$RTF_{MIXTURE}$	$RTF_{MIXTURE2}$	RTF_{NT10}	RTF_{NT20}	RTF_{NT30}	RTF_{RSPEED}
$RTF_{INSPECT}$[a]	1.00							
$RTF_{INSPECT2}$[a]	0.70*	1.00						
$RTF_{MIXTURE}$[b]	0.65*	0.67*	1.00					
$RTF_{MIXTURE2}$[c]	0.10	0.30	0.31	1.00				
RTF_{NT10}[a]	-0.21	-0.21	-0.24	-0.32	1.00			
RTF_{NT20}[a]	0.37*	0.12	0.20	-0.31	0.30	1.00		
RTF_{NT30}[a]	0.50*	0.22	0.33	-0.25	0.21	0.86*	1.00	
RTF_{RSPEED}[a]	0.43*	0.27	0.31	-0.07	-0.09	0.40*	0.50*	1.00

Note. All correlations were statistically significant ($p < .01$). RTF = Response Time Fidelity; $RTF_{INSPECT}$ = RTF visual inspection; $RTF_{INSPECT2}$ = visual inspection with information; $RTF_{MIXTURE}$ = lognormal mixture modeling; $RTF_{MIXTURE2}$ = lognormal mixture modeling with information; RTF_{NT10} = 10% normative threshold; RTF_{NT20} = 20% normative threshold; RTF_{NT30} = 30% normative threshold; RTF_{RSPEED} = reading speed; SD = standard deviation; Min = minimum; Max = maximum; Skew = skewness; Kurt = kurtosis.

[a] $N = 53$.
[b] $N = 52$.
[c] $N = 50$.
*$p < .01$.

Table 31

Omnibus Test Results for GEEs Analyzing the Relationship between MFLS Item Characteristics and the Logit of RTF

Item characteristic	Effect	df_1	df_2	F	p
Item position	Calculation Method	7	353	163.01	<.0001
	Position	1	51	5.15	0.028
	Position*Calculation Method	7	353	6.40	<.0001
Item length	Calculation Method	7	353	152.34	<.0001
	Length	1	51	26.17	<.0001
	Length*Calculation Method	7	353	44.59	<.0001

Note. Statistical significance was assessed using $\alpha = .01$.

Table 32

Descriptive Statistics of MFLS Item Characteristics

Item Characteristics[a]	Mean	*SD*	Median	Min	Max	Skew	Kurtosis
Item position	27.00	15.44	27.00	1.00	53.00	0.00	-1.20
Item length	11.13	3.95	11.00	4.00	21.00	0.61	0.45

Note. Min = minimum; Max = maximum; Skew = skewness.
[a] $N = 53$.

Table 33

Correlations Between RTF and MFLS Item Characteristics, by Threshold Calculation Method

RTF	Item position	Item length
$RTF_{INSPECT}$[a]	-0.28	-0.42*
$RTF_{INSPECT2}$[a]	0.03	-0.16
$RTF_{MIXTURE}$[b]	-0.02	-0.29
$RTF_{MIXTURE2}$[c]	0.22	0.31
RTF_{NT10}[a]	-0.01	-0.05
RTF_{NT20}[a]	-0.49*	-0.65*
RTF_{NT30}[a]	-0.49*	-0.71*
RTF_{RSPEED}[a]	-0.34	-0.87*

Note. RTF = Response Time Fidelity; $RTF_{INSPECT}$ = RTF visual inspection; $RTF_{INSPECT2}$ = visual inspection with information; $RTF_{MIXTURE}$ = lognormal mixture modeling; $RTF_{MIXTURE2}$ = lognormal mixture modeling with information; RTF_{NT10} = 10% normative threshold; RTF_{NT20} = 20% normative threshold; RTF_{NT30} = 30% normative threshold; RTF_{RSPEED} = reading speed.

[a] $N = 53$.
[b] $N = 52$.
[c] $N = 50$.
*$p < 0.01$.

Table 34

Results of the GEE Examining the Relationship between Item Position and the Logit of RTF and Simple Slopes Examining the Relationship between Item Position and the Logit of RTF, by Threshold Calculation Method

Effect	*b*	*SE*	*df*	*t*	*p*
Intercept, β_0	1.29*	0.12	51	10.77	<.001
INSPECT, β_1	1.73*	0.11	353	15.65	<.001
INSPECT2, β_2	1.40*	0.11	353	12.39	<.001
MIXTURE, β_3	1.42*	0.13	353	11.26	<.001
MIXTURE2, β_4	0.17	0.14	353	1.24	0.214
NT10, β_5	5.65*	0.25	353	22.26	<.001
NT20, β_6	3.44*	0.13	353	27.48	<.001
NT30, β_7	2.31*	0.11	353	21.19	<.001
Position, β_8	-0.02*	0.01	51	-3.02	0.004
Position*INSPECT, β_9	0.01	0.01	353	2.28	0.023
Position*INSPECT2, β_{10}	0.02*	0.01	353	3.48	0.001
Position*MIXTURE, β_{11}	0.02*	0.01	353	2.74	0.007
Position*MIXTURE2, β_{12}	0.03*	0.01	353	3.56	<.001
Position*NT10, β_{13}	0.02	0.01	353	1.57	0.118
Position*NT20, β_{14}	0.00	0.01	353	-0.51	0.607
Position*NT30, β_{15}	0.01	0.01	353	1.31	0.191

Threshold calculation method	Simple slopes	95% CI	*SE*	*t*	*p*
INSPECT	-0.01	[-0.01, 0.00]	0.00	-1.93	0.059
INSPECT2	0.00[a]	--	--	--	--
MIXTURE	0.00	[-0.01, 0.00]	0.01	-0.10	0.924
MIXTURE2	0.01	[0.00, 0.02]	0.00	1.94	0.058
NT10	0.00[a]	--	--	--	--
NT20	-0.02*	[-0.03, -0.01]	0.01	-4.87	<.0001
NT30	-0.01*	[-0.02, -0.01]	0.00	-4.22	<.0001
RSPEED	-0.02*	[-0.04, -0.01]	0.01	-3.02	0.004

Note. The threshold calculation method RSPEED was used as a reference variable.
[a] These simple slopes were calculated by hand. The standard errors and associated tests of significance were not calculated due to convergence problems in SAS.
*$p < .01$.

Table 35

Pairwise Comparison Results Examining the Relationship between Item Position and the Logit of RTF, by Threshold Calculation Method

Threshold Calculation methods		Estimate	*SE*	*df*	*t*	*p*
1. INSPECT	2	-0.008*	0.002	353	-3.24	0.001
	3	-0.006	0.005	353	-1.11	0.268
	4	-0.015*	0.005	353	-2.75	0.006
	5	-0.008	0.014	353	-0.56	0.576
	6	0.018*	0.005	353	3.46	0.001
	7	0.007	0.004	353	1.90	0.059
	8	0.015	0.007	353	2.28	0.023
2. INSPECT2	3	0.002	0.005	353	0.40	0.692
	4	-0.007	0.005	353	-1.31	0.191
	5	0.000	0.014	353	-0.01	0.991
	6	0.026*	0.006	353	4.66	<0.001
	7	0.015*	0.004	353	3.55	0.000
	8	0.023*	0.006	353	3.48	0.001
3. MIXTURE	4	-0.009	0.007	353	-1.24	0.217
	5	-0.002	0.015	353	-0.13	0.898
	6	0.024*	0.007	353	3.40	0.001
	7	0.013	0.006	353	2.10	0.037
	8	0.021*	0.008	353	2.74	0.007
4. MIXTURE2	5	0.007	0.015	353	0.48	0.635
	6	0.033*	0.007	353	4.42	<0.001
	7	0.022*	0.006	353	3.62	0.000
	8	0.030*	0.008	353	3.56	0.000
5. NT10	6	0.026	0.012	353	2.17	0.031
	7	0.015	0.012	353	1.22	0.223
	8	0.023	0.014	353	1.57	0.118
6. NT20	7	-0.011*	0.003	353	-3.89	0.000
	8	-0.003	0.007	353	-0.51	0.607
7. NT30	8	0.008	0.006	353	1.31	0.191

Note. Calculation method 1 = INSPECT; Calculation method 2 = INSPECT2; Calculation method 3 = MIXTURE; Calculation method 4 = MIXTURE 2; Calculation method 5 = NT10; Calculation method 6 = NT20; Calculation method 7 = NT30; Calculation method 8 = RSPEED.

*$p < .01$.

Table 36

Results of the GEE Examining the Relationship between Item Length and the Logit of RTF and Simple Slopes Examining the Relationship between Item Length and the Logit of RTF, by Threshold Calculation Method

Effect	*b*	*SE*	*df*	*t*	*p*
Intercept, β_0	1.42*	0.06	51	22.05	<.0001
INSPECT, β_1	1.59*	0.07	353	22.16	<.0001
INSPECT2, β_2	1.26*	0.07	353	18.80	<.0001
MIXTURE, β_3	1.30*	0.10	353	12.70	<.0001
MIXTURE2, β_4	0.04	0.07	353	0.54	0.589
NT10, β_5	5.48*	0.23	353	24.30	<.0001
NT20, β_6	3.32*	0.12	353	28.00	<.0001
NT30, β_7	2.18*	0.09	353	25.46	<.0001
Length, β_8	-0.21*	0.02	51	-8.90	<.0001
Length*INSPECT, β_9	0.18*	0.02	353	7.17	<.0001
Length*INSPECT2, β_{10}	0.20*	0.03	353	7.28	<.0001
Length*MIXTURE, β_{11}	0.16*	0.03	353	5.14	<.0001
Length*MIXTURE2, β_{12}	0.26*	0.03	353	9.53	<.0001
Length*NT10, β_{13}	0.22*	0.04	353	5.64	<.0001
Length*NT20, β_{14}	0.10*	0.03	353	3.02	0.003
Length*NT30, β_{15}	0.14*	0.03	353	4.99	<.0001

Threshold calculation method	Simple slopes	95% CI	*SE*	*t*	*p*
INSPECT	-0.03*	[-0.05, -0.01]	0.01	-3.43	0.001
INSPECT2	-0.01	[-0.04, 0.01]	0.01	-0.97	0.339
MIXTURE	-0.05	[-0.09, -0.01]	0.02	-2.63	0.011
MIXTURE2	0.05	[0.01, 0.09]	0.02	2.33	0.024
NT10	0.01	[-0.07, 0.09]	0.04	0.23	0.817
NT20	-0.11*	[-0.15, -0.07]	0.02	-5.26	<.0001
NT30	-0.07*	[-0.09, -0.04]	0.01	-5.60	<.0001
RSPEED	-0.21*	[-0.26, -0.16]	0.02	-8.90	<.0001

Note. The threshold calculation method RSPEED was used as a reference variable.

*$p < .01$.

Table 37

Pairwise Comparison Results Examining the Relationship between Item Length and the Logit of RTF, by Threshold Calculation Method

Threshold Calculation methods		Estimate	*SE*	*df*	*t*	*p*
1. INSPECT	2	-0.022*	-0.008	353	-2.88	0.004
	3	0.014	-0.017	353	0.83	0.405
	4	-0.083*	-0.021	353	-3.91	0.000
	5	-0.043	-0.040	353	-1.09	0.277
	6	0.073*	-0.021	353	3.48	0.001
	7	0.034	-0.014	353	2.42	0.016
	8	0.177*	-0.025	353	7.17	0.000
2. INSPECT2	3	0.036	-0.017	353	2.18	0.030
	4	-0.060*	-0.022	353	-2.78	0.006
	5	-0.021	-0.042	353	-0.50	0.617
	6	0.095*	-0.022	353	4.24	0.000
	7	0.056*	-0.015	353	3.64	0.000
	8	0.199*	-0.027	353	7.28	0.000
3. MIXTURE	4	-0.096*	-0.026	353	-3.75	0.000
	5	-0.057	-0.043	353	-1.34	0.180
	6	0.059	-0.024	353	2.41	0.016
	7	0.020	-0.019	353	1.02	0.307
	8	0.163*	-0.032	353	5.14	0.000
4. MIXTURE2	5	0.039	-0.045	353	0.86	0.389
	6	0.156*	-0.029	353	5.40	0.000
	7	0.116*	-0.022	353	5.20	0.000
	8	0.259*	-0.027	353	9.53	0.000
5. NT10	6	0.116*	-0.031	353	3.81	0.000
	7	0.077	-0.034	353	2.26	0.024
	8	0.220*	-0.039	353	5.64	0.000
6. NT20	7	-0.039*	-0.011	353	-3.54	0.000
	8	0.104*	-0.034	353	3.02	0.003
7. NT30	8	0.143*	-0.029	353	4.99	0.000

Note. Calculation method 1 = INSPECT; Calculation method 2 = INSPECT2; Calculation method 3 = MIXTURE; Calculation method 4 = MIXTURE 2; Calculation method 5 = NT10; Calculation method 6 = NT20; Calculation method 7 = NT30; Calculation method 8 = RSPEED.

*$p < .01$.

Table 38

Simple Slopes (in Logits) from the GEEs Reflecting a Significant Interaction between an External Characteristic and Threshold Calculation Method, by Analysis

Analysis	External Characteristic	INSPECT	INSPECT2	MIXTURE	MIXTURE2	NT10	NT20	NT30	RSPEED
RTE									
	Walk-in	-0.79*	-0.73*	-0.69*	-0.24	-0.67	-1.11*	-0.92	-0.17
	Effort	0.15*	0.13*	0.13*	0.06*	--	0.11*	0.16*	0.05*
	Index	0.99*	0.86*	0.84*	0.43*	--	0.68	1.01*	0.26
RTF									
	Position	-0.01	0.00	0.00	0.01	0.00	-0.03*	-0.01*	-0.02*
	Length	-0.04*	-0.01	-0.05	0.05	0.01	-0.11*	-0.07*	-0.21*

Note. Dashed lines indicate the threshold method was not included in the model. RTE = Response Time Effort; RTF = Response Time Fidelity; INSPECT = visual inspection; INSPECT2 = visual inspection with information; MIXTURE = lognormal mixture modeling; MIXTURE2 = lognormal mixture modeling with information; NT10 = 10% normative threshold; NT20 = 20% normative threshold; NT30 = 30% normative threshold; RSPEED = reading speed.

*$p < .01$.

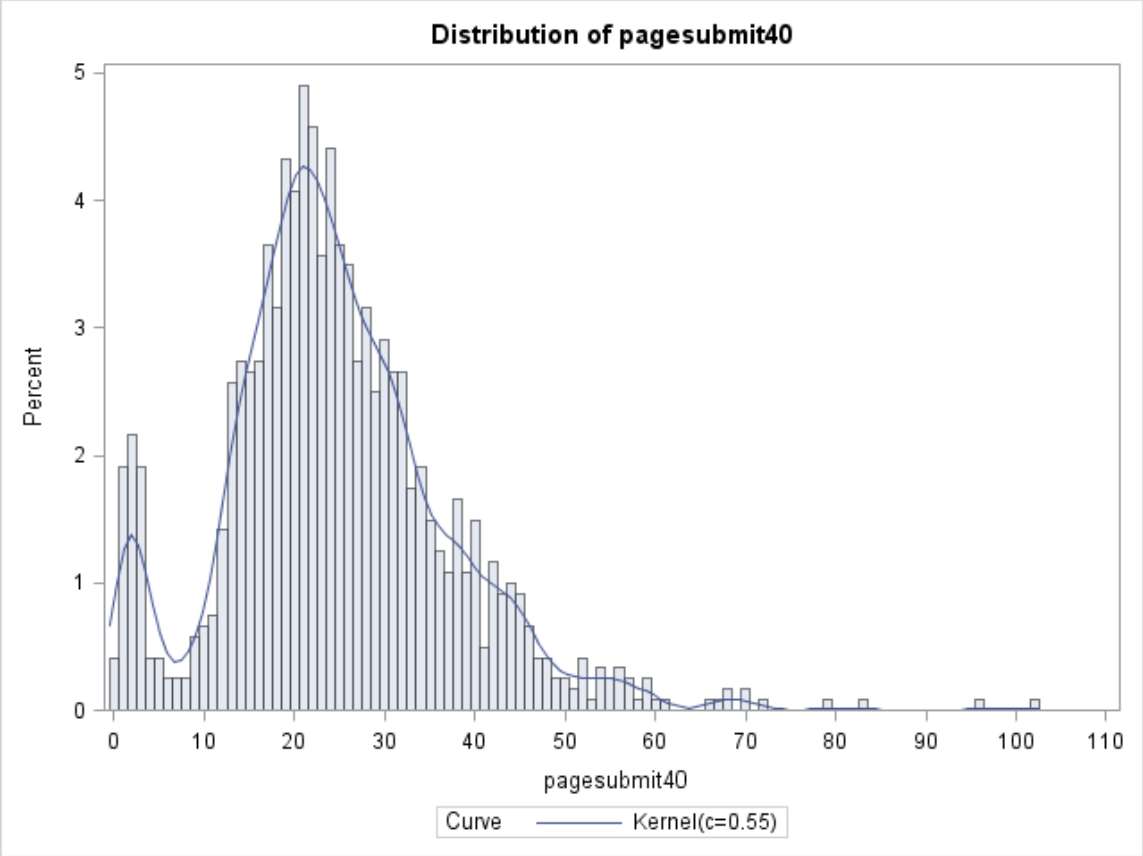


Figure 1. Example of a bimodal response time distribution.

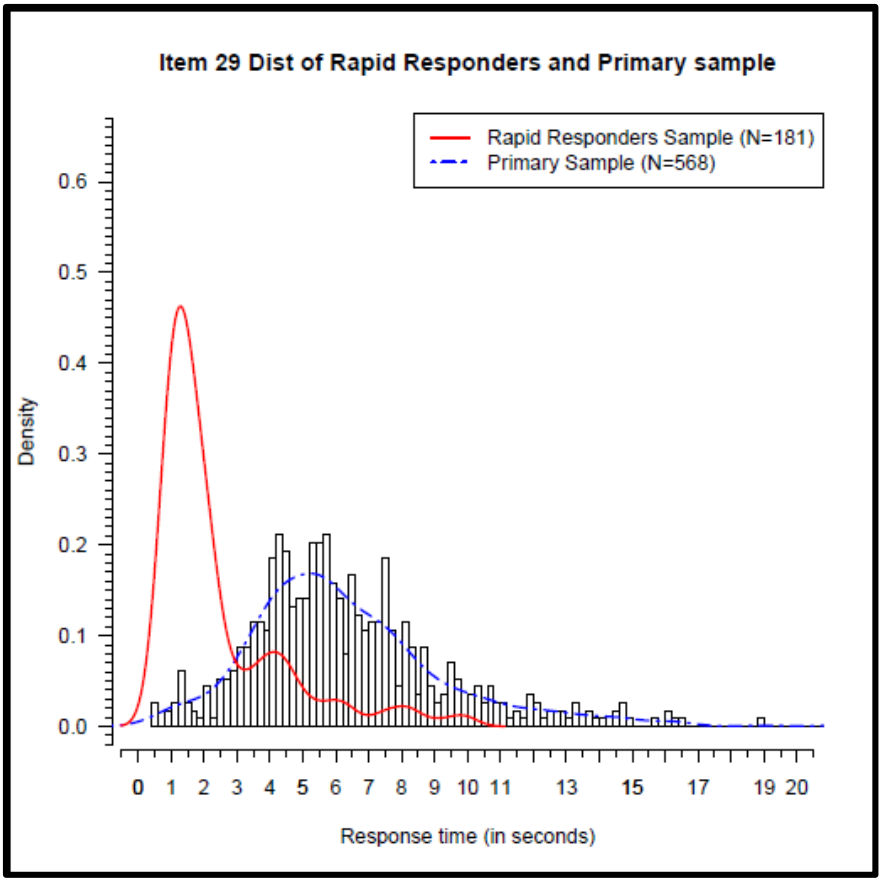


Figure 2. Example of response time distributions examined for the Visual Inspection with Information threshold calculation method.

	ID	Repeated_Measure	Method	RTE	INSPECT	INSPECT2	MIXTURE	MIXTURE2	NT10	NT20	NT30	RSPEED
1	1	1	INSPECT	1	1	0	0	0	0	0	0	0
2	1	2	INSPECT2	1	0	1	0	0	0	0	0	0
3	1	3	MIXTURE	1	0	0	1	0	0	0	0	0
4	1	4	MIXTURE2	0.96	0	0	0	1	0	0	0	0
5	1	5	NT10	1	0	0	0	0	1	0	0	0
6	1	6	NT20	1	0	0	0	0	0	1	0	0
7	1	7	NT30	1	0	0	0	0	0	0	1	0
8	1	8	RSPEED	0.925	0	0	0	0	0	0	0	1
9	2	1	INSPECT	1	1	0	0	0	0	0	0	0
10	2	2	INSPECT2	1	0	1	0	0	0	0	0	0
11	2	3	MIXTURE	1	0	0	1	0	0	0	0	0
12	2	4	MIXTURE2	0.96	0	0	0	1	0	0	0	0
13	2	5	NT10	1	0	0	0	0	1	0	0	0
14	2	6	NT20	1	0	0	0	0	0	1	0	0
15	2	7	NT30	1	0	0	0	0	0	0	1	0
16	2	8	RSPEED	0.943	0	0	0	0	0	0	0	1
17	3	1	INSPECT	0.075	1	0	0	0	0	0	0	0
18	3	2	INSPECT2	0.038	0	1	0	0	0	0	0	0
19	3	3	MIXTURE	0.077	0	0	1	0	0	0	0	0
20	3	4	MIXTURE2	0	0	0	0	1	0	0	0	0
21	3	5	NT10	1	0	0	0	0	1	0	0	0
22	3	6	NT20	0.491	0	0	0	0	0	1	0	0
23	3	7	NT30	0.132	0	0	0	0	0	0	1	0
24	3	8	RSPEED	0.019	0	0	0	0	0	0	0	1
25	4	1	INSPECT	0.962	1	0	0	0	0	0	0	0
26	4	2	INSPECT2	0.962	0	1	0	0	0	0	0	0
27	4	3	MIXTURE	0.962	0	0	1	0	0	0	0	0
28	4	4	MIXTURE2	0.86	0	0	0	1	0	0	0	0
29	4	5	NT10	0.981	0	0	0	0	1	0	0	0
30	4	6	NT20	0.962	0	0	0	0	0	1	0	0
31	4	7	NT30	0.962	0	0	0	0	0	0	1	0
32	4	8	RSPEED	0.868	0	0	0	0	0	0	0	1

Figure 3. Snapshot of the respondent level data with Response Time Effort scores analyzed in Phase Two.

	ItemID	Repeated_Measure	Method	RTF	INSPECT	INSPECT2	MIXTURE	MIXTURE2	NT10	NT20	NT30	RSPEED
1	1	1	INSPECT	0.9929577465	1	0	0	0	0	0	0	0
2	1	2	INSPECT2	0.9894366197	0	1	0	0	0	0	0	0
3	1	3	MIXTURE	0.9894366197	0	0	1	0	0	0	0	0
4	1	4	MIXTURE2	.	0	0	0	1	0	0	0	0
5	1	5	NT10	0.9982394366	0	0	0	0	1	0	0	0
6	1	6	NT20	0.9929577465	0	0	0	0	0	1	0	0
7	1	7	NT30	0.9823943662	0	0	0	0	0	0	1	0
8	1	8	RSPEED	0.9947183099	0	0	0	0	0	0	0	1
9	2	1	INSPECT	0.9771126761	1	0	0	0	0	0	0	0
10	2	2	INSPECT2	0.9718309859	0	1	0	0	0	0	0	0
11	2	3	MIXTURE	0.9753521127	0	0	1	0	0	0	0	0
12	2	4	MIXTURE2	0.8257042254	0	0	0	1	0	0	0	0
13	2	5	NT10	0.9964788732	0	0	0	0	1	0	0	0
14	2	6	NT20	0.9929577465	0	0	0	0	0	1	0	0
15	2	7	NT30	0.9771126761	0	0	0	0	0	0	1	0
16	2	8	RSPEED	0.9049295775	0	0	0	0	0	0	0	1
17	3	1	INSPECT	0.9718309859	1	0	0	0	0	0	0	0
18	3	2	INSPECT2	0.9330985915	0	1	0	0	0	0	0	0
19	3	3	MIXTURE	0.9788732394	0	0	1	0	0	0	0	0
20	3	4	MIXTURE2	0.7588028169	0	0	0	1	0	0	0	0
21	3	5	NT10	1	0	0	0	0	1	0	0	0
22	3	6	NT20	0.9964788732	0	0	0	0	0	1	0	0
23	3	7	NT30	0.9841549296	0	0	0	0	0	0	1	0
24	3	8	RSPEED	0.8732394366	0	0	0	0	0	0	0	1

Figure 4. Snapshot of the item level data with Response Time Fidelity scores analyzed in Phase Three.

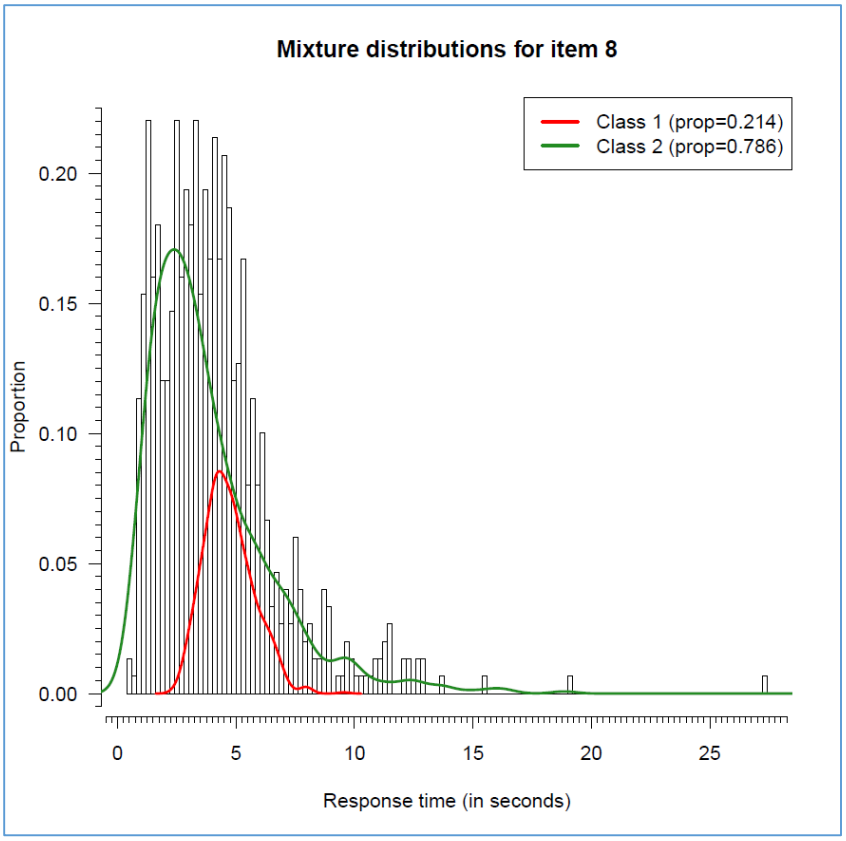


Figure 5. Histogram of the response time distribution for item 8 including the Class One and Class Two mixture densities estimated using the Lognormal Mixture Modeling with Information threshold calculation method.

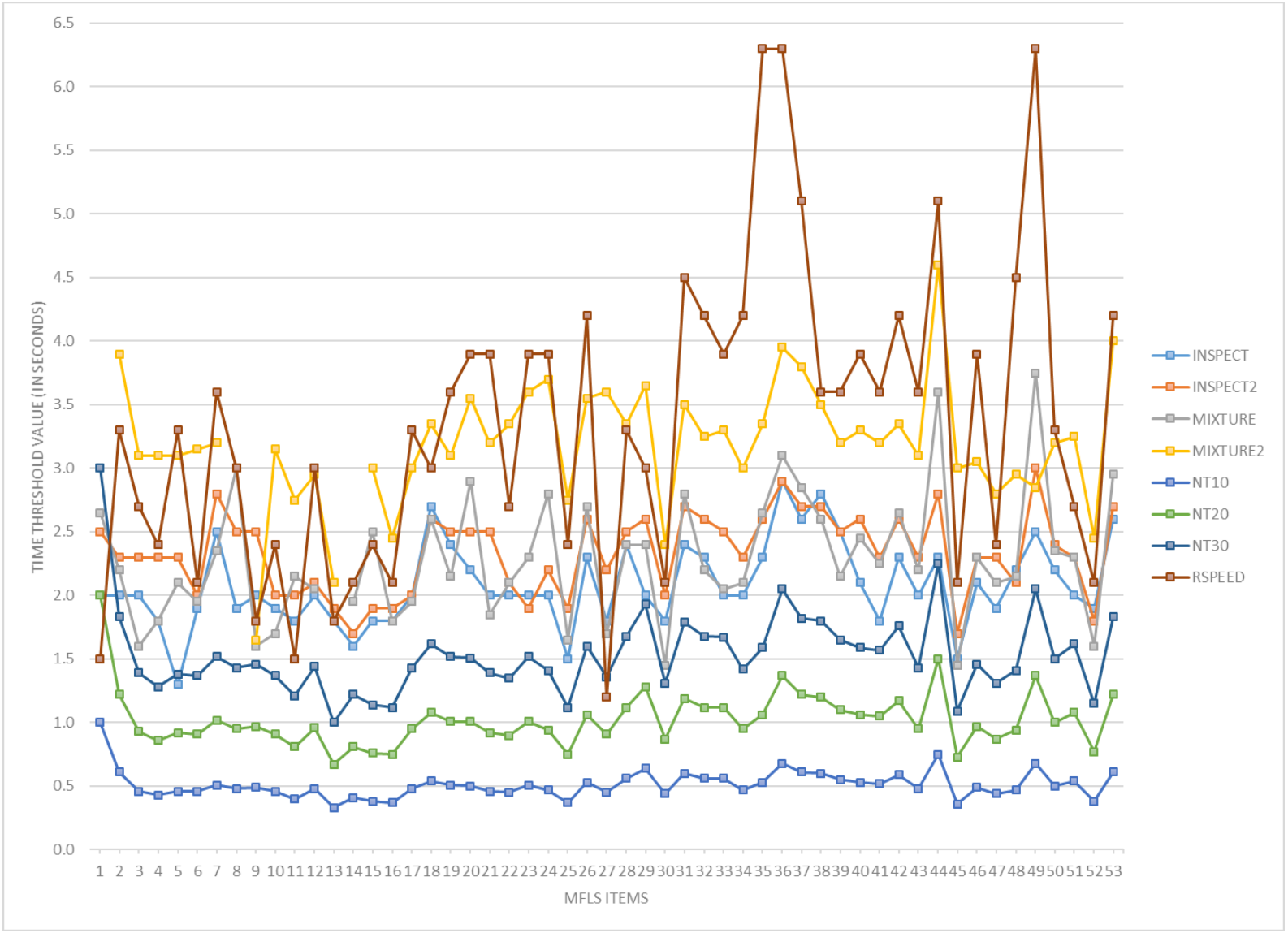


Figure 6. Graph of the defined time thresholds for MFLS items, by threshold calculation method.

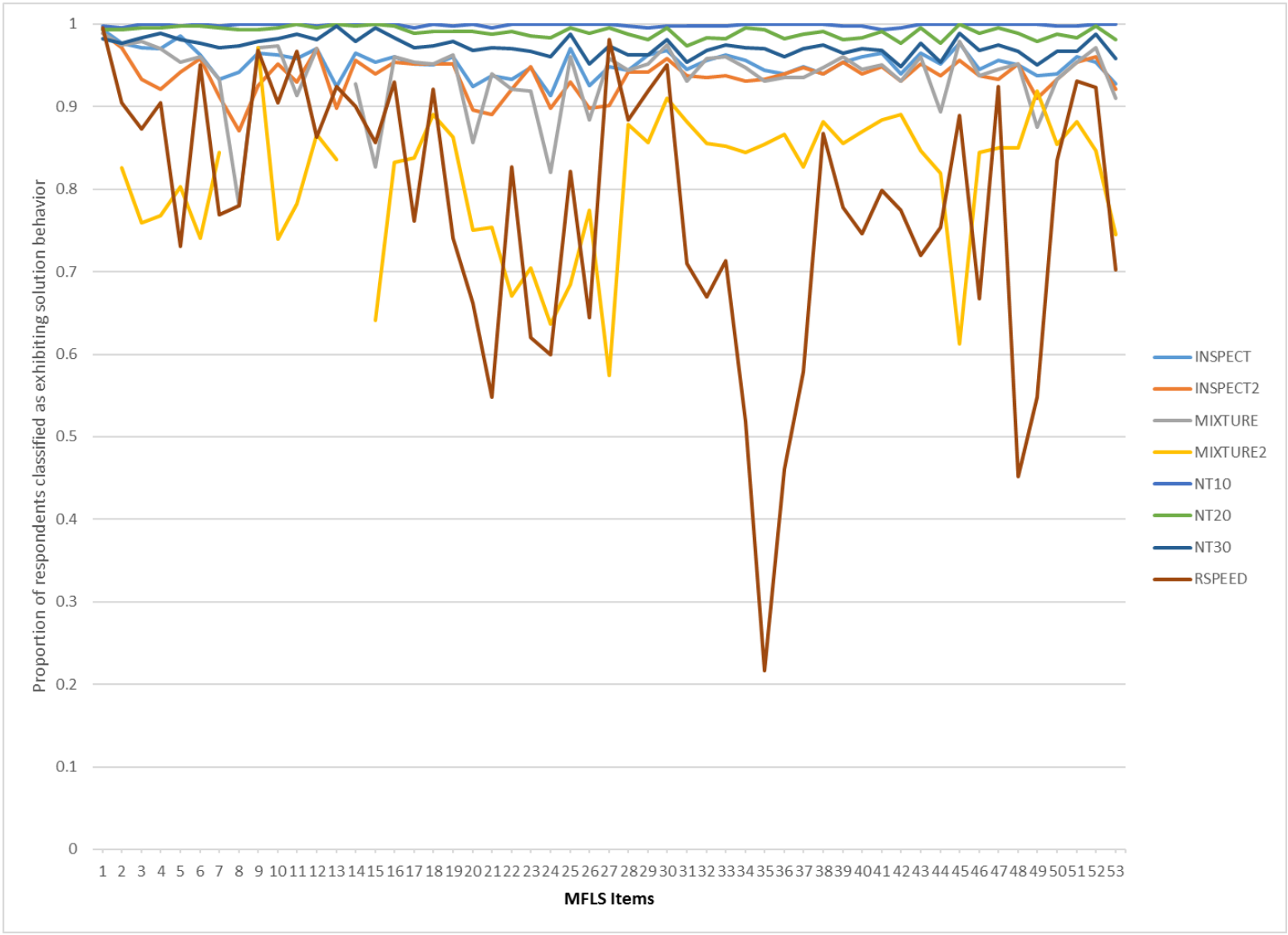
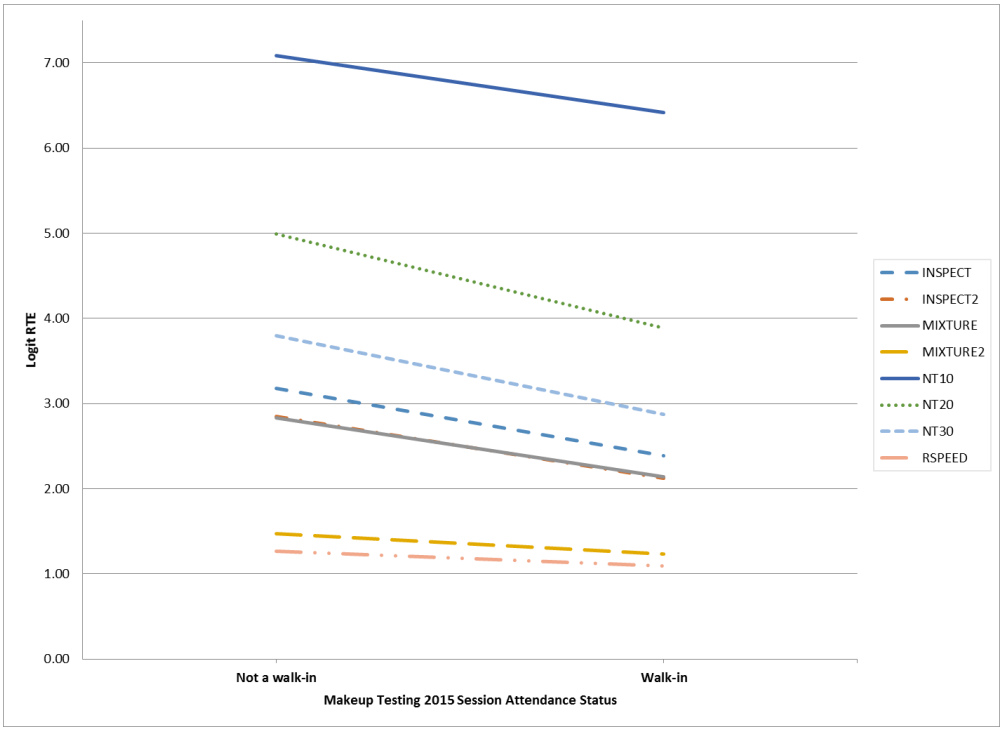


Figure 7. Proportion of respondents classified as exhibiting solution behavior on MFLS items, by threshold calculation method.



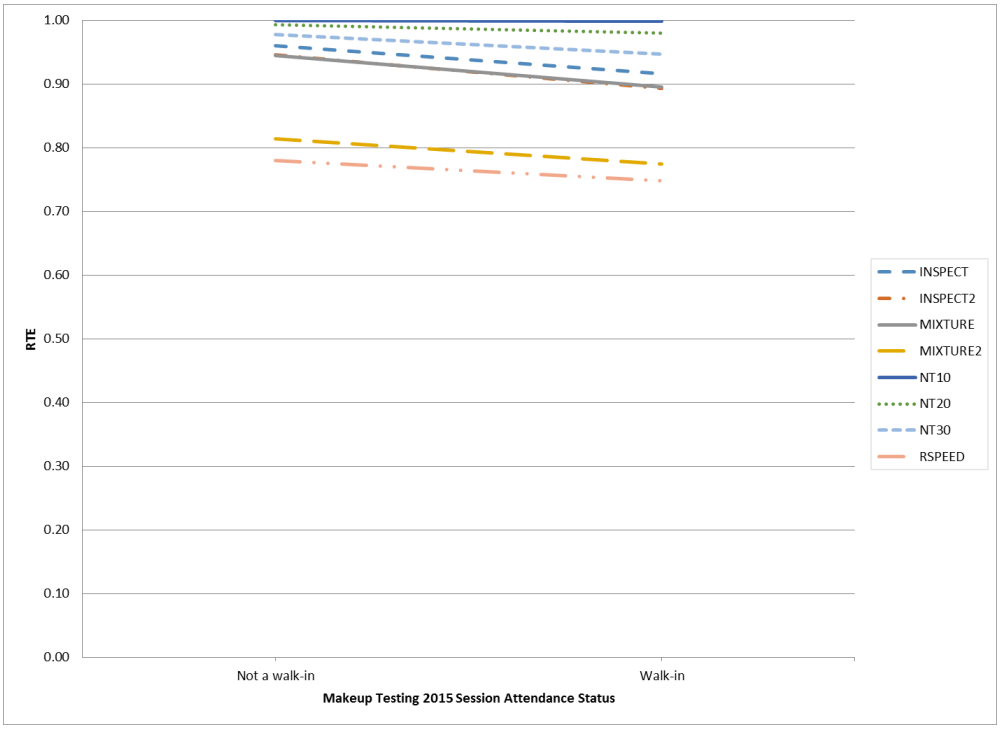


Figure 8. Graphs of the interaction between the logit of RTE and predicted RTE (top and bottom graphs, respectively) and its relationship with Makeup Testing session attendance status (walk-in), by threshold calculation method.

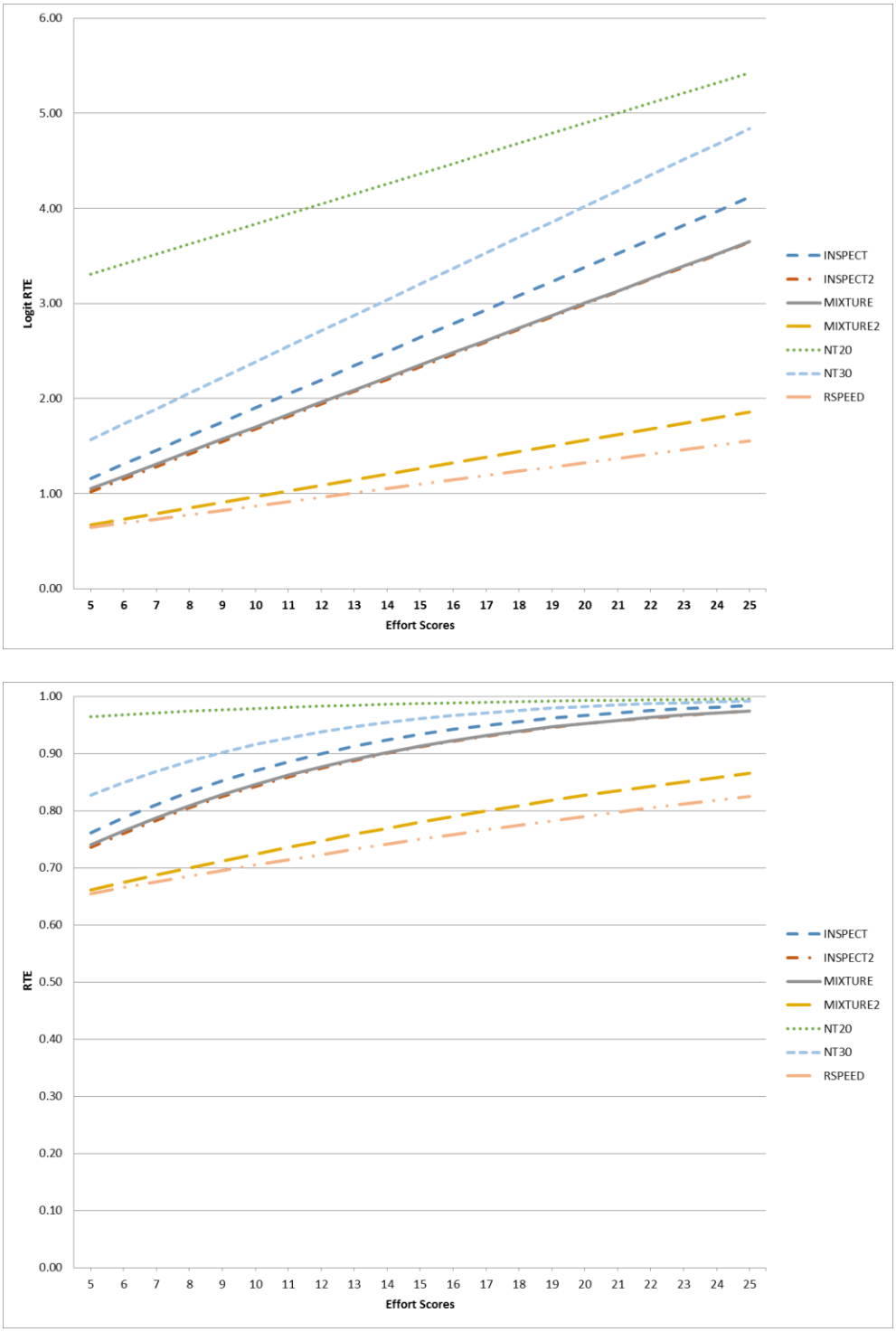


Figure 9. Graphs of the interaction between the logit of RTE and predicted RTE (top and bottom graphs, respectively) and its relationship with effort, by threshold calculation method

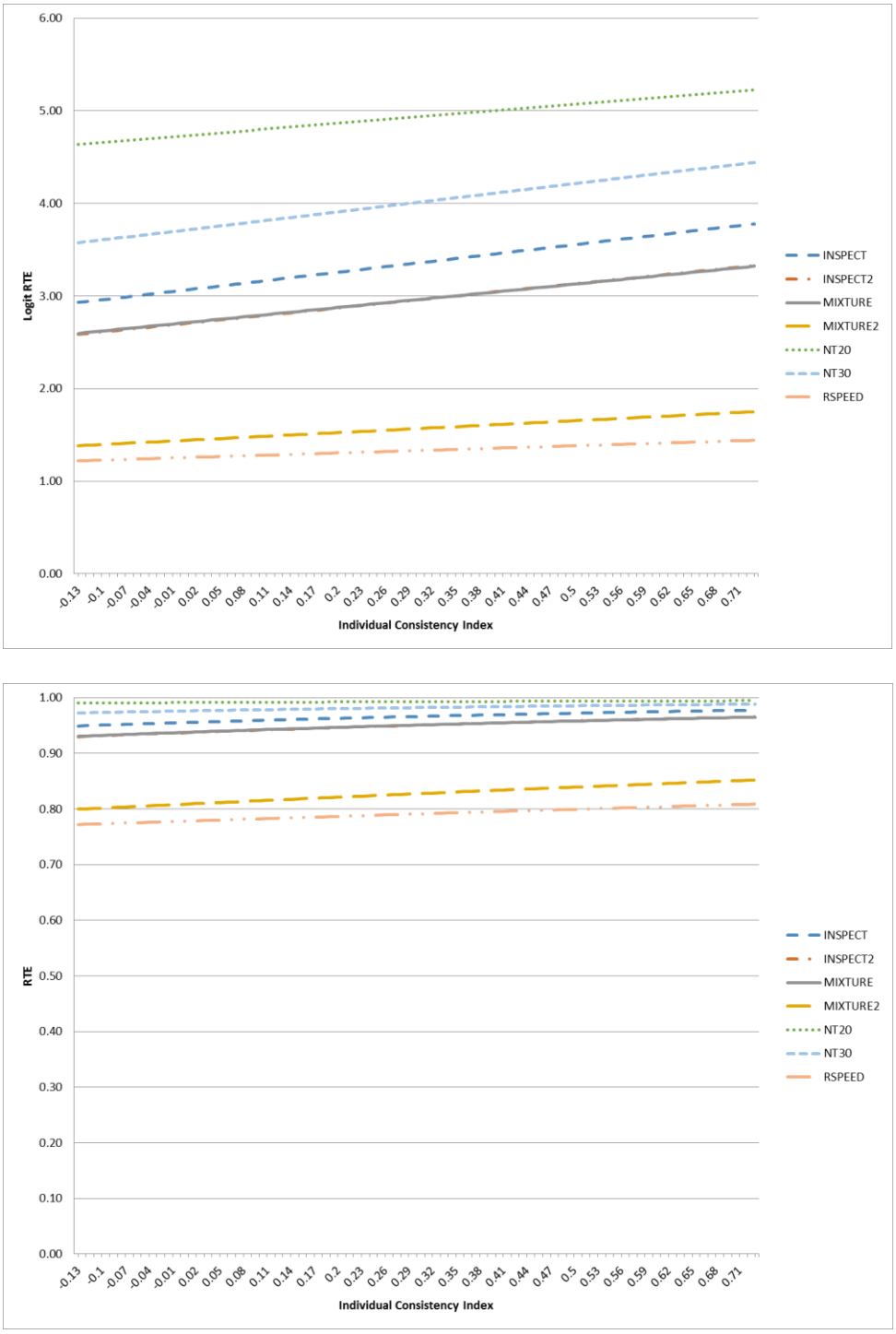


Figure 10. Graphs of the interaction between the logit of RTE and predicted RTE (top and bottom graphs, respectively) and its relationship with the individual consistency index, by threshold calculation method

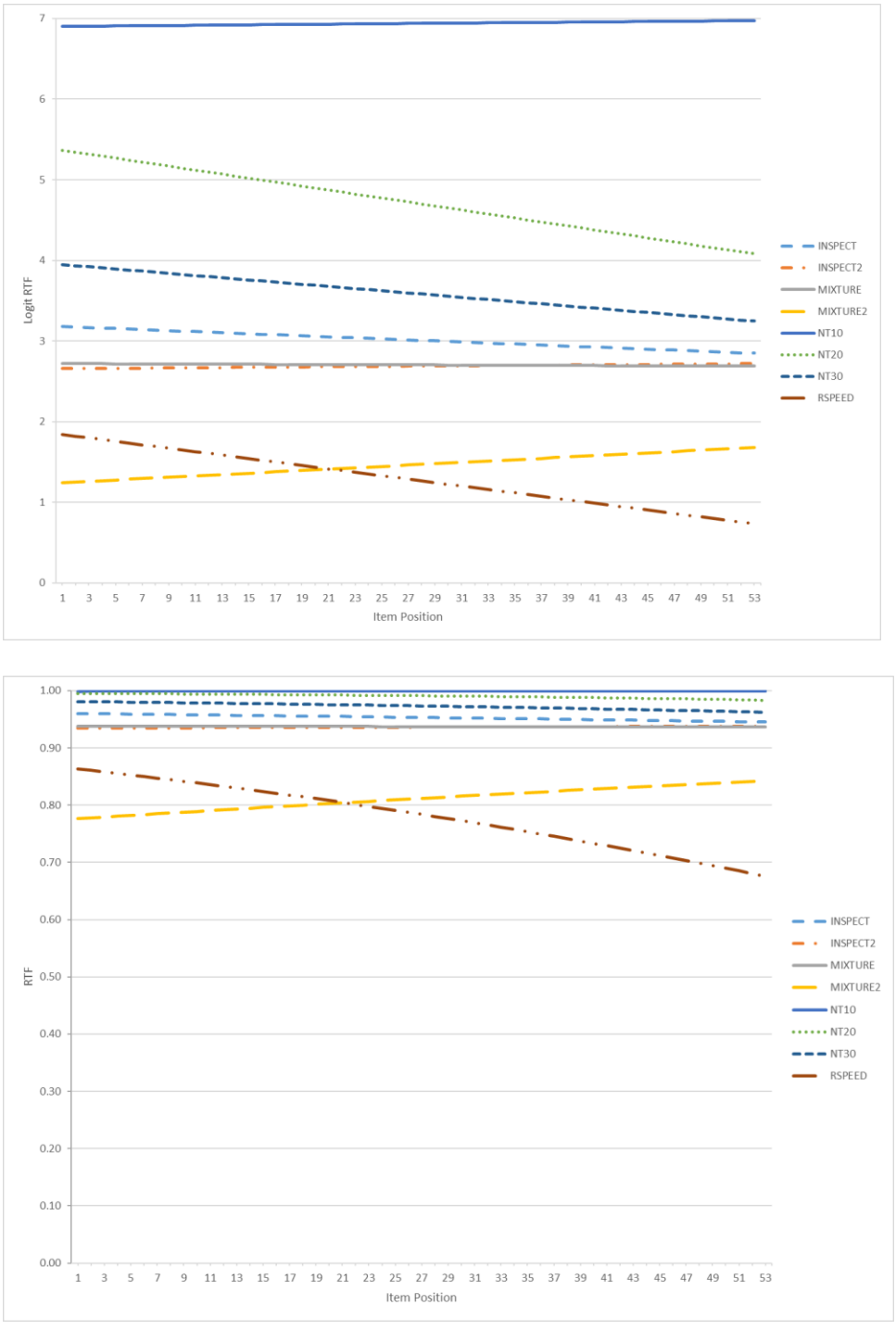


Figure 11. Graphs of the interaction between the logit of RTF and predicted RTF (top and bottom graphs, respectively) and its relationship with item position, by threshold calculation method

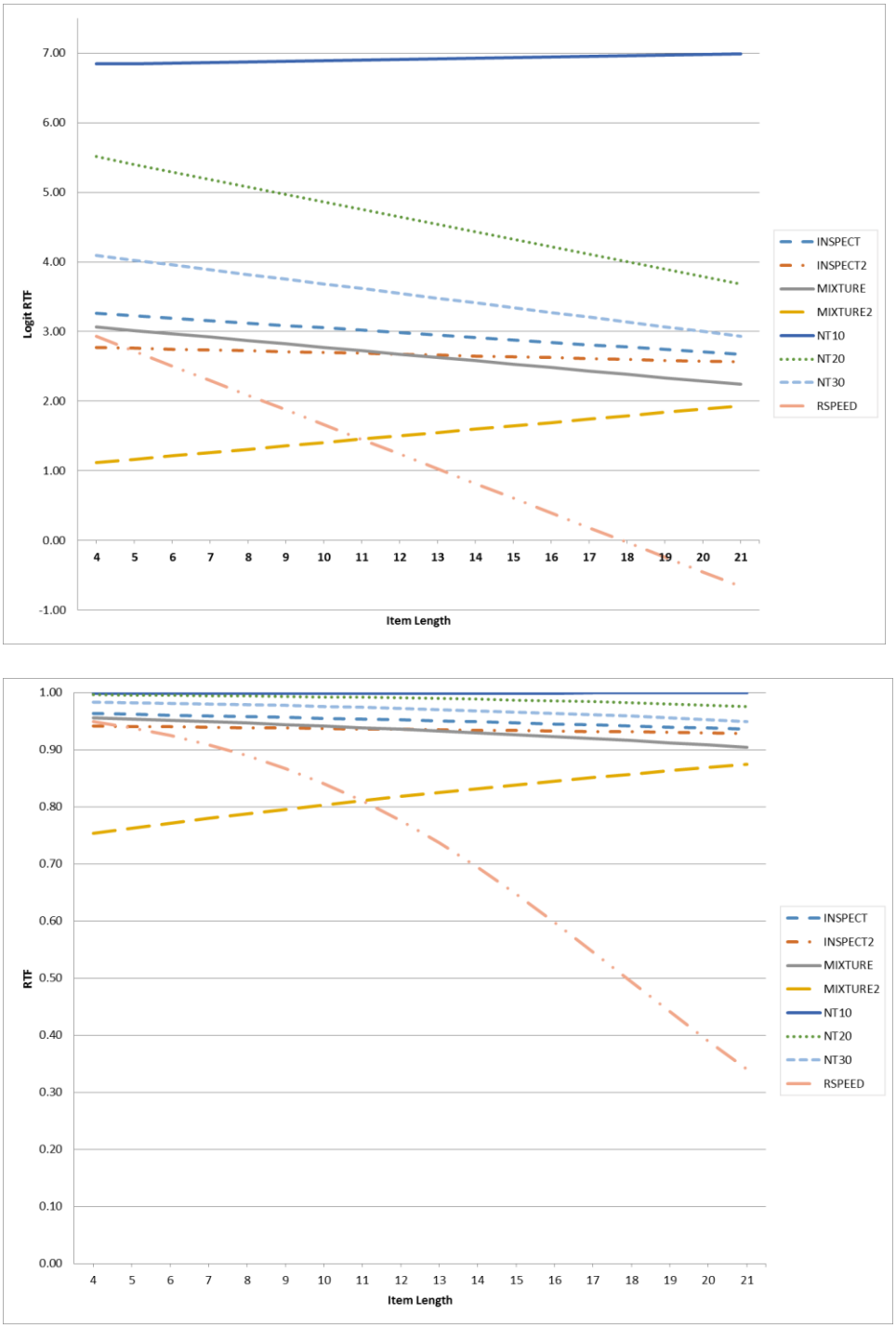


Figure 12. Graphs of the interaction between the logit of RTE and predicted RTF (top and bottom graphs, respectively) and its relationship with item length, by threshold calculation method