


Spring 2015

Extending an IRT mixture model to detect random responders on non-cognitive polytomously scored assessments

Mandalyn R. Swanson
James Madison University

Follow this and additional works at: <https://commons.lib.jmu.edu/diss201019>

 Part of the [Higher Education Commons](#), and the [Quantitative Psychology Commons](#)

Recommended Citation

Swanson, Mandalyn R., "Extending an IRT mixture model to detect random responders on non-cognitive polytomously scored assessments" (2015). *Dissertations*. 24.
<https://commons.lib.jmu.edu/diss201019/24>

This Dissertation is brought to you for free and open access by the The Graduate School at JMU Scholarly Commons. It has been accepted for inclusion in Dissertations by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.

Extending an IRT Mixture Model to Detect Random Responders on Non-Cognitive
Polytomously Scored Assessments

Mandalyn R. Swanson

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

May 2015

Acknowledgements

The completion of this dissertation would not have been possible without the support and encouragement I received from others throughout the process. To these individuals, I would like to express my deepest gratitude.

First, I would like to recognize and thank Dr. Dena Pastor, my advisor and mentor. Thank you, Dena for taking me on as your advisee, for all of your advice and guidance along the way, for your positive and encouraging attitude, and most of all, for believing in me. Not only would this dissertation not have come to fruition without your feedback and direction, but I would not be the person I am today. You are an amazing teacher and mentor and I am forever indebted to you. Thank you.

To my committee members, Dr. Jeanne Horst and Dr. Monica Erbacher, thank you for all of your time, advice, and support. You both have been instrumental in my completion of this dissertation and in my doctoral studies. It has been such a pleasure to work with you both and I am so lucky to have been afforded the opportunity.

To all of the Center for Assessment and Research Studies (CARS) faculty and staff, thank you for your personal contributions to my growth and development over the past three years. I would like to specifically thank Dr. Donna Sundre for her support, encouragement, praise, and sharing of her infinite assessment knowledge with me. Thank you, Donna for always reminding me to have fun.

To my peers and closest friends, Dr. Matthew Swain, Dr. Rory Lazowski, and Cathryn Richmond, thank you for your friendship. I'm so glad to have met you.

To my family and friends outside of CARS, thank you for believing in me and encouraging me from the beginning. Dr. Melanie Marks and Dr. Ling Whitworth, thank you for your mentorship, your encouragement and support, and for believing in me. To my parents and sister, thank you for always being there for me, for loving me, and for understanding when most of my time and effort was dedicated to not only this dissertation, but my doctoral program. And, to Jordan Gilles; I am eternally grateful for your unconditional love, support, and encouragement throughout this process. You never once doubted me, even when I doubted myself. I could not have done this without you. Thank you.

Table of Contents

I. Introduction	1
High-Stakes versus Low-Stakes Settings	3
Levels of Motivation Associated with Low-Stakes	5
Cognitive, Non-Cognitive, and Survey Instruments	7
Response Styles	8
Optimizing versus Satisficing	10
Methods for Detecting Amotivated Examinees	11
The Random Responding Model (RRM)	14
Purpose Statement	15
II. Review of Relevant Literature	17
Purpose	17
Organization of Literature Review	17
Techniques for Addressing Low-Stakes Response Data	18
Motivation Filtering	18
RTE	22
Statistical Models	27
Item Response Theory (IRT) Mixture Modeling	30
RRM with dichotomous items in low stakes testing	35
<i>Examinee classification with the two-class RRM 2PL model</i>	41
<i>Comparison of the one-class 2PL and two-class RRM 2PL models</i>	42
Non-Cognitive Models	48
Need for Study	49
III. Methods	53
Study 1	53
Data Generation	53
Valid Responders	54
Random Responders	57
Simulation Study Design	57
Phase 1	57
Phase 2	59
Software	61
Local maxima	62
Study 2	63

Low-stakes Assessment Dataset	63
Measures	64
Unified Measure of University Mattering (UMUM-15)	64
Student Opinion Scale (SOS)	65
External Validity Analyses	65
IV. Results	68
Study 1	68
RQ1:	68
Factor loadings	68
Thresholds	71
Thetas	74
RQ2	76
RQ3	77
RQ4	78
Factor loadings	78
Thresholds	81
Thetas	83
Study 2	88
RQ1	88
RQ2	90
RQ3	90
RQ5	91
V. Discussion	94
Study 1	94
Magnitude and direction of bias observed	98
GRM	98
RRM-GRM	98
Implications	99
Limitations	102
Study 2	103
Future research	105
Comparing Study 1 and Study 2 Results	106
Conclusions	107
Appendix A.....	109

Appendix B	111
Appendix C	114
Appendix D	126
References	134

List of Tables

Table 1	54
Table 2	55
Table 3	69
Table 4	72
Table 5	75
Table 6	76
Table 7	77
Table 8	78
Table 9	79
Table 10	81
Table 11	84
Table 12	86
Table 13	87
Table 14	90
Table 15	90
Table 16	91
Table 17	92
Table 18	93
Table 19	96

List of Figures

Figure 1. Factor Loading Bias for the GRM.....	70
Figure 2. Factor Loading RMSE for the GRM.	71
Figure 3. Threshold bias for the GRM.....	73
Figure 4. Threshold RMSE for the GRM.	73
Figure 5. Factor loading bias for the RRM-GRM.....	80
Figure 6. Factor loading RMSE for the RRM-GRM.	80
Figure 7. Threshold bias for the RRM-GRM.....	82
Figure 8. Threshold RMSE for the RRM-GRM.	83
Figure 9. Differences between GRM and RRM-GRM factor loading estimates.....	89
Figure 10. Differences between GRM and RRM-GRM threshold estimates.	89
Figure 11. Probability of membership in the random responder class.	93
Figure 12. Test information function (TIF) for all conditions in the GRM.....	101

Abstract

This study represents an attempt to distinguish two classes of examinees – random responders and valid responders – on non-cognitive assessments in low-stakes testing. The majority of existing literature regarding the detection of random responders in low-stakes settings exists in regard to cognitive tests that are dichotomously scored. However, evidence suggests that random responding occurs on non-cognitive assessments, and as with cognitive measures, the data derived from such measures are used to inform practice. Thus, a threat to test score validity exists if examinees' response selections do not accurately reflect their underlying level on the construct being assessed. As with cognitive tests, using data from measures in which students did not give their best effort could have negative implications for future decisions. Thus, there is a need for a method of detecting random responders on non-cognitive assessments that are polytomously scored.

This dissertation provides an overview of existing techniques for identifying low-motivated or amotivated examinees within low-stakes cognitive testing contexts including motivation filtering, response time effort, and item response theory mixture modeling, with particular attention paid to an IRT mixture model referred to in this dissertation as the Random Responders model – Graded Response model (RRM-GRM). Two studies, a simulation and an applied study, were conducted to explore the utility of the RRM-GRM for detecting and accounting for random responders on non-cognitive instruments in low-stakes testing settings. The findings from the simulation study show considerable bias and RMSE in parameter estimates and bias in theta estimates when the proportion of random responders is greater than 5%. Use of the RRM-GRM with the

same data sets provides parameter estimates with minimal to no bias and RMSE and theta estimates that are essentially bias free. The applied study demonstrated that when fitting the RRM-GRM to authentic data, 5.6% of the responders were identified as random responders. Respondents classified as random responders were found to have higher odds of being males and of having lower scores on importance of the test, as well as lower average total scores on the UMUM-15 measure used in the study. Limitations of the RRM-GRM technique are discussed.

I. Introduction

Within the last decade, higher education institutions have experienced increasing pressure from external stakeholders to demonstrate compelling empirical evidence of institutional quality. Spellings (2006) made the call for greater accountability and transparency in higher education apparent, stating that there “...is a lack of clear, reliable information about the...quality of postsecondary institutions, along with a remarkable absence of accountability mechanisms to ensure that colleges succeed in educating students” (p.vii).

Consequently, implementation of assessments to evaluate student learning and provide evidence of institutional quality have increased. The data collected from these assessments are not only reported to external stakeholders for accountability purposes, but they are also frequently used at the program or institution level to aid in augmentation of curriculum and to facilitate decision making for programmatic issues.

Tests administered for accountability and assessment purposes by postsecondary institutions are generally focused on measuring academic student learning outcomes, such as those associated with individual majors or the general education curriculum (Suskie, 2009). For example, student ability in the domains of critical thinking, quantitative reasoning, written and oral communication skills, and major-specific content knowledge is commonly assessed. Thus, many academic student learning outcomes are cognitive, or knowledge-based (Abedi & O’Neil, 2005; Heckman & Rubinstein, 2001). When administering achievement tests, we are attempting to measure examinees’ proficiency, or what they know and can do. In evaluating the scores produced from such tests, we make an implicit assumption that examinees put forth their best effort in

demonstrating their proficiency (Wise & Kong, 2005; Zerpa, Hachey, van Barneveld, & Simon, 2011). However, researchers have questioned this assumption, as inferences made on the basis of test scores are dependent upon construct-irrelevant factors, such as the amount of effort examinees exerted while completing the test (Wise & Kong, 2005; Zerpa et al., 2011). Essentially, if test scores are not consequential or important to examinees, it is reasonable to assume examinees may not put forth their best effort (Liu, Bridgeman & Adler, 2012; Wise & DeMars, 2005; Wise & Kong, 2005; Wise, Wise & Bhola, 2006; Zerpa et al., 2011). Without sufficient effort, examinee performance suffers and scores on the tests do not reflect examinees' actual proficiency. In fact, test scores would actually under-represent examinees' true ability on the construct, thereby negatively biasing proficiency estimates (Wise & DeMars, 2010). Thus, when examinees exert low effort on tests, a potential threat to test score validity exists (Liu et al., 2010; Wise & DeMars, 2005; Wise & DeMars, 2010; Wise & Kong, 2005).

The same threat to validity occurs with the administration of non-cognitive instruments, such as self-report measures that use Likert-type scales. In addition to the cultivation of knowledge, or cognitive skills, the mission of institutions of higher education includes development of non-cognitive skills such as leadership and character (Schmitt et al., 2011). For example, motivation, trustworthiness, beliefs, personality, and perseverance are only a few of the non-cognitive domains assessed by higher education institutions (Abedi & O'Neil, 2005; Heckman & Rubinstein, 2001). As with cognitive measures, the data derived from non-cognitive measures are used to inform practice. Thus, a threat to test score validity exists if examinees' response selections do not

accurately reflect their underlying level on the construct being assessed. As with cognitive tests, using data from measures in which students did not give their best effort could have negative implications for future decisions.

This chapter begins by describing the differences between high and low stakes testing settings. The varying levels of motivation associated specifically with measures administered in a low stakes context are then described, with particular attention given to amotivated examinees. The distinction between three types of instruments: cognitive, non-cognitive, and survey, is then made by comparing and contrasting the measures' overall purpose and actual instrument design. Because response styles and behaviors of optimizing and satisficing are associated with examinee motivation and validity of respondent scores, this chapter briefly discusses both concepts in relation to non-cognitive assessments. Because amotivated examinees are an issue of focus, methods for detecting this type of examinee are also examined. This chapter concludes by explaining the purpose of the current study, which seeks to build upon existing literature by examining the implications of detecting and modeling amotivated examinees on non-cognitive tests.

High-Stakes versus Low-Stakes Settings

When an examinee needs a high test score in order to gain a desired benefit (e.g. to obtain medical licensure or gain admission to a program) the test is considered a "high-stakes" test (Wise & Kong, 2005). Because high-stakes tests involve personal consequences associated with examinee performance, it can be assumed that examinees will exert good effort when completing the test. Examples of high-stakes tests include graduation, admissions, and licensure exams.

Although it is possible that not all examinees will put forth their best effort on high-stakes tests, test administrators and researchers place little focus on examinee effort on high-stakes exams, as it is assumed that examinees will try their best due to the associated personal consequences (Wise & Kong, 2005). Essentially, it is the responsibility of the examinees to put forth their best effort on a high-stakes test. Exerting low effort is considered a personal choice of the examinee to forego the benefits associated with high-stakes test scores, and therefore, is not considered a concern of the test administrator (Wise & Kong, 2005).

In contrast, “low-stakes” tests are characterized by their lack of personal consequences to the examinee for test performance (Liu et al., 2010; Wise & Kong, 2005; Wise et al., 2006). In some instances testing examinees is necessary, but attaching personal consequences to results is not possible. Essentially, there are three common low-stakes testing situations: 1) program evaluation; 2) test development; and 3) basic research (Wise & Kong, 2005). Regarding program evaluation, assessment programs that have potential consequences for institutions, but not individual examinees, exist. For instance, sometimes assessments are conducted for determining quality of instruction, for funding purposes, or for general accountability reasons. In the case of test development, administration of a high-stakes test in low-stakes settings is common practice. For example, this practice may occur when piloting test items for standardized tests, such as the SAT or GRE, and to collect validity evidence prior to widespread use of a test (Wise & Kong, 2005). Another low-stakes testing situation occurs for research purposes. For instance, students are sometimes required by a professor to participate in a university study as a requirement of a course (Lau, 2009; Wise & DeMars, 2005; Wise & Kong,

2005). There are additional instances in which low-stakes testing can occur, but those discussed previously are the most prevalent (Wise & Kong, 2005).

When examinees are given a non-consequential assessment test, some individuals may not be as concerned about achieving the highest score possible and subsequently, their scores may not represent their true level of proficiency on a construct. This score attenuation can be attributed to the fact that examinees will not be penalized for their performance nor will they receive any individualized benefit (Lau, 2009; Wise & DeMars, 2005; Wise & Kong, 2005). In low-stakes testing situations, examinees vary in the amount of effort they expend on completing such assessments. Despite the lack of personal consequences, many examinees still give good effort in completing low-stakes tests (Wise & Kong, 2005). However, researchers and test administrators are aware that some examinees give low, or even no effort at all. In low-stakes testing situations, the effort exerted by examinees is a serious issue and the responsibility for obtaining valid test scores is not considered to be that of the examinee, but of the test administrator (Wise & Kong, 2005).

Levels of Motivation Associated with Low-Stakes

When administered a low-stakes test, some examinees will still put forth their best effort on the test and fully engage in responding to all of the items. This may be because they are interested in the test, value the test's purpose, or because they have been trained to give their best effort when completing a test, among other things (Lau, 2009; Wise & DeMars, 2005). Essentially, some aspect of the testing scenario must support examinees' reasons for trying on the test. These examinees are considered to be *motivated*.

Other examinees may exert some effort, but not as much as if the test were for a grade or associated with other types of personal consequences (Lau, 2009; Wise & DeMars, 2005; Wise & Kong, 2005). In this scenario, some examinees may start out by expending high levels of effort, but their effort will wane during the test. Others may choose to answer some items, but not others (e.g., easy items that require little effort to answer) (Cao & Stokes, 2008; Wise & DeMars, 2005). Such examinees are considered *moderately-motivated* or *low-motivated* (Lau & Pastor, 2010).

Even more extreme, some examinees may exhibit such severely low levels of effort that they fail to engage in responding to any of the items on the test. This may include omitting their responses or answering items randomly without even opening the test booklet (Wise & DeMars, 2005). These examinees are referred to as *amotivated* (Lau, 2009; Lau & Pastor, 2010).

In sum, the levels of motivation associated with low-stakes testing situations will be relatively lower on average than if the same test were administered in a high-stakes testing situation. In fact, Wise and DeMars (2005) found an average of a .59 standard deviation difference between motivated and unmotivated groups on test performance in a review of 12 empirical studies. Because some examinees will not exert full effort on low-stakes tests, they will not perform to their potential and therefore the scores will not accurately represent examinees' true proficiency (Wise & DeMars, 2005). The average effect size found by Wise and DeMars (2005) demonstrates that motivation differences can be translated into real differences in performance. This issue is incredibly important in relation to validity of test scores, as test results will underestimate examinees' proficiency to the degree to which they fail to give their best effort on the test (Wise &

DeMars, 2005; Wise et al., 2006). That is, when low-motivated or amotivated examinees are present, scores will not accurately reflect examinees' true proficiency and may not be valid indicators of what they know and can do (Wise & DeMars, 2005; Wise et al., 2006). For example, when performance of students is underestimated, assessment results could lead institutions to erroneously conclude that their programing is ineffective or that major changes in curriculum are necessary for pupils to achieve the desired student learning outcomes set forth by the university. Furthermore, underestimations of test scores could potentially affect funding and external stakeholders' perceptions of what students are actually learning.

Cognitive, Non-Cognitive, and Survey Instruments

Whereas much of the focus surrounding student motivation has been centered on cognitive achievement tests, motivation is also a concern with non-cognitive tests. Cognitive skills are often integral to academic and professional success and are associated with thinking, reasoning, and communication. Essentially, cognitive skills require an individual to exhibit purposeful effort intellectually (ACT, 2013). In contrast, non-cognitive skills include motivation, interpersonal interaction and values, among others. Essentially, non-cognitive skills are related to an individual's personality, behaviors, and feelings (ACT, 2013).

For cognitive measures, a correct response exists. In order for an examinee to select the correct response, they must execute a specified skill. For example, on a quantitative reasoning test, examinees are administered items containing math problems. In order to solve each item, examinees must utilize their knowledge of mathematics. Accordingly, there is a specified right and wrong answer. Non-cognitive instruments

attempt to measure the underlying level of an examinee on a particular construct, such as a trait or characteristic (Marsh, 2013). For example, an examinee could be asked to select a value from a rating scale that indicates how important a particular value is to them.

It should be noted that non-cognitive and survey instruments are not the same. Non-cognitive and survey instruments are similar in that they both typically use Likert-type scales and self-report instruments to evaluate outcomes but, they differ in two key ways: what they aim to measure, or in other words, their purpose, and in the inferences made from the resulting data (Marsh, 2013). In contrast to non-cognitive measures, the purpose of survey instruments is to provide specific information about attitudes, beliefs or actions (Marsh, 2013).

In terms of the actual instrument, non-cognitive measures contain multiple items that attempt to measure the same construct; thus a response to a single item is not considered to be meaningful in isolation. To establish a respondent's level of a construct, the responses to several items measuring the same construct are taken into consideration. With survey measures, responses to single items are considered to be of interest to the researcher. Such items are generally concerned with frequency of behaviors or their beliefs and attitudes (Marsh, 2013).

Response Styles

Despite the differences previously noted between surveys and non-cognitive assessments, research regarding the ways in which respondents complete a survey is relevant to non-cognitive assessments. In addition to respondent motivation, the rating scale and wording associated with an item could also be an additional source of construct-irrelevant variance for non-cognitive measures (Baumgartner & Steenkamp,

2001). In this situation, the trait or characteristic being measured could be confounded with response style. Response styles are defined as systematic responses that are not based on content. In other words, response styles are essentially a set of responses made on some basis independent of what the items were designed specifically to measure (Baumgartner & Steenkamp, 2001; Cronbach, 1946). For example, some examinees may have a tendency to agree or disagree with items irrespective of their content, endorse the most extreme options, or respond to items randomly, among a variety of other response styles (Baumgartner & Steenkamp, 2001; Cronbach, 1946; Moustaki & Knott, 2014).

One of the response styles most commonly studied by researchers is that of acquiescence, or the tendency of examinees to agree with items regardless of content (Cloud & Vaughan, 1970; Coleman, 2013; McPherson & Mohr, 2005). In order to detect examinees exhibiting this response style, the practice of “balancing the scale” by including negatively worded or keyed items on an instrument along with positively worded and keyed items, frequently occurs (Cloud & Vaughan, 1970). The use of balanced scales has been thought to improve the psychometric properties of an instrument by averaging out bias so scores are not confounded with response style, specifically that of acquiescence or disacquiescence (Cloud & Vaughan, 1970; McPherson & Mohr, 2005).

A less studied, but frequently recognized response style is that of random responders. Random responders, also referred to as amotivated, are characterized by their tendency to respond to items carelessly or arbitrarily (Baumgartner & Steenkamp, 2001; Lau, 2009; Wise & DeMars, 2005). For example, random responders may not

even open the testing booklet, read the instructions, or interpret items as intended (Baumgartner & Steenkamp, 2001; Lau, 2009; Wise & DeMars, 2005).

Response styles, like acquiescence and random responding, can contaminate respondents' scores and create construct-irrelevant variance in several ways. For example, observed responses can be inflated or deflated. Moreover, the correlation between examinees' scores on instruments purporting to measure the same construct can also be inflated or deflated (Baumgartner & Steenkamp, 2001). Like with motivation, such contaminations of responses can lead to biased conclusions, thus influencing inferences made from scores (Baumgartner & Steenkamp, 2001; Coleman, 2013, Cronbach, 1946).

Optimizing versus Satisficing

When administering an instrument to examinees, researchers aim to acquire high quality data. Tourangeau (1984) proposes a model that contains four stages of cognitive processing that examinees ideally utilize when completing an instrument. Although the context of the model applies to administration of survey instruments, it could also be applied to administration of non-cognitive measures. In the initial stage, stage one, examinees carefully comprehend the meaning of each item. Once they understand the item, they proceed to the second cognitive processing stage, stage two, which involves retrieving all applicable information from memory. Stage three involves integrating the knowledge retrieved from memory with the item to make summary judgments. The summary judgments are then used in stage four to select and report an answer. If examinees execute the four steps of cognitive processing precisely and comprehensively, they are said to be *optimizing*. Optimizing occurs when examinees provide the optimal

(i.e. as accurate as possible, not most socially desirable) answer to each item on a measure (Krosnick, 1999). For just one single item, the task of optimizing requires a great amount of cognitive effort particularly compare to other tasks, like satisficing; thus, there is a substantial amount of mental work required to complete a sequence of questions, much less a series of instruments. Consequently, optimizing behavior requires significant motivation from the examinee.

While some examinees are motivated to expend the cognitive effort required to optimize throughout the entirety of an instrument or series of measures, others may drop-off at some point due to fatigue, loss of interest, or distractions, or never even engage in optimizing from the start. This behavior is termed *satisficing*. With satisficing, examinees could execute all four steps of Tourangeau's (1984) cognitive processing model, but less diligently than those exhibiting optimizing behavior. Instead of exerting maximum effort in providing the optimal answer, such examinees settle for answers that are simply satisfactory. This behavior has been termed "weak satisficing." More drastically, examinees could skip steps in the cognitive processing model or just arbitrarily answer items without completing any of the steps at all. This behavior is categorized as "strong satisficing." Krosnick (1991, 2011) identified three factors that increase the likelihood that an examinee will exhibit satisficing behavior. These factors include: tasks or items with increased levels of difficulty, low ability on the construct being measured, and low motivation. Random responders, or amotivated examinees, are considered to be strong satisficers because no retrieval or judgment is used to select their answers.

Methods for Detecting Amotivated Examinees

The *Standards of Educational and Psychological Testing* (AERA, APA, NCME, 2014), state that "...a test taker's score should not be interpreted in isolation; other relevant information that may lead to alternative explanations for the examinee's test performance should be considered" (Standard 9.13). Therefore, test administrators have a responsibility to document examinee motivation levels *and* consider them when interpreting examinee scores. A variety of methods for identifying amotivated examinees in low-stakes testing currently exist. Some of these methods include the reporting of test taking motivation. Two approaches to measuring and reporting test taking motivation include self-report motivation and response-time effort measures (Wise & DeMars, 2005; Wise & Kong, 2005). Self-report motivation measures attempt to discern examinees' opinions about how important the test was to them and the amount of effort they exerted when completing it (Wise & DeMars, 2005). Response-time effort (RTE) measures identify the amount of time examinees spend completing each item on computer-based tests in an attempt to differentiate examinees with different levels of motivation (Wise & DeMars, 2005). Both self-report and RTE measures can be used along with a "cutoff score" to classify an examinee as low-motivated. Sometimes this method is used to study characteristics of low-motivated examinees, whereas other times, it is used simply to identify low motivated examinees so they can be dropped from the data set, a technique known as *motivation filtering* (Steedle, 2014; Sundre & Wise, 2003; Wise & DeMars, 2005; Wise et al., 2006).

There are several reasons why low motivated examinees should be identified, even if they are not subsequently removed from the data set. One reason for detecting amotivated examinees is to estimate the proportion of random responders present in the

data set, which is helpful in better understanding and making inferences from the data. For example, if the proportion of amotivated examinees is extremely small, a researcher or test administrator could use such information as evidence for keeping and analyzing all examinee data because their effect on parameter estimates would be minute. In other words, low motivated examinees may not necessarily need to be removed from the dataset to make valid inferences. In contrast, if the proportion is relatively large, such information could be used to justify the decision to remove examinees from the data set. Some researchers may also be interested in studying the characteristics of amotivated examinees. For example, if random responders are able to be detected and identified, demographic, academic or other types of information could be useful in detecting differences in motivated and amotivated examinees. If differences are detected, such information could be used to provide early interventions to examinees with characteristics similar to that of random responders. Moreover, qualitative studies, such as focus groups or individual interviews, can be conducted in an attempt to determine why examinees were amotivated, and what might make them put forth more effort on similar instruments in the future. Detection of low motivated examinees is also important for exploring the relationship between testing conditions or test characteristics and proportion of random responders. Interactions between the measures and number of amotivated examinees could exist, and such information would be helpful for making changes to future testing conditions.

Statistical models also exist that either explicitly model the item response behavior of low- motivated and/or amotivated examinees or take into account information related to respondent effort, such as response time. Model based methods,

which are described more fully in chapter 2, include the threshold-guessing IRT model, difficulty-guessing IRT model (Cao & Stokes, 2008), and effort moderated item response models (Wise & DeMars, 2005). Many of these models can be used to identify examinees with low motivation. More often the reason for their use is to “purify” the item parameter estimates for valid responders. Having accurate item parameters is always important, but may be of utmost importance if the item parameter estimates are of primary interest because they will be used in deciding on item deletion or alteration. This typically happens during the initial phases of test development (e.g., item analysis, test construction from item information functions).

Another reason to account for random responders in a data set using statistical models is to “purify” the theta estimates for valid responders. In other words, by statistically accounting for the presence of low-motivated examinees, the thetas of valid responders provide more accurate estimates of their true ability.

The Random Responding Model (RRM)

One particular IRT model that can be used to identify amotivated examinees and obtain purified estimates of item parameters and theta estimates for valid responders is the Random Responding Model (RRM), which was first proposed by Mislevy and Verhelst (1990). The RRM is an IRT mixture model that specifies two unknown classes of examinees: one that responds in accordance with a traditional IRT model and another that responds with random guessing or responding. The RRM has since been applied to cognitive, low-stakes assessment data to detect the presence of amotivated examinees (e.g., Lau, 2009; Mislevy & Verheslt, 1990). The assessment data obtained from the cognitive instruments in the studies in which the RRM has been applied were scored

correct or incorrect; thus the RRM is appropriate for dichotomously scored items. Since the RRM is a relatively new technique that requires further study and has not yet been used with non-cognitive measures or polytomously scored items, it warrants further study in this context.

Purpose Statement

The intent of this study was to extend the RRM for use with Likert-type or polytomously scored items in a low-stakes testing context. The purpose for extending the RRM was to determine how item parameters and theta distributions are impacted when random responders are present. In Study 1, a simulation was conducted for the purpose of exploring the effect of random responders in the data set on item parameters and theta distributions. The simulation generated examinee response data that included various percentages of random respondents (1%, 5%, 10%, 20%). Initially, the Graded Response Model (GRM; Samejima, 1969) was fit to the data, ignoring the presence of random responders. In this phase of the study, the impact of random responders on the item parameter and theta estimates when the presence of random responders is ignored was investigated. A modified version of the RRM appropriate for polytomous responses was then fit to the data to determine how well the model identified the proportion of random responders. The extent to which item parameters and theta values were closer to their true values in the valid responder class was investigated when this model was used. Essentially, the first part of the simulation was used to illustrate the impact of the presence of unaccounted for random responders in polytomous data. The second part was to showcase the utility of the RRM to identify random responders and purify estimates for valid responders.

In Study 2, the RRM was applied to non-cognitive data gathered in a low-stakes testing setting from undergraduate students earning anywhere between 45 to 70 credit hours at a mid-sized, southern state university. The purpose of this study was to corroborate the results of using the RRM on real test data with those of the simulated data to provide evidence of the utility and appropriateness of the RRM for use with non-cognitive data collected in a low-stakes setting. In addition to the RRM, the GRM was also fit to the same authentic data set, enabling results from the one-class and two-class models to be compared. In addition, Study 2 also focused on identifying external validity evidence in an attempt to demonstrate that classes differ primarily as a result of test-taking motivation by evaluating differences between classes detected by the RRM on test-taking effort and importance as measured by the Student Opinion Scale (SOS; Sundre & Moore, 2002), total score on the scale, and total completion time of the measure.

II. Review of Relevant Literature

Purpose

The literature review will synthesize research related to existing methods used to identify and/or account for low-motivated or amotivated examinees (i.e. random responders) in low-stakes testing. The purpose of this literature review is to demonstrate that a considerable gap exists in the literature in regard to the detection of random responders on non-cognitive instruments administered in a low-stakes testing context, as the majority of existing studies are cognitive in nature. Within this review, techniques for identifying low-motivated or amotivated examinees within low-stakes cognitive testing contexts are explored. These methods include motivation filtering, response time effort (RTE) and item response theory (IRT) mixture modeling, with particular attention paid to an IRT mixture model known as the Random Responders Model (RRM). The strengths and weaknesses of each method in relation to detecting low-motivated or amotivated examinees, along with current related literature, are presented.

Organization of Literature Review

Methods for identifying amotivated examinees (i.e. random responders) are organized by technique. Each technique's section contains an analysis and synthesis of the related literature and includes associated advantages and disadvantages. Particular attention is paid to the Random Responding Model (RRM) used to identify amotivated examinees, which are those examinees who fail to give effort on any items and instead, randomly respond to all items on the assessment. The literature review concludes by describing how the RRM can be adapted for polytomous responses in order to identify random responders in non-cognitive assessments.

Techniques for Addressing Low-Stakes Response Data

Motivation Filtering. As mentioned in Chapter 1, two popular ways of identifying low motivated examinees include self-reported motivation (e.g., the Motivated Strategies for Learning Questionnaire [Pintrich, Smith, Garcia & Mckeachie, 1993], Student Opinion Scale [Thelk, Sundre, Horst, & Finney, 2009], Test-Taking Motivation Questionnaire [Eklöf, 2006], etc.) and response-time effort measures (Wise & DeMars, 2005; Wise & Kong, 2005). Self-report motivation measures attempt to discern examinees' opinions about how important the test was to them and the amount of effort they exerted when completing it (Wise & DeMars, 2005). Response-time effort (RTE), which is described more fully later in the chapter, measures the amount of time examinees' spend completing each item on computer-based tests in an attempt to differentiate examinees with different levels of motivation (Wise & DeMars, 2005). Both self-report and RTE measures can be used along with a "cutoff score" to classify an examinee as low-motivated.

After a test has been administered and examinees with low-motivation or amotivation have been identified (either through self-report or RTE), motivation filtering can be used to remove responses from examinees who did not put forth effort on the test from the dataset prior to analysis (Steedle, 2014; Sundre & Wise, 2003; Wise & DeMars, 2005; Wise et al., 2006). This specific technique operates under the logic that responses obtained from low- or amotivated students bias aggregate test scores by underestimating overall examinee ability, and that the sub-sample of examinees retained after motivation filtering will provide a more accurate estimate of overall examinee proficiency (Sundre & Wise, 2003; Wise & DeMars, 2005). Of all the techniques described in this chapter,

more research exists regarding motivation filtering than any other technique and this technique also appears to be the most widely used in practice (Steedle, 2014, Sundre & Wise, 2003; Swerdzewski, Harmes, & Finney, 2011; Wise & DeMars, 2005; Wise et al., 2006).

Sundre and Wise (2003) conducted a seminal motivation filtering study using two cognitive tests and one self-report motivation instrument administered in a low-stakes, higher-education setting. A random sample of over 700 undergraduate students from a mid-sized university with complete data on the two cognitive tests and the self-report motivation scale was used in the study. To identify examinees with low motivation, scores and response patterns from a 10-item, self-report instrument, the Student Opinion Scale (SOS; Sundre & Moore, 2002), which purports to measure effort and importance, were used. Examinees achieving at or below particular a priori threshold values or exhibiting “suspect” response patterns on the SOS, were filtered out of the dataset incrementally. For both tests, as the threshold values increased, an increase in average test scores and a decrease in the standard deviation of scores was observed. Coefficient alpha and the standard error of measurement both decreased slightly as more problematic examinees were removed, but an increased correlation between SAT score and test performance occurred. The correlation between SOS and SAT scores held steady near zero as the various filter levels were applied, indicating no relationship between an examinee’s ability and level of motivation. In sum, the findings from this study indicate that motivation filtering is an effective technique for reducing bias in test scores caused by low examinee motivation.

Wise and DeMars (2005) conducted a similar study with 330 randomly assigned undergraduate students to determine if validity of the data after filtering out low-motivated examinees was greater than that of the data when unfiltered. In the study, the examinees completed a cognitive test followed by the SOS. Four different motivation filters were then applied to the data and subsequently, the impact of the filters on average test scores, reliability, and correlations between test scores and SAT scores and test scores and SOS scores were evaluated. Wise and DeMars (2005) also compared their findings to those of Sundre and Wise (2003). Consistent with Sundre and Wise (2003), Wise and DeMars (2005) found that the validity coefficients of the data after filtering out low-motivated examinees was greater than when unfiltered. Moreover, as more strict filters were applied, average test scores increased, the correlation between test scores and SAT scores increased, and the reliability of test scores held constant (Wise & DeMars, 2005). Importantly, there was no correlation between self-reported motivation scores and SAT scores, indicating that motivation and academic ability are not related. That is, no evidence was present that indicated motivation filtering eliminated examinee data due to low ability (Wise & DeMars, 2005).

Wise, Wise and Bhola (2006) also conducted a motivation filtering study that expanded upon Wise and DeMars' (2005) study by applying a variety of motivation filters to five different cognitive content domains (information literacy, fine arts, quantitative reasoning, history and political science and sociocultural) to investigate generalizability of the technique. In addition to supporting the findings of Sundre and Wise (2003) and Wise and DeMars (2005), Wise et al. (2006) concluded that motivation filtering could be generalized to other content domains.

Although research has demonstrated that motivation filtering is an effective strategy, disadvantages of the technique exist. One issue evident in the Wise and DeMars (2005) and Wise et al. (2006) studies is that of acquiring adequate sample size. The original dataset ($N=330$) presented by Wise and DeMars's (2005) was reduced by 65% ($N=114$) when the most strict filter was applied. Likewise, Wise et al. (2006) experienced similar results with sample size reductions ranging from 65% to 76%. Such a significant reduction in sample size may make statistical analyses and consequently interpretations of results difficult, especially if the initial sample size is not adequately large. Related to this issue is that of overfiltering the data. Overfiltering occurs when too many examinees are filtered from the dataset and scores become biased based on the ability of examinees (Wise et al., 2006). That is, examinees who are not necessarily unmotivated, but are simply low ability, are filtered out of the data set and the distribution of examinee ability is impacted.

Motivation filtering also requires that information regarding student motivation be collected in conjunction with the test (Wise et al., 2006). Additionally, three assumptions must be met prior to the use of motivation filtering: 1) a valid measure of examinee motivation levels from the testing period must be obtained (Sundre & Wise, 2003); 2) there must be no (or a very low) correlation between motivation and examinee ability (Steedle, 2014; Sundre & Wise, 2003; Wise et al., 2006); and 3) motivation must be related to performance on the test (Steedle, 2014; Wise et al., 2006). Unfortunately, it can only be determined if these criteria are met after examinee data have been collected (Wise et al., 2006).

Other issues with motivation filtering include its use of a cut score to classify examinees by motivation level and that an additional measure, such as self-report or RTE (which will be described in more detail in the next section), is required to detect low- or unmotivated examinees (Wise & DeMars, 2005). Cut scores used with self-report and RTE measures have their own psychometric issues that must also be considered. For example, self-report instruments may not be accurate measures of motivation if examinees respond randomly or untruthfully (Grove & Geerken, 1977; Wise & DeMars, 2005; Wise & Kong, 2006) or if a spurious relationship between motivation and performance exists due to a shared correlation with examinee ability (Wise & DeMars, 2005; Wise et al., 2006). Furthermore, cut-scores are fairly arbitrary, as they are established by human judgment and techniques used to determine cut-scores have been shown to produce non-consistent results (Hambleton, 2012; Kane, 2012). Furthermore, cut-scores are somewhat sample-dependent, depending on how they are derived.

RTE. Instead of directly asking examinees to report their motivation levels through the use of a self-report instrument, an unobtrusive alternative is to collect data pertaining to the length of time it took examinees to respond to each item on the test. Response Time Effort (RTE) is a measure of motivation that assumes examinees who are motivated will respond to items using *solution behavior*, which requires adequate time to read each item and consider the available response options¹ (Schnipke 1995, 1996; Schnipke & Scrams, 1997; Swerdzewski et al., 2011; Wise & Kong, 2005). Conversely, RTE also assumes that an unmotivated examinee will respond to items using *rapid-guessing behavior*, which involves responding without taking sufficient time to consider

¹ Schnipke (1995, 1996) and Schnipke and Scrams (1997) discussed *solution behavior* in terms of test speededness, but not with respect to RTE.

the item and response options (Schnipke 1995, 1996; Schnipke & Scrams, 1997; Swerdzewski et al., 2011; Wise & Kong, 2005). RTE is defined as the proportion of items on an instrument for which an examinee is thought to have answered using solution behavior (Wise & Kong, 2005). In other words, it is the proportion of items with response times exceeding a set threshold (Swerdzewski et al., 2011; Steedle, 2014). Using RTE, test administrators can collect data regarding length of time it took an examinee to respond to each item on the test, along with their selected responses. This information can then be used to classify examinees as motivated or amotivated by specifying item thresholds to determine the presence of rapid-guessing behavior (Wise & Kong, 2005). By classifying examinees, the data collected from amotivated students can be eliminated from the culminating data analysis to prevent contamination of results with associated construct-irrelevant variance; that is, motivation filtering can be conducted using RTE.

Wise and Kong (2005), conducted a seminal RTE study using a cognitive, computer-based test administered in a low-stakes, higher-education setting. The sample included 472 randomly selected freshmen students from a mid-sized university. To identify examinees' academic proficiency, their Verbal and Quantitative SAT scores were obtained from a university database. Moreover, the SOS (Thelk et al., 2009) was electronically administered following the cognitive instrument and used as an additional measure of examinee motivation. Graphs of examinee response times were visually examined and for all items the distribution appeared bi-modal. Wise and Kong (2005) hypothesized that the smaller of the two modes that peaked at the lower response time was indicative of rapid-responding. They also noticed that the width of the part of the

distribution associated with the lower mode varied depending on the length of the item, with longer items having longer widths. Together this information was used to set cut points on each item; for instance, items with less than 200 characters had a threshold of 3 seconds whereas items with more than 1000 characters had a threshold of 10 seconds.

Wise and Kong (2005) found RTE scores to be reliable and were able to provide evidence of both convergent and divergent validity for the scores. RTE was found to be positively correlated with the self-report SOS data and almost uncorrelated with SAT scores. When motivation filtering was performed using the RTE and self-report SOS data, similar results were found: average scores on the test increased and the correlation between total test score and SAT scores increased. Even though the general trends were the same with motivation filtering using RTE and self-reported SOS data, RTE tended to remove fewer examinees and have slightly more favorable results (e.g. larger increases in means and correlations between total test scores and SAT scores).

Swerdzewski, Harmes and Finney (2011) conducted a study that expanded upon Wise and Kong's (2005) initial study by examining RTE and SOS data collected after each test given in a series of cognitive and non-cognitive tests. Swerdzewski et al. (2011) also explored an additional measure of RTE, global RTE, which spans the entire series of tests, as well as changes in the levels of test-level RTE and self-report SOS data over the battery of tests. A random sample of 303 second-year undergraduate students from a mid-sized institution completed a series of tests administered in a low-stakes, higher-education setting. Each examinee completed six to seven tests that varied in content and length. Of the tests administered, at least two were cognitive and four were non-cognitive in nature. At the end of each battery, each examinee was also required to

complete the SOS (Thek et al., 2009). Three motivation indices: test-level RTE, global RTE and global SOS were used to categorize examinees as motivated or unmotivated. To calculate test-level RTE, a cut score of 0.90 (i.e. 90% of items were completed using solution behavior) was selected based on Wise and Kong's (2005) study. The researchers acknowledged that the selection of .90 was fairly arbitrary. In the calculation of global RTE students were classified as unmotivated if at least one test-level RTE for their set of tests fell below 0.90.

Swerdzewski et al. (2011) found that approximately 66% of the examinee sample was classified consistently across methods (e.g. global RTE vs. global SOS, global SOS vs. test level RTE). Additionally, the researchers found the pattern of changes in the aggregate test scores that were similar with both self-report and RTE, thus concluding the two methods have equal utility. This finding was contrary to that of Wise and Kong (2005) who interpreted the differences in aggregate test scores when motivation filtering was employed between the two methods to be meaningful. Swerdzewski et al. (2011) also found the methods differed in the number of examinees they excluded from the data set, with the self-report measure removing more data than RTE. Thus, RTE appears to be a more parsimonious method than the self-report method.

Although research has demonstrated that RTE is an effective strategy, disadvantages of the technique exist. For instance, in order to acquire item response times, the test must be administered electronically; there is no way to accurately measure response time with paper and pencil tests. Computer-based testing may not be practical in all testing situations, especially if a large number of examinees are expected and resources are limited. Another issue lies with the assumption that an examinee who

exhibits rapid-guessing behavior by quickly responding to an item has low motivation. That is, examinee response time serves as a proxy for motivation (Swerdzewski et al., 2011). However, it is possible that a rapid responder may be a motivated examinee with faster than average processing speed (DeMars, 2007). Thus, RTE could potentially misclassify such motivated examinees.

In order to distinguish groups of examinees into motivational categories, item response time, a continuous variable, must be dichotomized and a cut-score, also referred to as a threshold, established in order to make such categorizations. Currently, multiple methods, such as visually inspecting plots (Schnipke, 1995) and distributions (Kong Wise, & Bhola, 2007; Wise, 2006) of response time frequency, using item surface information (Wise & Kong, 2005), setting a common threshold (Wise, Kingsbury, Thomason, & Kong, 2004; Kong et al., 2007), and using IRT mixture models (Kong et al., 2007; Schnipke, 1996; Schnipke & Scrams, 1997; Wise & DeMars, 2006), among others (Rios, Liu, & Bridgeman, 2014), exist for establishing a cut point, but there is no set standard. With no set standard, derivation of a cut-point depends upon the method selected for setting the threshold. Innumerable standard setting techniques exist throughout the literature, but again, no one method in particular is championed. Furthermore, loss of examinee information occurs with this strategy due to dichotomization of response time (MacCallum, Zhang, Preacher, & Rucker, 2002). However, dichotomization is necessary with RTE, as DeMars (2007) explained,

If motivation could be assumed to increase with response time, then response time itself, rather than the dichotomization of response time into rapid-guessing versus solution behavior, could be used in the model to capture varying degrees of

motivation. However, once a student has passed the threshold of adequate time to read the item, additional time spent may plausibly be due to differences in processing speed rather than to differences in effort (p.42).

Statistical Models. Self-report and RTE measures are used for the purpose of identifying students with low motivation and filtering them out of the dataset. In order to use these methods, additional examinee information must be gathered (e.g. response times or self-report data). An advantage of many statistical models is that they do not require the collection of supplementary information. Moreover, in contrast to self-report measures and RTE, these models estimate model parameters, such as theta and item difficulty and discrimination, while simultaneously accounting for the presence of low or amotivated examinees instead of completely eliminating them from the dataset.

At present, a plethora of statistical models exist for the purpose of detecting and accounting for low motivated examinees. Some statistical models actually integrate response time into the model. For example, the effort-moderated IRT model incorporates response time data with the 3-parameter logistic (3PL) IRT model to account for examinee rapid-guessing and provide more valid estimates of ability than a traditional 3PL IRT model (DeMars, 2007; Wise & DeMars, 2006). Essentially, the effort-moderated IRT model combines the item response functions for the probability of a correct response to an item using solution behavior (i.e. a traditional IRT model) and probability of a correct response to an item using rapid-guessing behavior (i.e. chance) into a single model (Wise & DeMars, 2006). This model is moderated by examinee response strategy for each item. For instance, if the response time for an item indicates that an examinee most likely engaged in solution behavior, the function for the traditional

IRT model is used. Conversely, if the response time indicates the examinee engaged in rapid guessing behavior, the probability of a correct response is set equal to chance level. Thus, the function used to model an examinee's response to an item is determined by their response time classification. This model is more flexible than other models that will be described in that examinees can switch from solution behavior to rapid-guessing behavior and back again. Other statistical models for identifying or controlling problematic examinee behavior due to low motivation that utilize response time include those proposed by Bovaird (2002), Meyer (2010) and Yang (2007). However, as with any use of response time, the assessment must be delivered electronically, which is not always practical. For this reason, statistical models incorporating response time will not be discussed further.

Other statistical models, known as partial guessing models, can capture different kinds of low motivated examinee response behaviors such as guessing on the hard items, a gradual decline in effort, or a sudden abandonment of solution behavior. These models do not utilize response time information; all the information that is needed to estimate the models are the students' scored responses to the items. These models are described as partial guessing models because some examinees are using solution behavior throughout, while other examinees exhibit guessing behavior in some form. The IRT difficulty-based guessing model (IRT-DG) assumes examinees guess on the more difficult items for their ability level, but try to answer the easy items (Cao & Stokes, 2008). As with the effort-moderated IRT model, examinees can switch their strategy from solution behavior to this kind of guessing behavior multiple times. Because multiple switches in behavior can be present, detecting this guessing pattern can be quite difficult.

Other partial guessing models suggest that the probability of answering an item correctly is related to the item's location on the test (Cao & Stokes, 2008). Such models include the IRT threshold guessing model (IRT-TG) and the IRT continuous guessing model (IRT-CG). Both models assume there are two types of examinees: motivated and unmotivated. The motivated examinee is thought to try on all items, whereas the unmotivated examinee is thought to decline in motivation throughout the testing session (Cao & Stokes, 2008). This decline can be contributed to fatigue or loss of interest in the test and does not necessarily result in guessing, but does result in low-effort and a decreased probability of a correct response. The IRT-TG model assumes examinees initially start out exhibiting solution-based behavior on the test, but suddenly switch abruptly over to guessing behavior (Cao & Stokes, 2008). For each examinee, this model specifies an item location threshold, or the point at which this switch occurs. That is, the IRT-TG model estimates the item in which the examinee switches behaviors and begins to guess (Cao & Stokes, 2008). This model is the same as Yamamoto (1995)'s HYBRID model, which was developed to model the behavior of examinees on speeded tests. (Yamamoto's model will be described more fully later in the chapter as the Random Responding Model and can be considered a constrained form of this model.)

Like the IRT-TG, the IRT-CG allows examinees to switch over to guessing behavior at some point in time in the test. In this model the switch is not abrupt, but instead characterized by a steady decline in valid response behavior as the test progresses. Models similar to the IRT-CG include those proposed by Goegebeur, DeBoeck, Wollack and Cohen (2008) and Jin and Wang (2014).

The IRT-TG, -CG, and -DG are all statistical models appropriate for identifying examinees that are at least somewhat motivated. At this point in their development, they can only be applied to dichotomously scored data. There are other, simpler models that distinguish motivated examinees from those that are not motivated. The latter examinees are considered amotivated, and are essentially randomly responding from the very beginning of the assessment onward. In order to better understand these models, which are the focus of this dissertation, a brief discussion of mixture modeling is necessary.

Item Response Theory (IRT) Mixture Modeling. IRT mixture modeling is a technique that can be used to capture the presence of unobserved differences between unknown groups (i.e. classes or subpopulations) of examinees in item responses. This technique permits unobserved heterogeneity of item and test characteristics not identified a priori to be examined by allowing IRT model parameters to vary across classes (Rost, 1990). Traditional IRT models assume all examinees come from the same population. Therefore, a single set of item parameters are appropriate. In contrast, IRT mixture models assume that examinees come from multiple subpopulations, with each subpopulation requiring its own unique set of item parameters (Rost, 1990). In other words, with IRT mixture modeling, the observed data are hypothesized to represent a mixture of distinct groups.

Mislevy and Verhelst (1990) provided an example of an IRT mixture model, specifically a two-class Rasch model, with the purpose of demonstrating the ability of such a model to capture heterogeneity of item responses. In their example, examinees were able to solve items on an instrument by using one of two possible strategies (e.g. rotation or matching features). A mixture IRT model was needed in this situation

because all examinees may not use the same strategy to solve the items, and the mixture model allowed for differences in item difficulty due to the use of different strategies. Thus, this model allows item difficulties to vary across the two groups, without knowing a priori which examinees were using which strategy. Each examinee receives a theta estimate and posterior probabilities of membership in each class. If an examinee's posterior probability for a class is low, their theta estimate for that class may not be trustworthy.

IRT mixture models are not limited to only two groups of examinees or situations in which different strategies are being used. IRT mixture models can be used in any context in which IRT model parameters (difficulties, discriminations, theta means or variances) are thought to vary across unknown groups. For instance, a DIF analysis where group membership is not known is a situation where model parameters would be thought to vary across unknown groups and in fact, IRT mixture models have been proposed for this purpose (Cohen & Bolt, 2005; DeAyala, Kim, Stapleton & Dayton 2002).

An equation representing a 2PL IRT mixture model with two classes is shown in Equation 1 using the factor model parameterization (Kamata & Bauer, 2008)². This model indicates that the marginal probability of an examinee's correct response to an item ($P(X_i=1)$) is the weighted sum of the conditional probability of obtaining a correct response in each class, which is equal to a 2PL model with class-specific parameters.

² In IRT parameterization, a is discrimination, b is difficulty, and the correspondence between factor model parameters of loadings (λ) and thresholds (τ) to discrimination (a) and difficulty (b) is:

$$\begin{aligned} l_i &= a_i \\ t_i &= a_i b_i \end{aligned}$$

$$P(u_i = 1) = \pi_1 \left| \frac{\exp(-\tau_{1i} + \lambda_{1i}\theta_{1j})}{1 + \exp(-\tau_{1i} + \lambda_{1i}\theta_{1j})} \right| + \pi_2 \left| \frac{\exp(-\tau_{2i} + \lambda_{2i}\theta_{2j})}{1 + \exp(-\tau_{2i} + \lambda_{2i}\theta_{2j})} \right| \quad (1)$$

The weight for each class (π_1, π_2) represents the proportion of examinees in the population contained in the class. In a two class solution, only one class weight is estimated because the weights are constrained to sum to 1.0 (e.g., $\pi_1 = 1 - \pi_2$).

With an IRT mixture model, more than one class can be specified. In deciding which model to retain, model fit indices such as information criteria (e.g. AIC, BIC) are often used along with a priori expectations and interpretability of the solutions. If only one class is retained, the mixture model reduces to a traditional IRT model. Thus, IRT models are nested within mixture models. That is, IRT models are more parsimonious forms of mixture models.

Although Mislevy and Verhelst's (1990) IRT mixture model specifies that each class follows an IRT model, others specify different kinds of models for varying classes. For instance, the HYBRID model presented by Yamamoto (1989) allows one class to follow an IRT model, and a second class to follow a latent class model³. In the latent class model, item responses are a function of item thresholds, but not of item loadings or the examinee's theta level. Yamamoto's (1989) HYBRID model is shown in Equation 2, with the conditional probability of a correct response in the first class represented using an IRT model and the conditional probability of a correct response in the second class represented using a latent class model.

³ A full latent class model uses a latent categorical variable to model relationships between dichotomous variables that are observed, whereas a factor model uses a latent continuous variable.

$$P(u_i = 1) = \pi_1 \frac{\exp(-\tau_{1i} + \lambda_{i1}\theta_1)}{1 + (\exp(-\tau_{1i} + \lambda_{i1}\theta_1))} + \pi_2 \frac{\exp(-\tau_{2i})}{1 + (\exp(-\tau_{2i}))} \quad (2)$$

IRT mixture models that specify different kinds of models for varying classes can be used in situations where two or more qualitatively different classes of examinees are present to make quantitative comparisons among examinees in each class. For example, if examinees in class 1 are using one solution strategy and examinees in class 2 are using a different solution strategy, quantitative comparisons among those examinees within each class, such as levels of ability (i.e. theta estimates), can be made. To take the example further, perhaps the examinees in class 1 vary in their ability levels, but examinees in class 2 do not. That is, examinees in class 2 all have the same ability level. This situation would be equivalent to constraining the theta variance in the IRT mixture model presented in Equation 1 to 0, which results in the HYBRID model presented in Equation 2.

The next model, a full latent class model presented in Equation 3, only reflects qualitative differences. There is no within class variability; thus, no quantitative differences in examinee ability exist within each class. Essentially, this is the same as setting the factor variance to zero for each class in a full IRT mixture model. In essence, Equations 2 and 3 could be thought of as constrained, more parsimonious versions of the IRT mixture model presented in Equation 1.

$$P(u_i = 1) = \pi_1 \frac{\exp(-\tau_{1i})}{1 + (\exp(-\tau_{1i}))} + \pi_2 \frac{\exp(-\tau_{2i})}{1 + (\exp(-\tau_{2i}))} \quad (3)$$

Some researchers adopt an exploratory approach, where many of the models previously discussed are fit to the data with varying numbers of classes. With this approach, model fit indices, a priori expectations, and interpretability of the solution are

used to guide model selection. Other researchers are more intentional and select a particular specification, often with a particular number of classes, to describe examinee behavior. For instance, an extended form of Yamamoto's (1989) HYBRID model was proposed for use with speeded tests by Yamamoto (1995), where examinees switch from valid responding to random responding behavior due to time limitations. The extended form of Yamamoto's HYBRID model is different than the HYBIRD model shown in Equation 2, in that item thresholds in the latent class are constrained to be a function of guessing on the item. The item thresholds, which will be referred to as guessing thresholds (g) from here on out, for items in the latent class are constrained to be a function of the number of response options for an item (r_i). That is, the probability of a correct response for an item if an examinee randomly responds is $1/r_i$. The associated threshold is equal to g_i which is calculated using Equation 4:

$$g_i = -\ln \left[\frac{\frac{1}{r_i}}{1 - \left(\frac{1}{r_i}\right)} \right] \quad (4)$$

As an example, if an item has three response options, $1/r$ would equal .33 and the guessing threshold (g) would equal .69317.

Another difference is that a parameter is included in the model to characterize the item at which an examinee switches from the valid responding class (IRT class) to the random responding class during the test. This model has been used in low-stakes testing to identify the point at which examinees switch from valid responding to random responding behavior (e.g. Cao & Stokes, 2008).

A simplified version of the model uses Equation 2 and constrains the thresholds in the latent class to be a function of guessing on the item (Equation 4), but does not estimate the "switch" point. This model can be used when examinees are considered to

engage in the same response behavior (either valid responding or random responding) for the entire test and is called the Rapid Responding Model (RRM) in this dissertation.

Specifically, this dissertation will focus on the applicability of the RRM for use with polytomous data, because it has yet to be used in practice for this purpose. Thus far, the RRM has only been used with dichotomous data.

RRM with dichotomous items in low stakes testing. The RRM was first presented by Mislevy and Verhelst (1990), who demonstrated how item parameters changed when using a two-class RRM versus a one-class Rasch IRT model with a sample of 1,906 examinees in a low-stakes testing setting. The sample was visually observed to include amotivated examinees. For example, it was reported that some didn't even open the testing booklet, yet provided responses on the answer sheet. The analysis included 12 dichotomous items with four alternative options. Mislevy and Verhelst (1990) found that the RRM ($-2LL = 2,606$) fit better than the one-class Rasch model ($-2LL = 2,752$)⁴ and the proportion of valid responders was estimated to be 0.955, indicating that 4.5% of the examinees were randomly responding on all items.

Mislevy and Verhelst (1990) also compared the item difficulties obtained using the one-class Rasch versus those obtained for the valid responding class using the RMM. This is an important question to answer as it indicates how item difficulties might be impacted in the one-class Rasch model when random responders are present in the data. They found that the item difficulties of the Rasch and RRM model were related monotonically. That is, little difference in item difficulties was seen between the two

⁴ The $-2LL$ values will always look better for the more complex model when models are nested. Information criteria and appropriate likelihood ratio tests (LRTs) (e.g. bootstrap LRT, Lo-Mendell Rubin LRT, etc.) are typically used when comparing models that differ in number of classes. Mislevy and Verhelst (1990) only used magnitude of difference in $-2LL$ values for model selection.

models with the harder items, but large differences were present with easier items.

Overall, it was found that the presence of random responders makes items (especially the easier ones) look harder in the one-class Rasch compared to the RRM.

Mislevy and Verhelst's (1990) applied example with the RRM illustrates two of its advantages over using the one-class IRT model and ignoring the presence of random responders in the data. In the one-class IRT model, IRT item parameters are estimated including the random responders, which results in tainted item parameters. The RRM estimates IRT item parameters, controlling for the presence of random responders (thus purifying the item parameter estimates). It can also be used to identify random responders in the data. Use of the RRM also has advantages over motivation filtering techniques. The RRM is advantageous in that it models amotivated examinee data by weighting them differently than motivated examinee data instead of deleting them (Lau, 2009). This technique is also more parsimonious than other statistical techniques for addressing low-stakes response data (e.g., a 3PL IRT model), in that only one additional parameter, the weighting parameter, is estimated (Lau, 2009; Mislevy & Verhelst, 1990). Furthermore, with other statistical techniques, such as the 3PL IRT model, thetas differ by degree and all thetas are comparable to one another; however, in the RRM, thetas are only comparable for examinees in the IRT class. To elaborate further on this point, in a Rasch model, all examinees with the same number of items correct will have the exact same theta. This includes both valid and random responders. Thus, we say that they have the same ability. With the RRM, these two examinees will have the same theta in the IRT class, but very different posterior probabilities of membership in the IRT class.

It is the posterior probabilities that indicate just how trustworthy their theta estimates are in the IRT class.

Using the Mislevy and Verhelst (1990) example as a model, Lau (2009) conducted two studies employing the RRM: a Monte Carlo simulation and an applied study using real archival data acquired from administration of a low-stakes test to sophomore and junior undergraduates at a mid-sized, public, southeastern university.

Study 1, the Monte Carlo simulation, was conducted to determine the utility of Mislevy and Verhelst's (1990) proposed model. More specifically, the purpose of Study 1 was to explore the effect of random responders on IRT parameter estimates as well as the efficacy of the RRM for detecting and accounting for amotivated examinees on cognitive instruments in low-stakes testing settings. In Study 1, the one-parameter logistic (1PL) and two-parameter logistic (2PL) IRT models were fit to simulated data consisting of both valid and random responders to answer *RQ 1: "How well are item parameters estimated when fitting a one-class model to a mixture of valid responders and random guessers?"* In Study 1, Lau fit the RRM to the same data to address the following two research questions: *RQ2: "How well are item parameters estimated when fitting a two-class model to a mixture of valid responders and random guessers?"* and *RQ 3: "Does the two-class model fit data that is a mixture of valid responders and random guessers substantially better than the one-class model?"* That is, Lau (2009) fit a one-class 1PL, a one-class 2PL, a two-class RRM 1PL and a two-class RRM 2PL to the data. Varying proportions of amotivated simulees (.9%, 9% and 20%) were incorporated into a large data set of valid responder simulees and the impact of amotivated examinees on IRT parameter estimates was evaluated by type of measurement model (1PL or 2PL),

proportion of amotivated examinees (.9%, 9% and 20%), and number of classes (1 or 2). In this simulation study, Lau (2009) conducted a total of 1,200 analyses.

The one-class models were fit to data that included both amotivated and motivated examinees. With the 1PL model, as the proportion of amotivated examinees in the data set increased, bias, percent bias and RMSE values increased for the item difficulty parameters. Because bias and RMSE were similar in value, it was concluded that bias was more of an issue than precision. Specifically, in comparing estimated parameters with true parameters, Lau (2009) found that the direction and magnitude of bias for item difficulty depended on the true difficulty of the item. For example, items with thresholds above 1 appeared to be easier than their true value when amotivated examinees were present, whereas items with thresholds below 1 appeared harder. Moreover, estimation of harder items was less biased than estimation of easier items, as easier items tended to be more biased with greater proportions of amotivated examinees. That is, the bias appeared more pronounced with larger proportions of random responders.

Results of fitting the 2PL model to the dataset were similar to those of the 1PL in that the greater the proportion of amotivated examinees, the weaker the recovery of parameters. With the 2PL model, Lau (2009) found that the magnitude of bias with factor loadings was greater for items that were more discriminating, and that the bias could be positive or negative. Specifically, the direction of the bias (e.g. positive or negative) was found to be associated with the difficulty of the item. That is, easy items were found to have a positive bias (i.e. discriminations were overestimated), whereas more difficult items had a negative bias (i.e. discriminations were underestimated).

Across all three proportions of amotivated examinees, items that were on the extreme ends of the difficulty continuum (i.e. extremely easy or hard) were harder to accurately estimate than items toward the middle of the continuum.

In contrast to the one-class models, the two-class models estimated one additional parameter, the proportion of examinees categorized in each latent class. In fitting the 1PL RRM to the data, the class proportions of all of the amotivated population conditions (.9%, 9% and 20%) were underestimated by a small percentage. As well, the classification accuracy in each condition was considered to be sufficiently high (average entropy $> .90$). In contrast to Lau's (2009) hypothesis, greater classification accuracy was found for the .9% amotivated examinee condition (.99) than for the 20% amotivated examinee condition (.96). In regard to item parameter estimation, the use of two separate classes for examinees resulted in more accurate and less biased item thresholds that more closely resembled true values. Specifically, in the .9% condition, bias was close to zero and was only -0.004 in the 20% condition. To compare back to the one class 1PL model, bias was 0.32 in the 20% condition. Thus, bias decreased from 0.32 to approximately 0 with the use of two classes instead of one.

In fitting the 2PL RRM to the data, the class proportions of all of the amotivated population conditions (.9%, 9% and 20%) were estimated accurately and the classification accuracy for each group was considered to be sufficiently high (average entropy $> .90$). As with the 1PL RRM, greater classification accuracy was found for the .9% amotivated examinee condition (.99) than for the 20% amotivated examinee condition (.96). Again, the use of two separate classes for examinees resulted in more accurate and less biased item parameter estimations. As a matter of fact, irrespective of

the proportion of amotivated examinees present, the model did a good job of estimating item parameters. Bias values were around zero for both factor loadings and thresholds.

Using model comparison indices (e.g. LL, AIC, BIC and SSA-BIC) that will be further described in Chapter 3, Lau (2009) compared the one-class and two-class models to determine if one fit better than the other. For both the 1PL and 2PL models, each model comparison index showed improvement with the addition of the second class, which supports inclusion of an additional class. In addition, as the proportion of amotivated examinees increased, the difference in fit between the one- and two- class models was greater, indicating even more support for the use of a two-class model when large proportions of amotivated examinees are present.

Study 2, the applied study, was conducted to demonstrate the application of the RRM to authentic cognitive data collected in a low-stakes setting. More specifically, the purpose of this study was to corroborate the results of using the RRM on real test data with those of the simulated data, and to add evidence of utility of the technique. In Study 2, a one-class 2PL and two-class RRM 2PL model were fit to authentic low-stakes data acquired from sophomore and junior undergraduates at a mid-sized, public, southeastern university to answer *RQ 4: What proportion of examinees are classified as amotivated?*, *RQ 5: How certainly can random and valid responders be distinguished from one another?*, *RQ 6: Which model best fits the data (2PL IRT or RRM)?*, *RQ 7: Do greater differences exist between classes in test-taking motivation or ability level?*, and *RQ 8: Are examinees still categorized in the same classes if the RRM is fit to only a portion of the items?*

The data utilized in this study (Lau, 2009), was collected from 4,391 undergraduate sophomores and juniors (students earning 40 to 75 credit hours) between the years of 2002 and 2006 who were required by their university to participate in a campus-wide testing series designed to assess general education and student affairs programs. The results of the testing series held no consequences for individual examinees, as scores were used in the aggregate; thus, the testing context was low-stakes and it was assumed that amotivated students were present. Within the series of tests, examinees completed the Global Experience (GLEX) instrument and Student Opinion Survey (SOS; Sundre & Moore, 2002). The GLEX is a 32-item, multiple choice, cognitive instrument assessing knowledge of global history with three to five response options. The 2PL and RRM models were applied to the data collected from this instrument. The SOS is a 10-item, non-cognitive instrument that uses a five-point Likert scale. Data collected from this instrument was used as a measure of test-taking motivation.

Examinee classification with the two-class RRM 2PL model.

In applying the RRM to real data, Lau (2009) found that approximately 1.2% of examinees were classified as amotivated by evaluating three methods of determining class membership: model-based (1.28%), posterior probabilities (1.30%), and modal assignment (1.18%). This proportion of examinees is less than the 4.5% found by Mislevy and Verhelst (1990). To provide validity evidence for the classes, Lau (2009) examined descriptive statistics for the total score of the GLEX for each class. It was found that the mean total score on the GLEX for the amotivated class was around chance level, Cronbach's coefficient alpha was negative, which demonstrated that item responses

were not correlated in general, and total score variance was lower than in the valid responder class, which indicated that fluctuation in scores was more likely due to chance than to systematic variance. Moreover, classical item statistics showed item difficulties to be around chance level and item discriminations to be at zero or a negative value for the amotivated class. Classification accuracy was also examined via classification table and the entropy statistic. Overall, classification accuracy was good; entropy for the model was .983, which is close to 1. Even though overall classification accuracy was good, classification of amotivated examinees was relatively more difficult than classification of motivated examinees. Specifically, the average posterior probability associated with the motivated class for examinees classified in the amotivated class was higher (average probability of .16) than the average posterior probability (average probability of less than .003) associated with the amotivated class for examinees classified in the motivated class.

Comparison of the one-class 2PL and two-class RRM 2PL models.

Models were compared based on relative fit indices (LL, AIC, BIC, SSA-BIC), likelihood ratio tests (LMR and Bootstrap LRT), and changes in item parameter statistics. All of the relative fit statistics were lower for the two-class model than for the one-class model. Although values will always look better for the more complex model when models are nested, only one additional parameter was estimated with the RRM, thus it was concluded that the change in indices given the small difference in the complexity of the models provided evidence of heterogeneity in the data. The likelihood ratio tests corroborated support for use of the two-class model, as the LMR and Bootstrap LRT test statistics were 101.59 and 113.703 respectively, with probabilities less than .0001.

In comparing item parameter estimates, factor loadings decreased .05 logits on average with the inclusion of the additional latent class for amotivated examinees. The change in factor loadings when going from a one-class to a two-class model was generally greater for easier items than more difficult items. Incorporating the latent amotivated class also resulted in threshold values that were an average of .01 lower than with the one-class model, indicating very little change in the threshold estimates. Recall that in Mislevy and Verhelst's (1990) applied example, they found little differences between item difficulties estimated using the 1- and 2-class models for the harder items, but larger differences with easier items. In contrast to Mislevy and Verhelst's (1990) findings, Lau (2009) found that items appearing more difficult when using the mixture model were on both of the extreme ends of the difficulty continuum (very easy and very hard), thus resulting in a curvilinear relationship between the threshold values of the one-class model and change values. Perhaps this discrepancy has to do with the differences regarding the measurement model used in the study. That is, Lau (2009) used a 2PL model, whereas Mislevy and Verhelst (1990) used a 1PL.

To provide further evidence that the classes detected in the RRM were those of motivated and amotivated examinees and not some other group of individuals, such as high-ability and low-ability examinees, Lau (2009) explored applicable validity evidence related to test-taking effort and academic ability. As previously stated, test-taking effort was measured using the self-report SOS measure of effort and scores ranged from 5 to 25. The amotivated class had lower mean scores (12.767) than the motivated examinees (17.198) and the difference was statistically significant ($p < .0001$). Regarding academic ability, GPA was compared. Again, the amotivated class had lower GPAs (2.67) than the

motivated examinees (2.95) and the difference was statistically significant ($p < .0001$). The difference in academic ability between classes was not congruent with Lau's (2009) hypothesis that class membership was due to differences in motivation only (e.g. motivated and amotivated classes). The strongest support for class membership being due to motivation only would have been no difference in academic ability between the two classes. Lau (2009) also divided the item set in half (items 16 to 32) and re-analyzed the data to determine if the same examinees were still categorized in the same classes, as this would be an indicator that the amotivated examinees were truly amotivated. It was found that 98.2% of examinees were classified the same way regardless of using the first-half or second-half of the test. It was also found that effort scores for the motivated class were consistently higher than for the amotivated class regardless of item set used.

Swanson (2013) also performed an applied study using a one-class 2PL and a two-class RRM 2PL model to examine the proportion of random responders detected, as well as which model fit the data best. The data collected for this study was similar to that of Lau (2009). The sample contained 805 undergraduate sophomores and juniors (45 to 70 credits) who were required by their university to participate in a campus-wide testing series designed to assess general education and student affairs programs in February of 2013. The results of the testing series held no consequences for individual examinees, as scores were used in the aggregate; thus, the testing context was low-stakes and it was assumed that amotivated students were present. Within the series of tests, examinees completed the Sociocultural Dimension Assessment - Version 6, (SDA-6), which is a 32-item cognitive instrument. Each item contained three to five dichotomously scored, multiple-choice options.

In this study, both a 2PL model and the RRM were applied to the data. Results of the application of the RRM revealed that approximately 1.62% of examinees were classified as amotivated. Classification accuracy was examined via classification table and the entropy statistic. Overall, classification accuracy was good; entropy for the model was .977, which is close to 1. Even though overall classification accuracy was good, classification of amotivated examinees was relatively more difficult than classification of motivated examinees. Specifically, the average posterior probability associated with the motivated class for examinees classified in the amotivated class was higher (average probability of .201) than the average posterior probability (average probability of less than .003) associated with the amotivated class for examinees classified in the motivated class.

The models were compared by evaluating changes in parameter estimates and relative fit indices. In evaluating changes in parameter estimates with the addition of the second class, factor loading estimates decreased by approximately .067 logits on average and the change in factor loading values was found to be greater for easier items than for harder items.

On average, threshold estimates decreased by approximately .019 logits and a curvilinear relationship was found between the 2PL threshold value and the change in threshold values. That is, while most of the items appeared easier with the RRM than with the one-class 2PL model, items at the extreme ends of the threshold scale appeared to be more difficult. In evaluating the change in relative fit indices, the Information Criteria (IC) for the RRM were all smaller than the IC for the 2PL, which provided evidence that the data were heterogeneous. The LMR ratio test was also conducted to

compare the fit of the two models. The results indicated a need of at least a two-class model to describe the data ($p < .001$). The findings from this study were similar to those found in Lau's (2009) applied study.

Another applied study using comparable samples was conducted by Swanson and Pastor (2014) and provided similar results. The purpose of this study was to estimate the proportion of amotivated examinees across a variety of low-stakes assessments and to ascertain which model, a one-class 2PL or a two-class RRM 2PL, best fit the data. For this reason, differences in parameters between the two models were not investigated. The data collected for this study was similar to that of the previous studies by Lau (2009) and Swanson (2013). Multiple samples containing undergraduate sophomores and juniors (45 to 70 credits) who were required by their university to participate in a campus-wide testing series were utilized. Within the series of tests, examinees completed either the Natural World - Version 9, (NW-9), which is a 66-item cognitive instrument, containing three to five dichotomously scored, multiple-choice options, or the American Experience – Version 2 (AMEX2) which is a 40-item cognitive instrument, containing five dichotomously scored, multiple-choice options. All data was archival and varied in collection date and sample size. Specifically, both the 2PL and RRM 2PL models were fit to data collected from the regularly scheduled testing series in spring 2013 (NW-9, $N = 1,404$) and spring 2012 (NW-9, $N = 1072$; AMEX2, $N = 1015$). The models were also fit to data collected from a make-up testing session from spring 2012 for students who were unable to attend the regularly scheduled session (NW-9, $N = 178$). It was hypothesized that the sample from the make-up testing session would contain a higher proportion of

amotivated examinees than the samples collected on the regularly scheduled day because these examinees failed to attend the first mandatory session.

Results of the application of the RRM to the NW-9 revealed that approximately .64% to .89% of examinees were classified as amotivated in spring 2012 and spring 2013 respectively. When applied to the AMEX2 data, the model failed to converge, potentially because the class weight was too close to zero to estimate. In applying the RRM to the make-up data, 7.48% of examinees were found to be amotivated. This may be contributed to the fact the examinees may have been less motivated to put forth good effort on low-stakes assessments considering they missed the first mandatory session.

The models were compared by evaluating relative fit indices and a likelihood ratio test. In evaluating the change in relative fit indices, the Information Criteria (IC) for the RRM were all smaller than the IC for the 2PL (with the exception of the AMEX2, which did not converge), which was expected and provided evidence that the data were heterogeneous. The LMR ratio test was also conducted to compare the fit of the two models for each data set. The results indicated a one-class model was adequate to describe the data (NW-9 2013, $p = .37$; NW-9 2012 $p = .19$; NW-9 2012 make-up, $p = .78$). These findings are not consistent with Swanson 2013 and Lau 2009 in that the RRM was not championed.

Overall, Lau's (2009), Swanson's (2013) and Swanson and Pastor's (2014) studies help to demonstrate the effect of random responders on IRT parameter estimates as well as the efficacy of the RRM for detecting and accounting for amotivated examinees on cognitive instruments in low-stakes testing settings. In all of the studies, the RRM was found to have good classification accuracy with entropy values greater than

.90. Results of Lau's (2009) simulation study showed that classification accuracy was greater when a smaller proportion (.9%) of amotivated examinees were present in the dataset than a large proportion (20%), and Lau's (2009) and Swanson's (2013) applied studies found that classification of amotivated examinees was relatively more difficult than classification of motivated examinees. Swanson and Pastor (2014) did not evaluate classification accuracy.

In comparing fit of the 2PL versus RRM 2PL models, Lau (2009), Swanson (2013) and Swanson and Pastor (2014) found each model comparison index to show improvement with the addition of a second class, supporting the inclusion of a second class. Further, results of the LMR likelihood ratio test indicated the need of at least a two-class model to describe the data in Lau's (2009) and Swanson's (2013) studies, but not in Swanson and Pastor's (2014) study. In evaluating changes in parameter estimates, Lau (2009) and Swanson (2013) found the addition of a second class resulted in decreased factor loading estimates (by approximately .05 and .067 logits, respectively), with the change in loading values being greater for easier items than harder items. It was also found that thresholds decreased (by approximately .01 and .019 logits respectively), and while most items appeared easier with the RRM than with the one-class 2PL model, items at the extreme ends of the threshold scale appeared to be more difficult. Changes in parameter estimates were not evaluated in the Swanson and Pastor (2014) study.

Non-Cognitive Models

As described in Chapter 1, evidence suggests that random responding occurs in non-cognitive assessments. To date, the RRM has yet to be applied to non-cognitive data, however, models similar to the RRM, such as the hybrid Rasch-Latent Class (LC)

model (von Davier & Rost, 1995), have. Even though HYRBID models have been fit to polytomous data, the latent classes utilized in these models were not constrained to capture random responders. In other words, mixture measurement models including those having an IRT class and latent class have been utilized with polytomous data, but not for the purposes of identifying random responders.

Need for Study

The purpose of this study therefore represents an attempt to distinguish two classes of examinees – random responders and valid responders – on non-cognitive assessments in low-stakes testing. The majority of existing literature regarding the detection of random responders in low-stakes settings exists in regard to cognitive tests that are dichotomously scored. However, evidence suggests that random responding occurs in non-cognitive assessments, and as with cognitive measures, the data derived from such measures are used to inform practice. Thus, a threat to test score validity exists if examinees' response selections do not accurately reflect their underlying level on the construct being assessed. As with cognitive tests, using data from measures in which students did not give their best effort could have negative implications for future decisions. Thus, there is a need for a method of detecting random responders on non-cognitive assessments that are polytomously scored.

To facilitate the introduction of the RRM appropriate for polytomous responses, the equation for the RRM based on the 2PL is provided below and hereafter referred to as the RRM-2PL.

$$P(u_i = 1) = \pi_1 \frac{\exp(-\tau_{1i} + \lambda_{i1}\theta_1)}{1 + (\exp(-\tau_{1i} + \lambda_{i1}\theta_1))} + \pi_2 \frac{\exp(-g_i)}{1 + (\exp(-g_i))} \quad (5)$$

Recall that the thresholds for the random responders (τ_{2i}) are not freely estimated, but fixed equal to $g_i = -\ln \left[\frac{\frac{1}{r_i}}{1 - (\frac{1}{r_i})} \right]$, with r_i representing the number of response options for the item. To extend the RRM to polytomous responses, an IRT model appropriate for polytomous responses is needed for the class of valid responders. Although there are a variety of models that could be used for this purpose, the Graded Response Model (GRM; Samejima, 1969) was chosen due to its ability to accommodate scales where the number of response options differ across items. Utilizing the GRM for the class of valid responders, a version of the RRM appropriate for polytomous responses is shown in Equation 6 and is hereafter referred to as the RRM-GRM:

$$P(u_i \geq x) = \pi_1 \frac{\exp(-\tau_{1ik} + \lambda_i \theta_j)}{1 + \exp(-\tau_{1ik} + \lambda_i \theta_j)} + \pi_2 \frac{\exp(-g_{ik})}{1 + \exp(-g_{ik})} \quad (6)$$

Whereas the RRM-2PL was used to ascertain the probability of a correct response on an item, the RRM-GRM is used to ascertain the probability of a response in category x or higher. Another difference is the presence of multiple thresholds per item in the RRM-GRM; in fact, there are $k=1$ to m thresholds per item, with $m+1=M$ being the number of response categories. As in the RRM-2PL, the thresholds in the RRM-GRM for the random responding class are fixed, not freely estimated. For example, for an item with five categories ($M = 5$), the m thresholds are set equal to -1.386 for category 2 or higher, -0.405 for category 3 or higher, 0.405 for category 4 or higher, and 1.386 for category 5. The thresholds are a function of $1/M$, which is the proportion of respondents expected to respond to each category if responses were selected randomly. How to arrive at the specific values at which the guessing thresholds are fixed will be described in Chapter 3.

To explore the functioning and utility of the RRM-GRM, the following research questions will be pursued:

Research Question 1 (RQ1): How are item parameter and theta estimates of the GRM impacted by the presence of random responders in the data set?

Research Question 2 (RQ2): Which model (the RRM-GRM or GRM) best fits the data?

Research Question 3 (RQ3): If the RRM-GRM is fit to the data, does it accurately estimate the proportion of random responders?

Research Question 4 (RQ4): Are parameter and theta estimates purified when the RRM-GRM is fit to the data?

Research Question 5 (RQ5): When the RRM-GRM is fit to real data, does evidence suggest that respondents in the random responding class are amotivated?

Two studies were conducted in an attempt to answer these research questions.

The purpose of the first study, Study 1, was to explore the utility of the RRM-GRM for detecting and accounting for random responders on non-cognitive instruments in low-stakes testing settings. Data were simulated such that different proportions of random responder simulees (1%, 5%, 10%, 20%) were incorporated into a large data set containing valid responder simulees following the GRM. Study 1 was divided into two phases. In the first phase, the GRM was fit to the simulated data to answer RQ1, and in the second phase, the RRM-GRM was fit to the same simulated data to answer RQ2, RQ3, and RQ4.

The purpose of the second study, Study 2, was to corroborate the results of using the RRM on real test data with those of the simulated data. Moreover, the results from

this phase can be used as evidence of the utility and appropriateness of the RRM-GRM. In Study 2, the RRM-GRM was fit to non-cognitive data collected in a low-stakes setting to demonstrate its application to authentic data. The GRM was also fit to the same data set, enabling results from the one-class and two-class models to be compared. In addition to answering research questions similar to RQs 1-3, Study 2 also focused on RQ5 by evaluating differences between the two classes on test-taking effort and importance (SOS; Sundre & Moore, 2002), total score on the test, and total time spent completing the measure.

III. Methods

Study 1

The purpose of Study 1 was to explore the utility of the RRM-GRM for detecting and accounting for random responders on non-cognitive instruments in low-stakes testing settings. In Study 1, data was simulated such that different proportions of random responder simulees were incorporated into a large data set of valid responder simulees following the GRM with the resulting data used in two phases of Study 1. In phase 1, the GRM was fit to the simulated data to answer *RQ1: How are item parameter and theta estimates of the GRM impacted by the presence of random responders in the data set?* In phase 2, the RRM-GRM was fit to the same simulated data to answer *RQ2: Which model (the RRM-GRM or GRM) best fits the data?, RQ3: If the RRM-GRM model is fit to the data, does it accurately estimate the percent of random responders?, and RQ4: Are parameter and theta estimates purified when we fit the RRM-GRM to the data?*

Data Generation. Separate samples of valid and random responders were generated according to their corresponding models and concatenated to simulate data sets containing a mixture of respondents. Data sets were created to consist of various percentages of random responders: 1%, 5%, 10%, and 20%, with this condition hereafter being referred to as %RR. The proportions of random responders were selected based on previous research; Wise and DeMars (2006) suggested that roughly 6% of examinees in low-stakes conditions may be unmotivated, whereas Mislevy and Verhelst (1990) and Lau (2009) found the proportion of amotivated examinees to be approximately 4.5% and 1.2% respectively, after applying the RRM to real data. A testing situation where 10% of examinees are unmotivated may be possible in some extreme situations, whereas an

amotivated examinee proportion of 20% does not really seem plausible. However, including this extreme condition should aid in identifying the point wherein the GRM really breaks down. That is, including a proportion of 20% of random responders in a data set will help to demonstrate how “off” parameter estimates might be when one-fifth of respondents do not try on the test.

All datasets included a total of 5,000 simulees (see Table 1) for each of the four levels of the %RR condition. For each of the four levels of random responders, 100 datasets were simulated, resulting in a total of 400 data sets used in both phases of Study 1.

Table 1
Simulee breakdown per %RR condition

% RR	# Valid Responders	# Random Responders	Total # of Simulees
1%	4,950	50	5,000
5%	4,750	250	5,000
10%	4,500	500	5,000
20%	4,000	1,000	5,000

Valid Responders. Data for valid responders were generated according to the Graded Response Model (GRM). The GRM is an extension of the 2PL model that is appropriate for polytomous items and commonly used with Likert scale data. Responses for 20 items were generated to simulate valid responders data on a non-cognitive, unidimensional assessment using a 5-point Likert scale. Population parameters for generating data representative of valid responder simulees, shown in Table 2, were obtained from Lautenschlager, Meade, and Kim (2006), who used data from an administration of the short form of the Minnesota Satisfaction Questionnaire (MSQ) to generate population GRM item parameters for their own simulations (Lautenschlager et

al., 2006). The data they acquired and used to populate the parameters were gathered from 891 manufacturing employees. The short form of the MSQ contains 20 items and uses a five-point Likert scale.

Table 2

Population Parameters for Generating Valid Responders

Item	λ	τ_1	τ_2	τ_3	τ_4
1	0.95	-4.05	-2.76	-1.19	1.91
2	1.48	-3.63	-2.13	-0.89	2.15
3	1.46	-3.02	-1.85	0.23	3.08
4	1.49	-2.61	-1.13	0.19	3.01
5	1.38	-3.02	-1.75	-0.48	2.10
6	1.35	-3.89	-2.66	-0.69	2.52
7	0.96	-3.62	-2.14	-1.22	1.29
8	1.32	-4.28	-3.02	-0.65	2.55
9	1.08	-3.54	-2.26	0.53	3.34
10	2.00	-3.14	-1.50	-0.26	3.36
11	1.22	-1.70	0.10	1.31	3.65
12	0.89	-2.64	-1.34	-0.36	2.17
13	2.05	-4.20	-2.44	-0.31	3.83
14	1.59	-1.91	-0.38	0.97	3.94
15	2.31	-3.88	-2.19	-0.58	3.90
16	2.07	-3.93	-2.24	-0.81	3.29
17	1.55	-2.79	-1.24	0.16	3.04
18	0.92	-3.51	-2.42	-1.10	1.54
19	1.64	-2.30	-0.82	0.57	3.44
20	2.35	-4.00	-2.12	-0.14	4.25

Note. Population parameters have been converted from IRT model parameterization to the factor model parameterization.

Source: Lautenschlager, Meade, & S. H. Kim (2006, p. 7).

Using the factor model parameterization, an equation representing the GRM is shown in Equation 7. The equation represents the marginal probability of an examinee scoring x or higher on item i , given theta ($P(u_i \geq x)$).

$$P(u_i \geq x) = \frac{\exp(-\tau_{ik} + \lambda_i \theta_j)}{1 + \exp(-\tau_{ik} + \lambda_i \theta_j)} \quad (7)$$

Specifically, in Equation 7, lambda (λ) represents loadings and tau (τ), thresholds⁵. A respondent's estimated ability is represented by θ and there are $k=1$ to m thresholds, with $m+1=M$ being the number of categories for an item. For this study, a five-point Likert scale was used; thus, there were four threshold parameters⁶.

To calculate the probability of selecting a particular option, Equation 7 cannot be directly used. Instead, the probability of selecting options 1 through 5 can be calculated with Equation 8 through Equation 12, respectively.

$$P(u_i = 1) = 1 - P(u_i \geq 2) \quad (8)$$

$$P(u_i = 2) = P(u_i \geq 2) - P(u_i \geq 3) \quad (9)$$

$$P(u_i = 3) = P(u_i \geq 3) - P(u_i \geq 4) \quad (10)$$

$$P(u_i = 4) = P(u_i \geq 4) - P(u_i \geq 5) \quad (11)$$

$$P(u_i = 5) = P(u_i \geq 5) - 0 \quad (12)$$

To generate item responses for valid responders according to the GRM, theta values were generated for each simulee by extracting a random number from a standard normal distribution with a mean of zero and standard deviation of one. The theta values were then used along with the true population parameters shown in Table 2 and the GRM model in Equation 7 to determine the probability of an examinee scoring at a particular category (x) or higher, given their simulated theta. As an example, consider the set of

⁵ The correspondence between factor model parameters and IRT parameters in Equation 7 is loadings (λ) = a and thresholds (τ) = (ab) .

⁶ In discussing the GRM in terms of the factor model parameterization, the term "thresholds" is used to describing τ , whereas the term "difficulties" is used when describing b as part of the IRT parameterization.

cumulative probability values for item 1 for a simulee with a theta of 0: $\geq 1 = 0.983$, $\geq 2 = 0.940$, $\geq 3 = 0.767$, $\geq 4 = 0.129$.

Item responses were generated using the SAS macro IRTGEN⁷ (Whittaker, Fitzpatrick, Williams, & Dodd, 2003). In this program, responses are generated by comparing each cumulative probability value to a random number generated from a uniform distribution in order to add a degree of realism to the data by incorporating random error. If the probability of a correct response for a category was at or higher than the number generated from the uniform distribution, the simulee was assigned that category score for the item. For instance, if the random number drawn were 0.45, the response for the example simulee would be 4 since this random number falls in between the cumulative probabilities associated with response options 3 and 4. This process was repeated for every simulee and item in the study.

Random Responders. Various proportions of the simulees (1%, 5%, 10%, and 20%) were generated to emulate random responders. Random responders, also referred to as amotivated respondents, are characterized by their tendency to respond to items carelessly or arbitrarily starting from the first item on the test. Population data for random responders were generated by selecting a random value from a multinomial distribution having an equal probability of discrete values between 1 and 5. The SAS syntax used to create the data sets of simulees for all conditions is located in Appendix A.

Simulation Study Design

Phase 1. In phase 1, a simulation was conducted to explore the impact of random responders on item parameter estimates and theta distributions when an IRT model is fit

⁷ Because IRTGEN utilizes the IRT parameterization of the GRM, the parameters in Table 2 were converted to the IRT parameterization prior to their input into the program.

to the data and the presence of random responders is ignored. Essentially, the GRM was fit to each of the simulated data sets, ignoring the fact that random responders were present. The percent of random responders in the data set was varied in an attempt to determine how the item parameters and theta estimates were impacted by the presence of varying amounts of random responders. To answer RQ1, the true item parameters and true theta values were compared to the estimated values to assess the impact of the presence of random responders.

To compare true and estimated parameters, bias and root mean squared error (RMSE) were evaluated. If an estimate is biased, then it is either consistently above or below the true value on average. To calculate bias for each individual parameter, the true population value (ξ) is subtracted from the average estimate value ($\hat{\xi}$) across replications, where r represents the number of replications. Equation 13 presents this computation. To calculate percent bias, bias is simply divided by the true population value (ξ), as shown in Equation 14. It was expected that the magnitude of bias would increase as the proportion of random responders increased. Bias in parameter estimates for different values of loadings and category thresholds was evaluated (e.g., does the direction and magnitude of bias in loadings depend on the true value of the loading?).

$$bias = \frac{\sum_{l=1}^r (\hat{\xi}_l - \xi)}{r} \quad (13)$$

$$\%bias = \frac{bias}{\xi} \quad (14)$$

RMSE is another way to evaluate if item parameters differ from true parameters. Not only does RMSE capture bias, but it also takes into account the amount of variability in the estimate, or how imprecise it is. Since there is a trade-off between bias and variability, RMSE expresses the degree to which they are balanced in parameter

estimates. To calculate RMSE, the empirical standard error (SE), or the standard deviation of the estimate across replications, is squared and added to the squared deviance of the mean parameter estimate from the true parameter value. This value is considered to be the mean squared error (MSE). To get the RMSE, the square root of the MSE is taken, effectively putting it on the same metric as the parameter. For the RMSE index, good estimation is signified when values are closer to zero. The computational formula is presented in Equation 15.

$$RMSE = \sqrt{(\hat{\xi} - \xi)^2 + SE^2} \quad (15)$$

It was hypothesized that the effect of random responders on item parameters and theta distributions would depend on the value of the true parameters and thetas. That is, the effect could depend on whether discriminations and category thresholds are high or low for each item. For this reason, the bias and RMSE were examined conditional on the true values of item parameters and thetas.

Phase 2. In Phase 2, the RRM-GRM was fit to the same 400 data sets as the GRM in Phase 1. The RRM-GRM is shown in Chapter 2 as Equation 6 and again here as Equation 16.

$$P(u_i \geq x) = \pi_1 \left[\frac{\exp(-\tau_{ik} + \lambda_i \theta_j)}{1 + \exp(-\tau_{ik} + \lambda_i \theta_j)} \right] + \pi_2 \left[\frac{\exp(-g_{ik})}{1 + \exp(-g_{ik})} \right] \quad (16)$$

The equation represents the marginal probability of an examinee scoring x or higher on item i , given theta ($P(u_i \geq x)$). More specifically, the marginal probability is expressed as the weighted sum of two terms. The first term in Equation 16 represents a single factor measurement model, which is used for the valid responders. Shown here, the first term is the GRM. With the GRM, the probability of an examinee scoring x or

higher on item i is a function of the ability of the examinee (θ) and the particular item's loadings (λ) and thresholds (τ). The second term in Equation 16 represents the probability of an examinee scoring x or higher on item i as equal to that of chance, which is the model used to represent the random responders. Essentially, this model is the GRM with the variance of theta set to zero, loadings set to zero, and the category thresholds fixed to a guessing threshold (g_{ik}), which is equal to $g_{ik} = \ln \left[\frac{1 - \frac{M-k}{M}}{\frac{M-k}{M}} \right]$.

Recall that there are $k=1$ to m thresholds per item, with $m+1=M$ being the number of response categories. Since the items in Study 1 have five categories ($M = 5$), the m thresholds are set equal to -1.386 for category 2 or higher, -0.405 for category 3 or higher, 0.405 for category 4 or higher, and 1.386 for category 5.

The weight of the class in Equation 16 represents the proportion of examinees contained in the class in the population. For valid responders, the weight of the class is represented by π_1 and for the random responders, the weight of the class is represented by π_2 , which is a function of π_1 ($\pi_2=1-\pi_1$) since weights are constrained to sum to one across classes.

To answer RQ2, model-data fit indices for the RRM-GRM and GRM were compared to assess which model best fit the simulated data. The fit of the measurement models were compared using log-likelihood based relative fit indices. The log-likelihood based relative fit indices that were examined included Akaike's Information Criteria (AIC; Akaike, 1973), Bayesian Information Criteria (BIC; Schwarz, 1978), and the sample-size adjusted BIC (SSABIC; Sclove, 1987), which were all obtained from Mplus (Muthén & Muthén, 1998-2006). The BIC and SSABIC indices both take into account the number of parameters, thereby penalizing models with greater numbers. The

SSABIC also accounts for sample size, which confounds the BIC and AIC indices. For AIC, BIC and SSABIC values, those closer to zero were indicative of better model fit, thus lower values were more desirable. For this study, all three indices were examined, but the SSABIC index was weighted more heavily since it also accounts for sample size and has been found to perform relatively better than the other indices in simulation studies (Henson, Reise, & Kim, 2007; Tofighi & Enders, 2008; Yang, 2006).

To answer RQ3, the average class weights across replications were compared to the true value to determine if the RRM-GRM model accurately estimated the class proportions in the dataset. The average entropy value was also used to evaluate classification accuracy. The class weight for each data set is the only additional parameter estimated by the RRM-GRM that is not estimated by the GRM. These values were obtained through Mplus (Muthén & Muthén, 1998-2006) and compared to the true proportion of random responders included in each dataset (1%, 5%, 10%, and 20%). The average entropy statistic was also obtained through Mplus and compared across conditions to evaluate if classification accuracy was greater with particular proportions of random responders than with others.

To answer RQ4, the same methods (e.g., bias, RMSE) used to answer RQ1 were used to determine whether the item parameter and theta estimates were purified, or in other words, closer to their true values, when the RRM was fit to the data.

Software. The software used for estimation in both phases was Mplus, version 7.11 (Muthén & Muthén, 1998-2006). The estimation method used in Mplus was the default maximum likelihood technique (ML) for categorical items. The datasets generated were analyzed two times, once with the GRM and once with the RRM-GRM.

To set the scale of the latent variable (for the GRM or for the valid responder class in the RRM-GRM), the mean and variance of the factor (theta) were set to zero and one, respectively. When the RRM-GRM was fit to the data, the loadings and thresholds were allowed to freely estimate for the valid responder class. For the random responder class, the variance of theta was set to zero, loadings were set to zero, and the category thresholds were fixed to be a function of the cumulative probability of selecting a particular category if a respondent were randomly responding, as described above. Data including resulting item parameters and global fit indices were imported into SAS, version 9.4 for further analyses. Appendix B contains the SAS syntax used to generate the Mplus syntax for the GRM and RRM-GRM. Appendix C contains the SAS syntax used to read in the datasets from Mplus and to complete computations.

Local maxima. When estimating item parameters, the goal is to identify the most likely solution by estimating the highest peak, or the global maximum, of the likelihood function. However, the likelihood function for mixture models is bumpy, with a multitude of peaks. Thus, the estimation process may have a difficult time detecting the highest peak, as it is possible to converge on a local maximum instead. If convergence on a local instead of a global maximum occurred, the results would not reflect the most likely parameterization of the data. Therefore, precautions must be taken to prevent convergence on a local maximum.

To assist Mplus in converging on the global maximum, a feature available in Mplus was used to generate random sets of starting values for the parameters. For each model, 200 sets of randomly generated starting values were used to estimate the model with a limited number of iterations. The best fitting 50 were retained and allowed to run

until convergence was obtained. The best fitting set of estimates of these 50 (assumed to be the solution associated with the global maximum) was used as the model's final set of estimates.

Study 2

The purpose of Study 2 was to corroborate the results of using the RRM-GRM on real data with those of the simulated data to provide evidence of the utility and appropriateness of the RRM-GRM for use with non-cognitive data collected in a low-stakes setting. In addition to the RRM-GRM, the GRM was also fit to the same data set, enabling results from the one-class and two-class models to be compared. It was expected that the differences between the models would resemble those observed with the simulated data. Study 2 was similar to Study 1 in that it answered RQs 1, 2, 3, and 4 using the same methods. Because the true parameter values are not known in Study 2, only the change in parameter estimates when the GRM versus the RRM-GRM were fit to the data were examined (as opposed to examining how parameter estimates compared to their true values). Study 2 also focused on answering *RQ5*: “*When the RRM-GRM is fit to real data, does evidence suggest that respondents in the random responding class are amotivated?*” by evaluating differences between classes detected by the RRM-GRM on test-taking effort and importance as measured by the Student Opinion Scale (SOS; Sundre & Moore, 2002), gender, and total score on the scale. In other words, external validity evidence for the class solution was obtained for RQ5. Further information about the dataset used in Study 2 along with external variables is provided below.

Low-stakes Assessment Dataset. Archival data collected in a low-stakes testing context were used in this study. The data were collected from 3,585 undergraduate

students with credits ranging from 45 to 70 (sophomores or juniors) in February of 2014 at James Madison University (JMU), a mid-sized, public, southeastern university. The examinees were required by JMU to participate in a three-hour, campus-wide testing series designed to assess general education and student affairs programs. If students missed the initial administration, they were still required to complete the assessments by either attending one of two make-up sessions or as a “walk-in” at JMU’s Assessment and Testing Center. The testing series was comprised of cognitive and non-cognitive tests, and was concluded with the administration of the SOS for all examinees. The results of the testing series held no consequences for individual examinees, as scores were used in the aggregate; thus, the testing context was low-stakes and it was assumed that random responders were present.

Measures.

Unified Measure of University Mattering (UMUM-15). The UMUM-15 (France, 2011) is an abbreviated version of the Revised University Mattering Scale (RUMS; France, 2011). The RUMS, a non-cognitive instrument with 34 items, was reduced to the 15 item UMUM-15 based on France’s (2011) model-data fit findings from a confirmatory factor analysis study. The UMUM-15 is a unidimensional instrument that seeks to measure university mattering, or the feeling of an individual that they are significant to and make a difference in their university (France, 2011). The items have six response options that range from Strongly Disagree (1) to Strongly Agree (6). The scale was administered along with three other scales as part of the Attitudes Toward Learning, Version 13 (ATL-13) instrument on Assessment Day. The UMUM-15 was

placed near the end of the measure (specifically, it was items 63-77 on the 93 item ATL-13). The placement of the ATL-13 in the succession of tests was variable.

Student Opinion Scale (SOS). The SOS (Sundre & Moore, 2002) is a self-report measure of test-taking motivation that is administered to examinees after completing a test, or in this case, a battery of tests. The SOS consists of 10 items that ask students to respond to statements about how much effort they exerted and their perceived importance of the test using a five-point Likert scale. Response options on the Likert scale range from one (strongly disagree) to five (strongly agree). SOS responses were summed to create total scores with a range of five to 25 points. Total scores on the lower end indicated low effort/perceived importance, whereas scores on the upper end indicated high effort/perceived importance. There is empirical support for a two-factor structure consisting of an “importance” and an “effort” factors (Thelk et al., 2009). Each factor contains five items and separate scores were reported for each subscale.

External Validity Analyses. Since the RRM-GRM is used to detect unknown groups, validity evidence for the composition of the classes must be acquired. To establish validity evidence, classes can be compared to variables (often called “auxiliary” variables) that previous research or theories have proposed to be related to evaluate if they are correlated as hypothesized. A straightforward approach to such an analysis is to classify respondents into classes using modal assignment (i.e., assign respondent to the class for which their posterior probability is the highest) and then relate this grouping variable to auxiliary variables using traditional statistical analyses (e.g., t-test, regression). A limitation of this approach is that it does not take into account the measurement error associated with the grouping variable. For instance, unless

classification accuracy is perfect (e.g., entropy is 1.0), the grouping variable based on modal assignment will be an imperfect representation of the latent categorical variable.

There are a variety of different analytical options available in Mplus to take the measurement error of the grouping variable into account when estimating its relationship with auxiliary variables (Asparouhov & Muthén, 2014a; Asparouhov & Muthén, 2014b). Unfortunately, many of the options associated with the best performance in simulation studies (e.g., the BCH method, Lanza's methods) cannot be used with this model in Mplus⁸. The only option available is the use of the manual-3-step procedure⁹ in Mplus proposed by Vermunt (2010). In this approach, the RRM-GRM is first fit to the data and information pertaining to the classification accuracy of the model is retained. In a second model, a grouping variable is still created using modal assignment, but its relationship with the latent categorical variable in this model is fixed to values that represent the classification accuracy of the RRM-GRM. Parameters from this second model that capture the relationships of auxiliary variables with the latent categorical variable are used to ascertain the validity of the latent categorical variable in the RRM-GRM. Effort, importance and gender were specified as predictors of the latent categorical variable and total score on the UMUM was specified as an outcome¹⁰.

⁸ Mplus has not yet made these options for auxiliary analyses available when numerical integration is used during estimation.

⁹ The 3-step procedure of Vermunt (2010) can be implemented in Mplus automatically, but not for models that use numerical integration during estimation. For this reason, the 3-step procedure had to be implemented manually.

¹⁰ Because class-switching can occur in the 3-step approach when auxiliary variables are specified as outcomes, the validity analyses for the outcome variables were monitored for class-switching. Specifically, the proportions of respondents in each class using modal assignment in the RRM-GRM were compared to the same proportions obtained in the validity model. If more than 20% of respondents change classes across the two models, the results of the validity model were considered inconsistent and not trustworthy (Asparouhov & Muthén, 2014a).

The results would provide support for the interpretation of a random responder class if the average UMUM-15 score was equal to random responding, which is 52.5 here. That is, because the UMUM-15 has a 6-point scale and there is a 0.17 chance of responding in each of the 6 categories, 0.17 can be multiplied by each response option ($0.17*1 + 0.17*2 + 0.17*3 + 0.17*4 + 0.17*5 + 0.17*6 = 3.5$) to get a total of 3.5 for each item. Since there are 15 items, 3.5 would then be multiplied by 15 to get a total score of 52.5. Additionally, validity evidence supporting the RRM-GRM would be acquired if the average number of males was found to be greater in the random responding class than in the valid responding class.

IV. Results

Study 1

RQ1: How are item parameter and theta estimates of the GRM impacted by the presence of random responders in the data set? Descriptive statistics for item parameter estimates of the GRM are presented in Tables 3 and 4 and for theta estimates in Table 5. An overview of the results is provided here, with more specific information in the paragraphs that follow. For theta estimates, bias increased along with the proportion of random responders in the dataset. For item parameter estimates, bias, percent bias, and RMSE values also increased along with the proportion of random responders. In other words, larger proportions of random responders were found to be associated with weaker estimation accuracy, including higher bias and RMSE.

Factor loadings. Factor loadings (see Table 3) in the 1%, 5%, 10%, and 20% random responder conditions, were underestimated on average by 0.02, 0.11, 0.21, and 0.37 units respectively. In other words, on average, factor loadings were estimated at a lower value than the true loadings, and as the proportion of random responders in the dataset increased, so did the amount of bias present. For example, for the 20% random responder condition, bias was -0.373, which is 23.7% of the parameter value. Thus, the presence of a large proportion of random responders makes accurate loading parameter estimation problematic, even for a low-stakes setting. Additionally, the RMSE value for each of the conditions is very similar to each condition's value for bias. For example, the average amount of bias for the 20% condition is -0.373 and average amount of RMSE is 0.375. This indicates that the amount that the estimates depart from their true value is a function of bias, not of sampling error.

Table 3
Average Performance Indices for the GRM – Loadings

Criterion	π	M	SD	Min	Max	
Loadings						
Bias	0.01	-0.023	0.013	-0.054	-0.008	
	0.05	-0.110	0.059	-0.239	-0.035	
	0.10	-0.207	0.107	-0.436	-0.068	
	0.20	-0.373	0.178	-0.740	-0.138	
Proportion	0.01	-0.015	0.004	-0.023	-0.006	
	Bias	0.05	-0.069	0.019	-0.102	-0.039
		0.10	-0.130	0.033	-0.186	-0.071
RMSE	0.20	-0.237	0.051	-0.315	-0.144	
	0.01	0.048	0.014	0.032	0.078	
	0.05	0.117	0.057	0.047	0.244	
	0.10	0.211	0.106	0.074	0.439	
	0.20	0.375	0.177	0.142	0.742	

Note. π is the true proportion of random responder simulees.

To determine if bias and RMSE are related to the true values of the factor loadings as opposed to the average value across all 20 loadings, the population values (i.e. true values) were plotted against the bias and RMSE values and are displayed in Figures 1 and 2, respectively. That is, does the amount of bias and RMSE present in each condition depend on the value of the loadings? Figures 1 and 2 demonstrate that the amount of bias and RMSE present in the conditions does depend on the value of the loadings. For instance, when examining bias in Figure 1, it can be seen that the higher the value of the factor loading, the worse the negative bias in the estimated loadings in the presence of random responders. For conditions containing higher proportions of random responders, more negative bias is present for higher loading values than in conditions with lower proportions of random responders. Figure 2 demonstrates a similar interaction with RMSE and loading values, which is expected as the amount that the

estimates depart from their true value appears to be a function of bias, and not of sampling error. The data used to construct the plots are located in Appendix D.

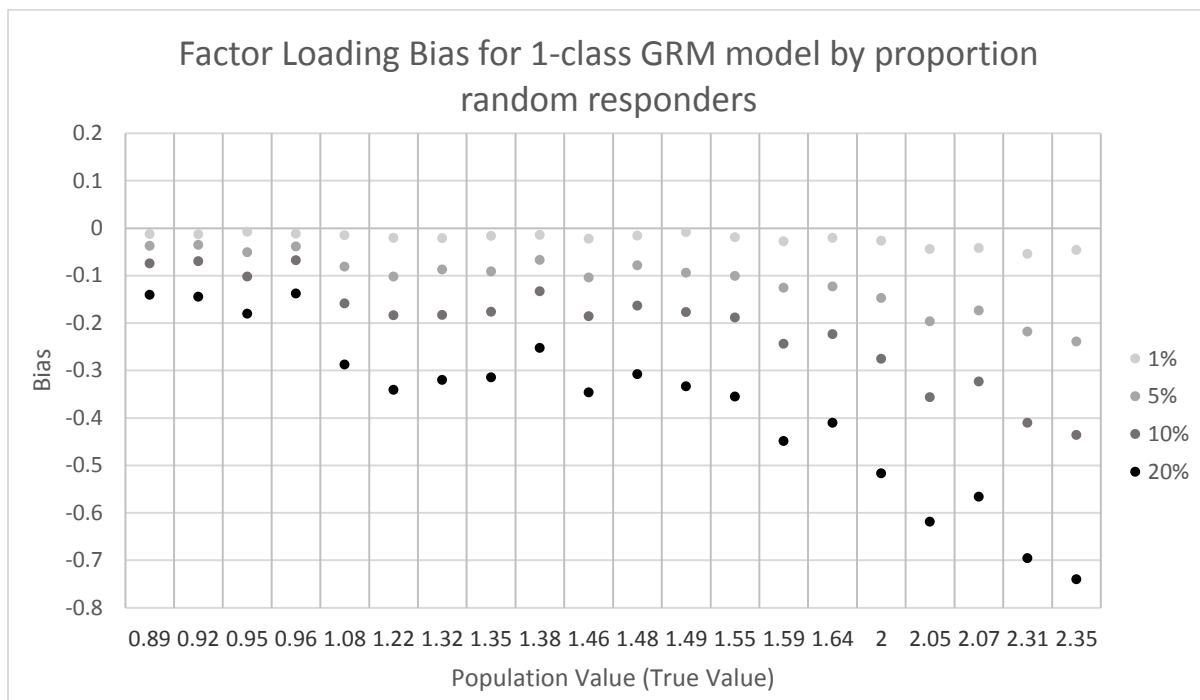


Figure 1. Factor Loading Bias for the GRM.

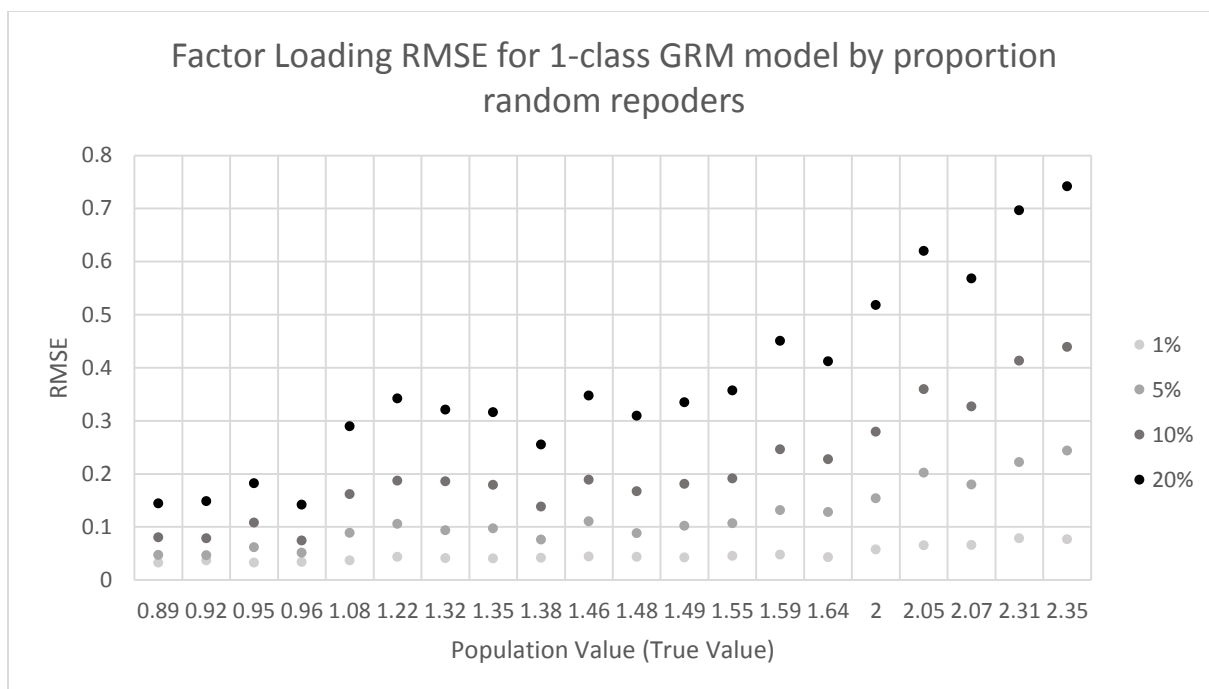


Figure 2. Factor Loading RMSE for the GRM.

Thresholds. The indices associated with thresholds are provided in Table 4. For the 1%, 5%, 10%, and 20% conditions, thresholds were overestimated on average by 0.02, 0.07, 0.12, and 0.21 units, respectively. In other words, on average, thresholds were estimated at a higher value than the true thresholds, and as the proportion of random responders in the dataset increased, so did the amount of bias present. For example, for the 20% random responder condition, bias was 0.212, which is approximately 24.7% of the parameter value. Thus, the presence of a large proportion of random responders makes accurate threshold estimation problematic. In contrast to the loadings, the RMSE value for each of the conditions is larger than each condition's value for bias. For example, the average amount of bias for the 20% condition is 0.212 and the average amount of RMSE is more than double at a value of 0.492. This indicates that the amount

that the estimates depart from their true value is a function of both sampling error and bias, on average.

Table 4
Average Performance Indices for the GRM - Thresholds

Criterion	π	M	SD	Min	Max
Bias	0.01	0.015	0.037	-0.073	0.092
	0.05	0.066	0.166	-0.368	0.386
	0.10	0.119	0.310	-0.682	0.680
	0.20	0.212	0.544	-1.141	1.172
Proportion	0.01	-0.015	0.020	-0.091	0.069
	0.05	-0.073	0.084	-0.541	0.154
	0.10	-0.136	0.140	-0.858	0.274
Bias	0.20	-0.247	0.252	-1.537	0.480
	0.01	0.066	0.026	0.029	0.130
	0.05	0.160	0.097	0.036	0.393
RMSE	0.10	0.282	0.182	0.036	0.686
	0.20	0.492	0.313	0.039	1.173

Note. π is the true proportion of random responder simulees.

To determine if bias and RMSE are related to the true values of the thresholds as opposed to the average across all 80 threshold values, the population values (i.e. true values) were plotted against the bias and RMSE values and are displayed in Figures 3 and 4, respectively. That is, does the amount of bias and RMSE present in each condition depend on the value of the thresholds? In Figure 3, it can be seen that thresholds at the extremes are most biased. That is, thresholds that are extremely low or high have weaker estimation accuracy than thresholds that are average. More specifically, negative thresholds are positively biased, or overestimated, whereas positive thresholds are negatively biased, or underestimated. In Figure 4, thresholds at the extremes also contain the most RMSE. That is, the presence of random responders contributes to sampling error and bias more when thresholds are really low or really high.

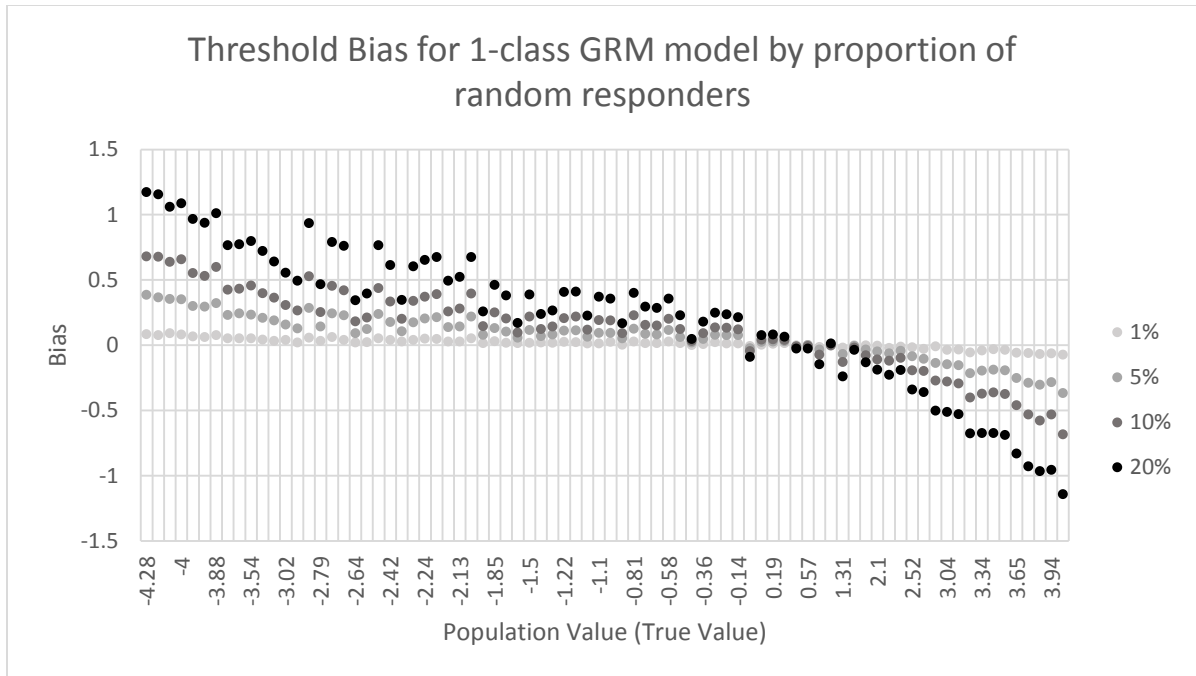


Figure 3. Threshold bias for the GRM.

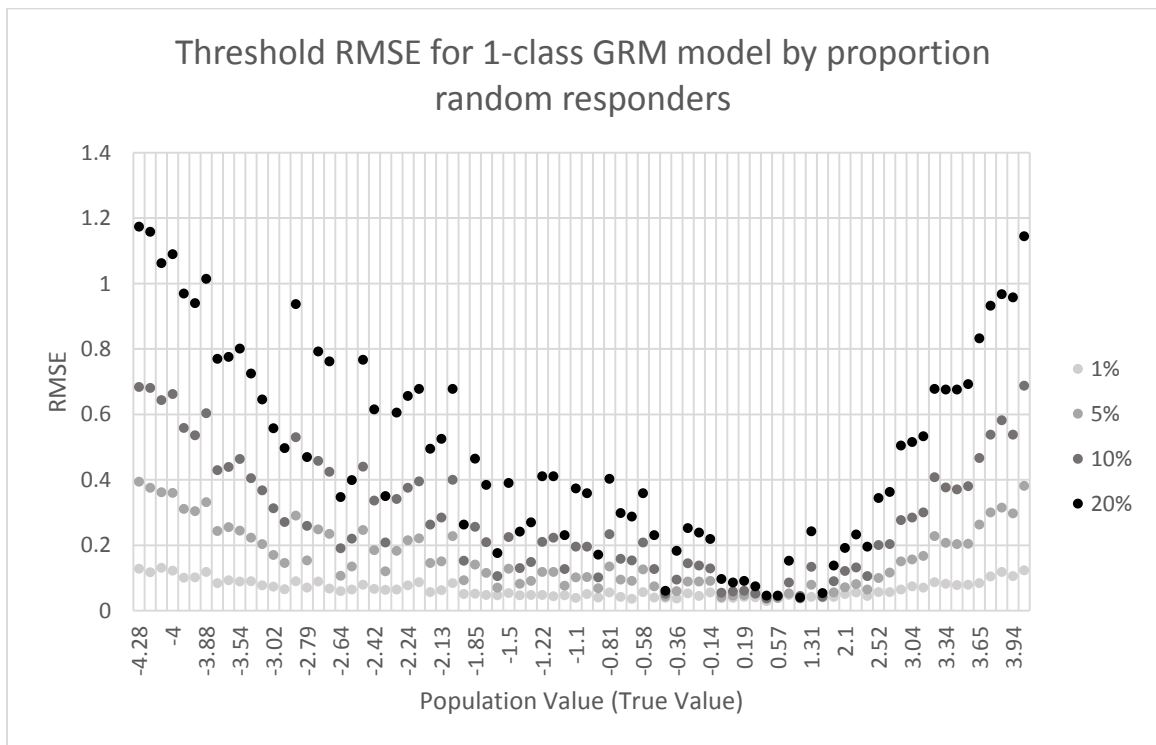


Figure 4. Threshold RMSE for the GRM.

Thetas. Only bias was examined for theta estimates. The average theta estimates by condition for true valid responders and true random responders is shown in Table 5. Recall that true values of theta only exist for simulees in the valid responder class and that the true thetas for such simulees were taken from a standard normal distribution. The average theta estimate for valid responders is therefore a measure of bias, which does appear to be a problem. In Table 5, it can be seen that average theta estimates for valid responders are positively biased in the GRM and that as the proportion of random responders increases, bias becomes more pronounced. More specifically, the average theta estimate for the valid responding class in the 5% condition is 0.004, 0.022 in the 10% condition 0.043 in the 15% condition, and 0.088 in 20% condition, when the true theta average is really zero.

The average theta estimate is also reported for random responders to ascertain what conclusions would be made about their theta levels if the GRM were used. In all conditions, the average theta estimate for random responders is below zero. Thus, use of the GRM when random responders are present in the dataset would lead one to conclude that random responders have lower than average theta levels on the construct being measured.

Table 5
Average Theta Estimates for the GRM

π	True Class	M	SD
0.01	Valid	0.004	0.964
	Random	-0.428	0.499
0.05	Valid	0.022	0.971
	Random	-0.412	0.495
0.10	Valid	0.043	0.981
	Random	-0.389	0.494
0.20	Valid	0.088	1.002
	Random	-0.351	0.503

To determine if the magnitude and direction of the bias depends on the true theta level for valid responders, true theta estimates were categorized by range and the average bias (estimated theta-true theta) computed for all conditions. The pattern of the results is the same for all conditions, so the results of only one of the conditions, the 20% condition, are located in Table 6. Table 6 demonstrates that true thetas at extreme values were more biased than thetas near the average. For example, the mean bias of theta estimates in the ≤ -3.51 and ≥ 3.51 ranges were 0.532 and -0.482 respectively, whereas the mean bias of theta estimates in the 0 to 0.49 range was 0.115. Furthermore, the direction of bias differs depending on whether theta is low or high. In other words, low thetas are positively biased (e.g. theta = ≤ -3.51 , mean = 0.532) and high thetas are negatively biased (e.g. theta = ≥ 3.51 , mean = -0.482).

Table 6

*Average Bias by True Theta Ranges Using
the GRM (20% Condition)*

Range Min	Range Max	<i>M</i>	<i>SD</i>
--	≤ -3.51	0.532	0.301
-3.5	-3.01	0.274	0.272
-3	-2.51	0.151	0.279
-2.5	-2.01	0.080	0.288
-2	-1.51	0.087	0.283
-1.5	-1.01	0.105	0.274
-1	-0.51	0.120	0.270
-0.5	-0.01	0.129	0.271
0	0.49	0.115	0.265
0.5	0.99	0.055	0.258
1	1.49	-0.008	0.274
1.5	1.99	-0.014	0.298
2	2.49	0.007	0.310
2.5	2.99	-0.036	0.319
3	3.49	-0.153	0.291
≥ 3.5	--	-0.482	0.287

RQ2: Which model (the RRM-GRM or GRM) best fits the data? Model fit

indices are displayed in Table 7 for the GRM and RRM-GRM. The fit indices for LL are higher and AIC, BIC, and SSA-BIC are lower for the RRM-GRM. That is, each index is improved with the RRM-GRM, which supports the use of the RRM-GRM over the GRM. Additionally, it should be noted that as the proportion of random responders increases, so does the difference between the model fit indices.

Table 7
Model Fit Indices Summary

Index	π	GRM	RRM-GRM	Difference
LL	0.01	-122,129.72	-121,906.29	-223.42
	0.05	-125,441.81	-124,132.50	-1,309.31
	0.10	-129,284.58	-126,674.17	-2,610.41
	0.20	-136,061.18	-131,438.88	-4,622.30
AIC	0.01	244,459.44	244,014.59	444.85
	0.05	251,083.62	248,466.99	2,616.63
	0.10	258,769.15	253,550.34	5,218.82
	0.20	272,322.36	263,079.77	9,242.59
BIC	0.01	245,111.15	244,672.82	438.33
	0.05	251,735.34	249,125.23	2,610.11
	0.10	259,420.87	254,208.57	5,212.30
	0.20	272,974.08	263,738.00	9,236.07
SSA-BIC	0.01	244,793.39	244,351.88	441.51
	0.05	251,417.57	248,804.29	2,613.29
	0.10	259,103.11	253,887.63	5,215.48
	0.20	272,656.31	263,417.06	9,239.25

RQ3: If the RRM-GRM is fit to the data, does it accurately estimate the proportion of random responders? The proportion of responders in each class (π) is the only additional parameter that is estimated with the RRM-GRM when compared to the GRM. The average estimated proportion for each of the conditions is located in Table 8. According to Table 8, the RRM-GRM estimated the proportion of random responders for the 1%, 5% and 10% conditions to be the true proportion. The estimated proportion of the 20% condition was only off from the true proportion by 0.001.

Classification accuracy for each of the conditions can be evaluated by the entropy statistic located in Table 8. Entropy is higher for the conditions with lower proportions of random responders than conditions with higher proportions, but we still considered to be sufficiently high, as all values are above 0.90.

Table 8
Estimated class proportions and average entropy for the RRM-GRM

π	Estimated π	Estimated $1 - \pi$	Difference	Entropy
0.01	0.010	0.990	0.000	0.991
0.05	0.050	0.950	0.000	0.974
0.10	0.100	0.900	0.000	0.961
0.20	0.199	0.801	0.001	0.943

Note. π is the true proportion of random responder simulees.

RQ4: Are parameter and theta estimates purified when the RRM-GRM is fit to the data? Descriptive statistics for item parameters from the RRM-GRM are presented in Tables 9 and 10 and for theta estimates in Table 13. An overview of the results is provided here, with more specific information in the paragraphs that follow. For theta estimates, essentially no bias was detected in true valid responder's theta values that are assigned to the correct class with the RRM-GRM. Additionally, it was determined that the magnitude and direction of the small amount of bias that existed for valid responders classified as valid depended on true theta level. For item parameter estimates, bias and percent bias values were low for all conditions. However, while RMSE values were also low, they were higher than bias values. In other words, as the proportion of random responders increases, sampling error appears to become more of a factor. Even in this situation, the sampling error values are not large enough to be problematic in practice.

Factor loadings. Factor loadings (see Table 9) in all of the random responder conditions were estimated on average with little to no bias. Average RMSE values for each of the conditions were also low. However, they were higher than estimates of bias,

indicating that the amount that the estimates depart from their true value is mainly a function of sampling error, not bias.

Table 9

Average Performance Indices for the RRM-GRM - Loadings

Criterion	π	M	SD	Min	Max
Bias	0.01	0.000	0.004	-0.009	0.011
	0.05	-0.001	0.004	-0.010	0.009
	0.10	0.000	0.005	-0.013	0.009
	0.20	0.001	0.004	-0.007	0.008
% Bias	0.01	0.000	0.003	-0.006	0.007
	0.05	-0.001	0.003	-0.007	0.007
	0.10	0.000	0.003	-0.006	0.004
	0.20	0.000	0.003	-0.007	0.004
RMSE	0.01	0.042	0.009	0.031	0.063
	0.05	0.041	0.007	0.030	0.057
	0.10	0.043	0.008	0.032	0.062
	0.20	0.047	0.009	0.035	0.066

Note. π is the true proportion of random responder simulees.

As with the GRM in RQ1, the population values (i.e. true values) were plotted against the bias and RMSE values and are displayed in Figures 5 and 6 respectively. In Figure 5, it can be seen that the factor loadings lie almost directly at zero. As previously noted in Table 9, little to no bias was present on average for the factor loadings. Figure 5 demonstrates that this is true for all of the population values. In Figure 6, even though there is very little RMSE present in the factor loadings, it does appear that as the factor loading population values increase, RMSE does as well. Thus, as the proportion of random responders increases, sampling error becomes more of a factor. That is, the presence of random responders contributes to sampling error more when factor loadings are high. Importantly, even for higher true factor loading values, the values of RMSE are not high enough to be problematic.

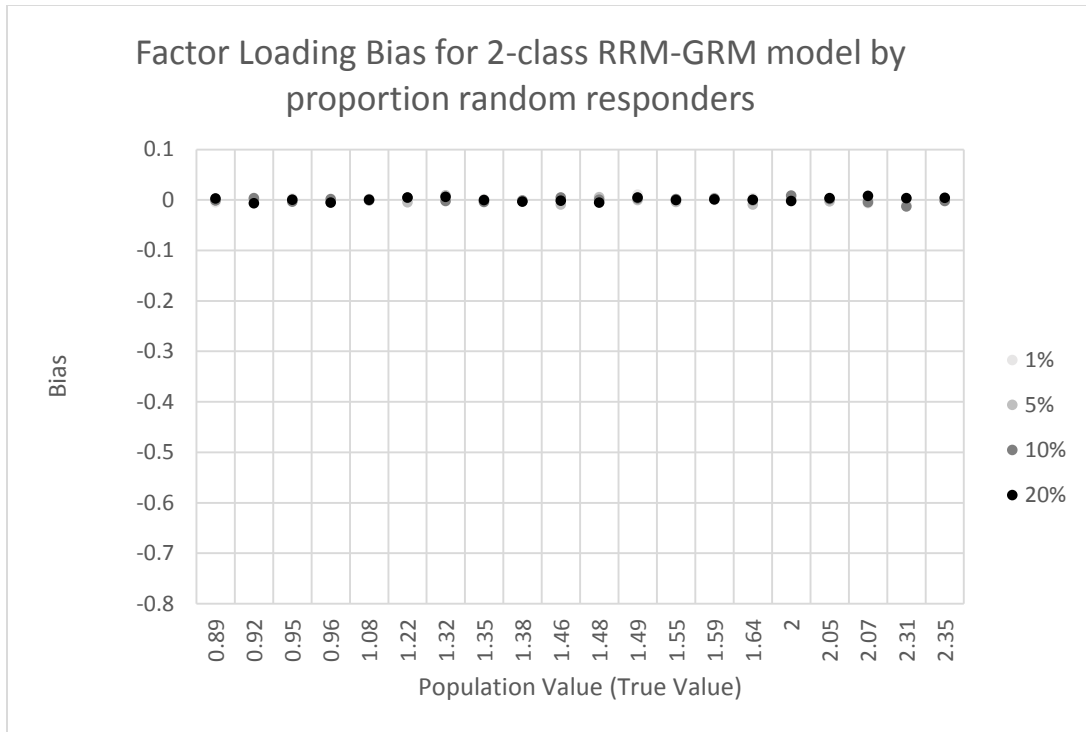


Figure 5. Factor loading bias for the RRM-GRM.

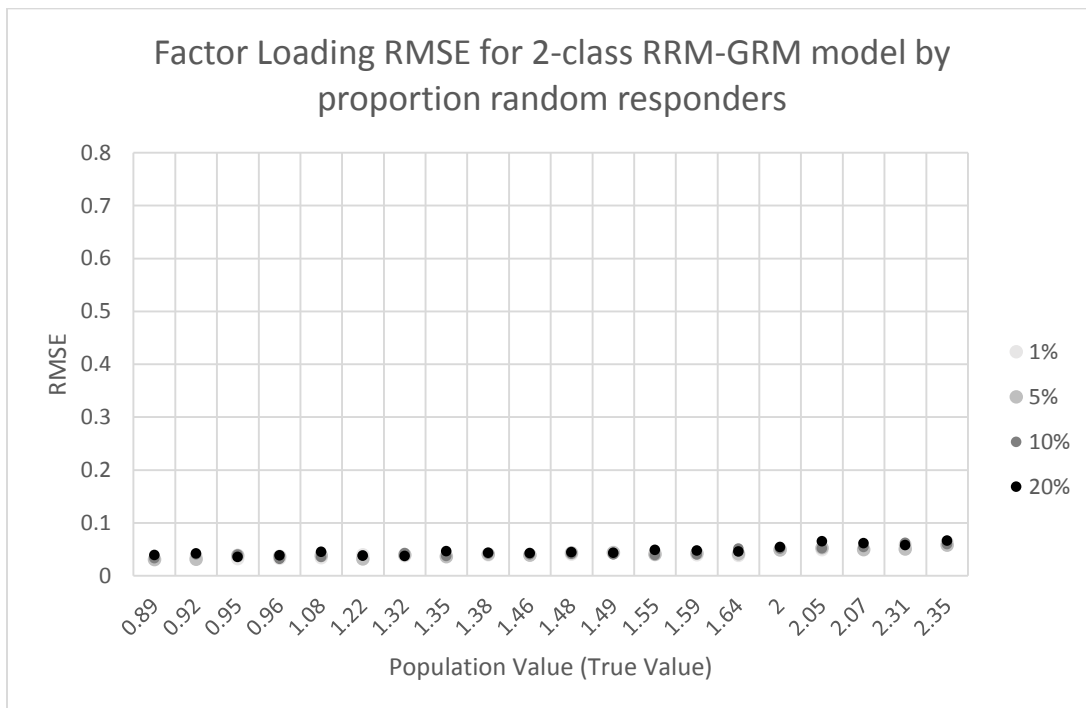


Figure 6. Factor loading RMSE for the RRM-GRM.

Thresholds. Threshold values (see Table 10) looked very similar to those of the loadings. That is, all of the random responder conditions were estimated on average with little to no bias and low RMSE values. As with loadings, RMSE values were higher than estimates of bias, indicating that the amount that the estimates depart from their true value is mainly a function of sampling error, not bias.

Table 10
Average Performance Indices for the RRM-GRM - Thresholds

Criterion	π	M	SD	Min	Max	
Bias	0.01	0.001	0.006	-0.011	0.021	
	0.05	0.002	0.006	-0.011	0.015	
	0.10	0.000	0.008	-0.020	0.023	
	0.20	0.001	0.007	-0.024	0.018	
Proportion	0.01	0.000	0.009	-0.036	0.052	
	Bias	0.05	-0.001	0.010	-0.050	0.046
		0.10	0.001	0.008	-0.041	0.037
RMSE	0.20	-0.001	0.012	-0.079	0.029	
	0.01	0.058	0.018	0.029	0.106	
	0.05	0.060	0.018	0.035	0.120	
	0.10	0.062	0.020	0.032	0.115	
	0.20	0.065	0.019	0.038	0.108	

Note. π is the true proportion of random responder simulees.

When plotting the population values (i.e. true values) against the bias and RMSE values (see Figures 7 and 8), it can be seen that the thresholds lie almost directly at zero, indicating essentially no bias was present for all of the conditions. As previously noted in Table 10, little to no bias was present on average for the thresholds. Figure 7 demonstrates that this is true for all of the population values, thus making it difficult to discern if bias is related to the true values of the thresholds. However, even though there is very little RMSE present for the thresholds on average, it does appear that RMSE increases for population values at the extremes (see Figure 8). Thus, as the proportion of

random responders increases, sampling error becomes more of a factor. That is, the presence of random responders contributes to sampling error more when thresholds are really low or really high. Even at these true threshold values, however, RMSE is not high enough to be problematic.

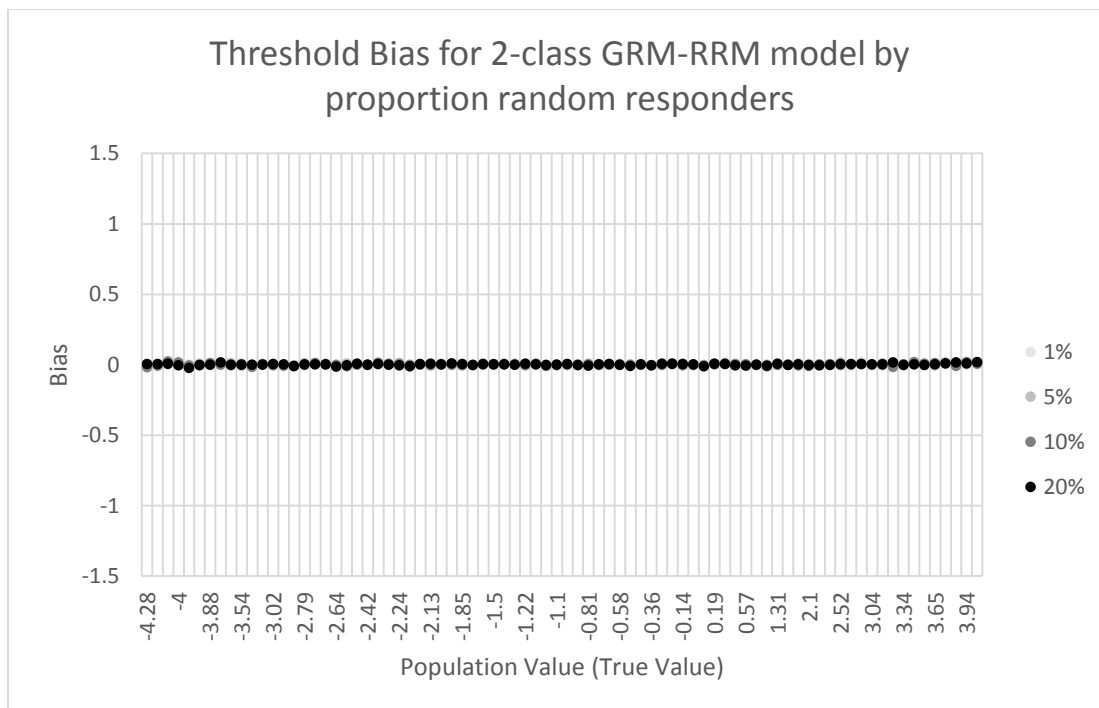


Figure 7. Threshold bias for the RRM-GRM.

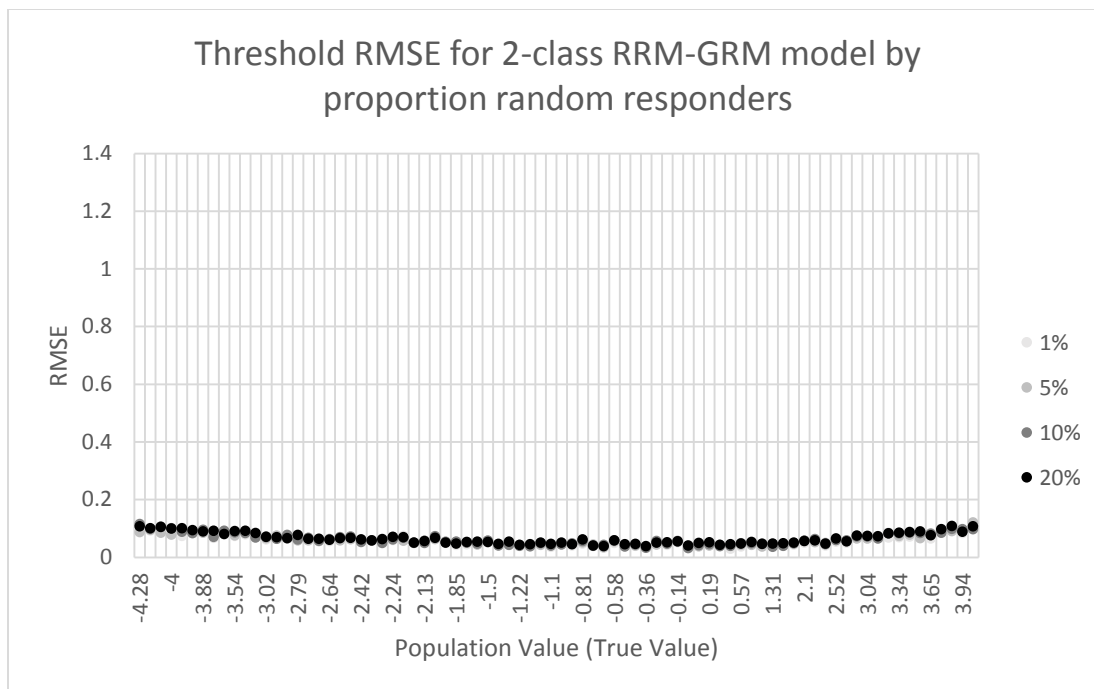


Figure 8. Threshold RMSE for the RRM-GRM.

Thetas. Regarding theta estimates, Tables 11, 12, and 13 contain information on the modal assignment of simulees, means of modal classification, and descriptive statistics for theta using the RRM-GRM, respectively. Prior to examining theta estimates under the RRM-GRM, it is important to consider how the results would be used in an authentic setting. In practice, the posterior probabilities of class membership would serve to assign each examinee to either the valid responding class or the random responding class based on modal assignment. Information pertaining to theta for subjects classified as random responders would not be used since the model identified their responses as random. However, theta information for subjects assigned to the valid responding class would be used; specifically, theta estimates conditional on membership in this class would be consulted.

Because the model would be used in this way in practice, how true valid responders and true random responders would be classified based on modal assignment is of interest. This information is provided in Table 11. The main diagonal includes simulees that have been classified correctly. The values on the main diagonal are very close to the true classification rates in the far right column for each condition, which isn't surprising given the high entropy values in Table 8. When misclassification occurs, there are slightly more true random responders classified as valid responders than there are true valid responders classified as random, but the differences in these two kinds of misclassifications are minor.

Table 11
Modal classification percentages

		Modal Classification					Modal Classification		
		Valid	Random				Valid	Random	
True Classification	Condition 1 Valid	98.95%	0.05%	99.00%	True Classification	Condition 2 Valid	94.78%	0.22%	95.00%
	Random	0.16%	0.84%	1.00%		Condition 2 Random	0.46%	4.54%	5.00%
		99.11%	0.89%	100.00%			95.24%	4.76%	100.00%
		Modal Classification					Modal Classification		
		Valid	Random				Valid	Random	
True Classification	Condition 3 Valid	89.64%	0.36%	90.00%	True Classification	Condition 4 Valid	79.40%	0.60%	80.00%
	Random	0.66%	9.34%	10.00%		Condition 4 Random	0.96%	19.04%	20.00%
		90.30%	9.70%	100.00%			80.36%	19.64%	100.00%

Note . In Conditions 1, 2, 3 and 4 the true proportion (p) of random responders equaled 0.01, 0.05, 0.10 and 0.20, respectively.

Average estimated (unbolded) and true (bolded) theta values are provided for the various classifications in Table 12. A comparison of the estimated and true averages for valid responders classified as valid indicates the extent to which thetas for properly classified valid responders are biased under the RRM-GRM. Table 12 demonstrates that the true and estimated theta average for valid responders assigned to the valid classes are the same, with the exception of the 5% condition where they differ by a value of 0.002. Thus, there is essentially no bias in true valid responder's theta values that are assigned to the correct class with the RRM-GRM. For example, for Condition 1 in Table 12, the average true theta values for the valid responders that were correctly classified as valid responders was 0.00 and the model correctly estimated this value. For valid responders that were misclassified as random responders, their true average theta value was -0.575. Thus, valid responders with lower than average theta values were misclassified as random responders. Likewise with true random responders who were misclassified as valid; their estimated average theta value was - 0.561. Thus, random responders misclassified as valid responders had estimated theta values that were slightly lower than average.

Table 12
Modal Classification Means

		<u>Modal Classification</u>				<u>Modal Classification</u>	
Condition 1		Valid	Random	Condition 2		Valid	Random
True Classification	Valid	0.000	-0.575	True Classification	Valid	0.000	-0.526
		0.000	---			0.002	---
True Classification	Random	---	---	True Classification	Random	---	---
		-0.561	---			-0.579	---

		<u>Modal Classification</u>				<u>Modal Classification</u>	
Condition 3		Valid	Random	Condition 4		Valid	Random
True Classification	Valid	0.003	-0.563	True Classification	Valid	0.005	-0.547
		0.003	---			0.005	---
True Classification	Random	---	---	True Classification	Random	---	---
		-0.598	---			-0.592	---

Note. Average for true theta values are shown in bold. Cells with dashes indicate that a theta average could not be calculated (e.g., because true valid responders assigned to the random responder class do not have estimated theta values, no estimated theta mean is reported for this group). In Conditions 1, 2, 3 and 4 the true proportion (π) of random responders equaled 0.01, 0.05, 0.10 and 0.20, respectively.

To determine if the magnitude and direction of the bias for valid responders classified as valid depends on true theta level, true theta estimates were categorized by range and the average bias (estimated theta-true theta) computed. The pattern of the results is the same for all conditions, so the results of only one of the conditions, the 20% condition, are located in Table 13. Table 13 demonstrates that true thetas at extreme values were more biased than thetas near the average. For example, the mean of theta estimates in the ≤ -3.51 and ≥ 3.51 ranges were 0.602 and -0.699 respectively, whereas the mean of theta estimates in the 0 to 0.49 range was -0.001. Furthermore, the direction

of bias differs depending on whether theta is low or high. In other words, low thetas are positively biased (e.g. theta = ≤ -3.51 , mean = 0.602) and high thetas are negatively biased (e.g. theta = ≥ 3.51 , mean = -0.699).

Table 13

Average Bias by True Theta Ranges using the RRM-GRM (20% Condition)

Range Min	Range Max	<i>M</i>	<i>SD</i>
--	≤ -3.51	0.602	0.306
-3.5	-3.01	0.361	0.276
-3	-2.51	0.234	0.271
-2.5	-2.01	0.137	0.267
-2	-1.51	0.102	0.257
-1.5	-1.01	0.074	0.251
-1	-0.51	0.051	0.250
-0.5	-0.01	0.028	0.257
0	0.49	-0.001	0.263
0.5	0.99	-0.053	0.265
1	1.49	-0.106	0.274
1.5	1.99	-0.128	0.278
2	2.49	-0.153	0.280
2.5	2.99	-0.237	0.297
3	3.49	-0.372	0.287
≥ 3.5	--	-0.699	0.296

Bias does not apply to the other simulees (because they don't have both estimated and true theta values). However, the average means can be inspected to understand true and estimated theta values for those simulees assigned to the wrong class. Table 12 demonstrates that for true random responders misclassified as valid responders, the average estimate thetas are low (e.g., -0.579 in the 5% condition). If this were real data, the practitioner would incorrectly conclude that these responders are low on the construct. The valid responders who have been misclassified as random responders have

an average true theta value that is also lower than the mean. If this were real data, the practitioner would incorrectly conclude that these responders are random responders, when in fact, they are truly low on the construct. Thus, the model has difficulty distinguishing valid responders that are low on the construct from random responders, which is not surprising.

Study 2

RQ1: How are item parameter and theta estimates of the GRM impacted by the presence of random responders in the data set? Because the true parameter values are not known in Study 2, only the difference in parameter estimates when the GRM versus the RRM-GRM was fit to the data were examined (as opposed to examining how parameter estimates compared to their true values). The loading and threshold parameter estimates for each model are displayed graphically in Figures 9 and 10, respectively. With the addition of a second class in the RRM-GRM, factor loading estimates increased by 0.116 on average. In Figure 10, items with negative thresholds appear to be higher in the GRM relative to RRM-GRM, whereas items with positive thresholds are lower using the GRM relative to the RRM-GRM. On average, the loadings for the UMUM-15 were larger by a value of 0.116 in the RRM-GRM relative to the GRM and thresholds were lower by a value of 0.274 in the RRM-GRM relative to the GRM.

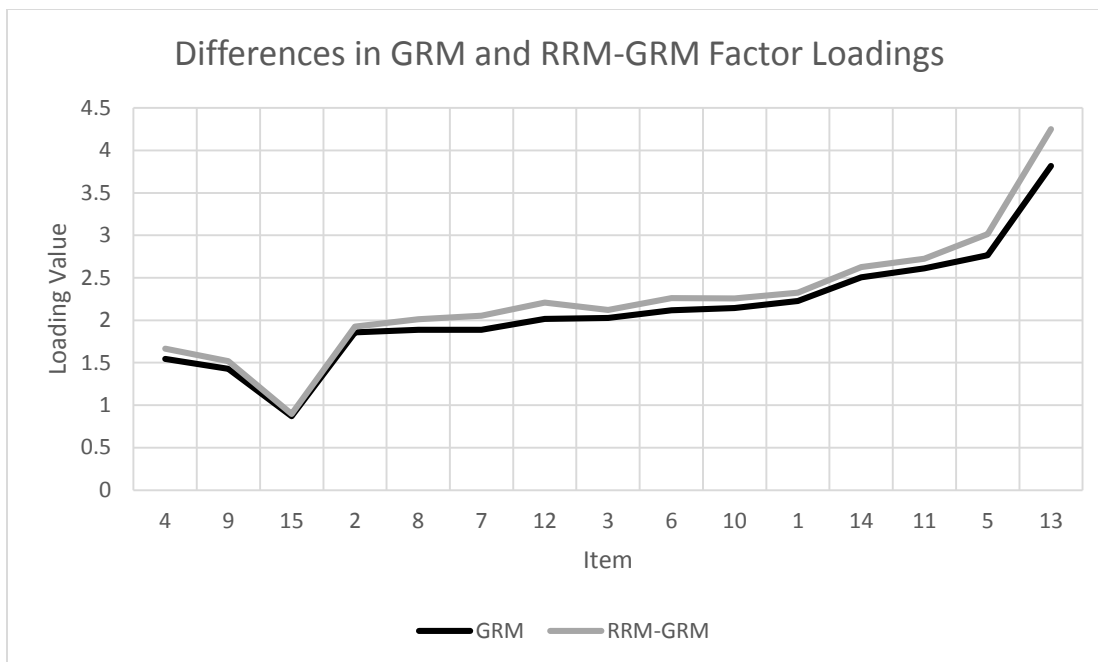


Figure 9. Differences between GRM and RRM-GRM factor loading estimates.

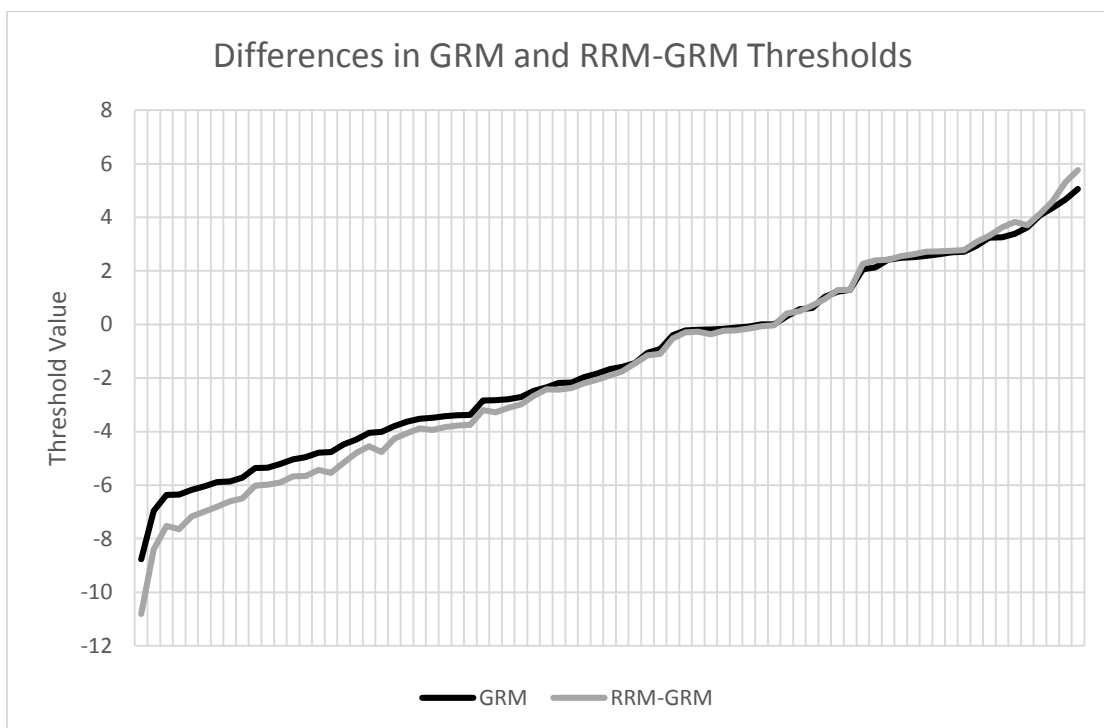


Figure 10. Differences between GRM and RRM-GRM threshold estimates.

RQ2: Which model (the RRM-GRM or GRM) best fits the data? Table 14 conveys the relative fit indices for the GRM compared to the RRM-GRM. The values for LL are higher and AIC, BIC, and SSA-BIC are lower for the RRM-GRM. That is, each index is improved with the RRM-GRM, which supports the use of the RRM-GRM over the GRM.

Table 14
Model fit indices

	GRM	RRM-GRM	Difference
Free parameters	90	91	1
LL	-65570.75	-64895.96	-674.78
AIC	131321.49	129973.93	1347.56
BIC	131878.10	130536.72	1341.38
SSA-BIC	131592.12	130247.57	1344.56

Note. Estimates are model-based.

RQ3: When the RRM-GRM is fit to the data, what is the estimated percentage of random responders? The percentages of valid and random responders in the classes that emerged when the RRM-GRM was fit to the UMUM-15 data are displayed in Table 15. According to the model-based estimates of class proportions, approximately 5.6% of respondents were classified as random responders.

Table 15
Number and percentage of responders in each class

	<i>N</i>	%
Random Responders	200.26	5.6
Valid Responders	3,384.7	94.4

Note. Estimates are model-based.

Looking more closely at responder classification, Table 16 contains the posterior probabilities of responders being classified in a different class than the one they were modally assigned for the RRM-GRM. The probability of a different assignment is small. That is, the average posterior probability of a random responder being classified as a valid responder is 0.092 and the average probability of a valid responder being classified as a random responder is 0.009. In other words, the RRM-GRM identified valid responders with more certainty than random responders. Thus, classification errors are more likely to be made when classifying a random responder. However, the overall classification accuracy is very good for the model, as conveyed by the entropy statistic, value of 0.955.

Table 16
Average posterior probabilities by modal assignment

	Random Responders	Valid Responders
Random Responders	0.908	0.092
Valid Responders	0.009	0.991

RQ5: When the RRM-GRM is fit to real data, does evidence suggest that respondents in the random responding class are randomly responding? For RQ5, the respondent's sex, scores on the effort and importance scales of the SOS, and total score on the UMUM-15 were examined for validity evidence. Sex, effort subscale score, and importance subscale scores were all considered to be potential predictors of group membership, whereas total UMUM-15 score were considered to be outcomes. That is, it was hypothesized that how important a respondent thought the assessments were, how much effort respondents put into them, and the respondent's sex would *predict* class

membership, whereas a respondent's total score on the UMUM-15 would be a *result* of their class membership.

Table 17 contains the coefficients associated with each of the hypothesized predictors. Both sex ($p < 0.001$) and importance ($p = 0.035$) significantly predicted membership in the random responding class. That is, sex is a significant predictor when controlling for effort and importance, and importance is a significant predictor when controlling for sex and effort. Effort ($p = 0.178$) was not a significant predictor. For the sex predictor, the odds of a male (1) being classified as a random responder are higher than those of a female (0) by a factor of 2.016. Additionally, for the importance predictor, for every unit increase in importance, the odds of being classified as a random responder decrease by a factor of 0.956.

Table 17
Predictors of Class Membership

	B	SE	Sig	Exp(B)
Intercept	-2.003	0.426	0.000	0.135
Sex	0.701	0.172	0.000	2.016
Effort	-0.031	0.023	0.178	0.969
Importance	-0.045	0.021	0.035	0.956

To help visualize the relationship between the significant predictors, the probability of membership in the random responder class for males and females for different levels of importance (holding effort at the average) is displayed in Figure 11. It can be seen that, when holding effort constant, the probability of males being classified in the random responder class is higher than for females. As well, probability of membership in the random responding class decreases as importance score increases.

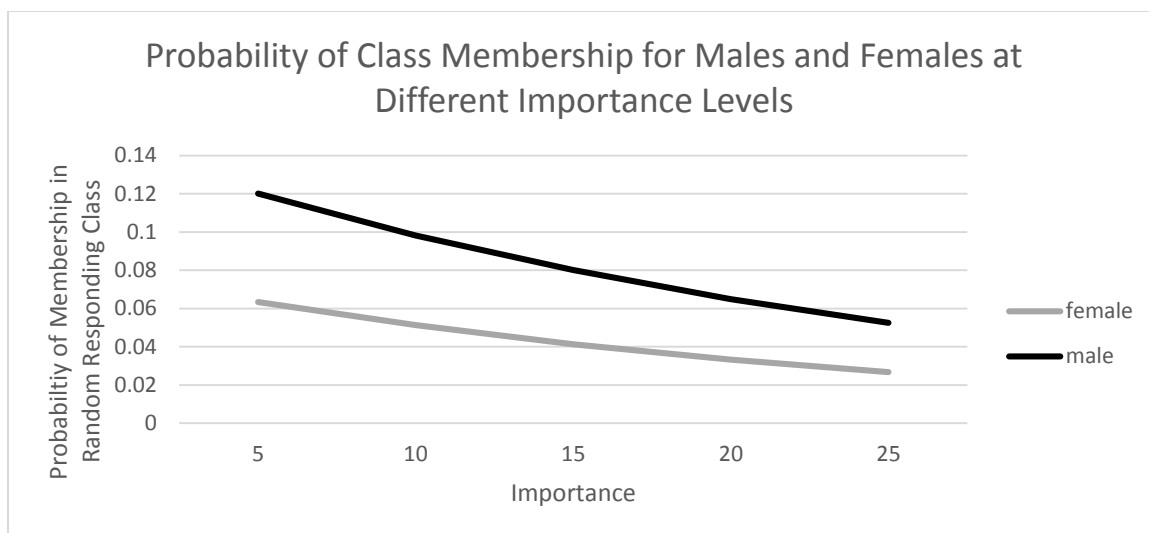


Figure 11. Probability of membership in the random responder class.

Table 18 contains estimated means and variances for total score on the UMUM-15 for the two classes. For the total UMUM-15 score, the average of respondents classified as random responders was lower (50.72) than responders classified in the valid responder group (66.47). The Wald test was performed to test whether the group means are equal across classes. According to the Wald test, the group means do significantly differ. That is, there is a significant difference between classes on total score on the UMUM-15.

Table 18

Means and Variances for total score on the UMUM-15

	Random Responders		Valid Responders	
	Estimate	SE	Estimate	SE
Mean	50.723	1.273	66.474	0.215
Variance	173.020	22.234	127.326	4.411
Wald Test	Value	df	<i>p</i> -value	
	137.063	1	0.000	

V. Discussion

Study 1

Study 1 aimed to answer four research questions that pertained to the differences between fitting two models, the GRM and RRM-GRM, to data containing four various proportions of random responders (1%, 5%, 10%, and 20%). The results contributed to understanding how item parameter and theta estimates are impacted by the presence of random responders when the GRM is fit to data (RQ1) and how they are purified with the use of the RRM-GRM (RQ4). The results also provided information as to the accuracy of the RRM-GRM in estimating the proportion of random responders present (RQ3) and whether the RRM-GRM is the best fitting model when random responders are present (RQ2).

Results from Study 1 indicate that both item parameter and theta estimates are biased when the GRM is fit to a data set containing random responders. This is especially true for loadings and theta estimates when the proportion of random responders present is greater than 0.01 and 0.05 for thresholds. On average, factor loadings were underestimated, thresholds were overestimated, and the average of the theta estimates for valid responders was overestimated. Additionally, larger proportions of random responders were found to be associated with weaker estimation accuracy and higher bias (for loadings, thresholds, and theta estimates) and RMSE (for loadings and thresholds).

Hoogland and Boomsma (1998) suggest that percent bias values lower than 5% are acceptable for parameter estimates. Using this rule to evaluate the minimum and maximum percent bias values in Tables 3 and 4, bias in item parameter estimates was present but minimal in the 1% random responder condition for the GRM. That is, if only

1% of responders are randomly responding, item parameter estimates may not be drastically affected. In the 5%, 10%, and 20% random responder conditions, the presence of a large proportion of random responders makes accurate item parameter estimation problematic, even for a low-stakes setting.

As for theta estimates, because the theta scale is fixed to have a mean of 0 and standard deviation of 1, the presence of random responders forces the thetas of valid responders, on average, to be high (because random responders are given lower thetas). The effect becomes more extreme as the proportion of random responders in the data increases. From a practical perspective, the issue with the use of the GRM for theta estimation in the presence of random responders is incorrect inferences about random responders (who shouldn't receive a theta value) and valid responders (whose thetas, on average, are higher than their true thetas). Inspection of bias in the theta values of valid responders by true theta level (Table 6) indicated overestimation of low theta values and underestimation of high theta values. However, this is not necessarily a function of the presence of random responders in the data set as the same pattern of bias (or nearly the same magnitude) was obtained (see Table 19) when data was generated for 10,000 simulees, all of which were valid responders, and the 1-class GRM was fit to the data. This pattern indicates shrinkage of theta estimates towards the mean and is likely a function of the estimation procedure used, expected-a-posteriori (EAP; Tong & Kolen, 2007).

Table 19
*Average Bias by True Theta Ranges
 when Estimating the GRM with 0%
 Random Responders*

Range Min	Range Max	<i>M</i>	<i>SD</i>
--	≤ -3.51	0.612	---
-3.5	-3.01	0.347	0.349
-3	-2.51	0.224	0.299
-2.5	-2.01	0.179	0.244
-2	-1.51	0.082	0.258
-1.5	-1.01	0.079	0.251
-1	-0.51	0.053	0.245
-0.5	-0.01	0.013	0.253
0	0.49	-0.021	0.261
0.5	0.99	-0.067	0.264
1	1.49	-0.131	0.274
1.5	1.99	-0.159	0.25
2	2.49	-0.206	0.283
2.5	2.99	-0.233	0.294
3	3.49	-0.500	0.216
≥3.5	--	-0.915	0.368

When the RRM-GRM was fit to the data set, item parameters and theta estimates were estimated with minimal to no bias for all proportions of random responders. For item parameter estimates, bias and RMSE in both loadings and thresholds were minimal for all conditions, but RMSE values were higher than bias values indicating that the amount that the estimates depart from their true value is mainly a function of sampling error, not bias. As the proportion of random responders increased, sampling error appeared to become more of a factor. However, even in the conditions with a large proportion of random responders, values of RMSE for the item parameters were not problematic.

For theta estimates, essentially no bias on average was detected in true valid responder's theta values that were assigned to the correct class with the RRM-GRM. When bias in thetas for valid responders was inspected by theta value, the same pattern of results as found with the GRM were observed. As noted, this same pattern occurred when GRM-generated data with no random responders was fit to the GRM (Table 20). It is therefore more a function of the estimation procedure used than of the RRM-GRM model itself.

Even though bias does not apply to the other simulees, the average thetas for simulees assigned to the wrong condition were inspected and it was found that for true random responders misclassified as valid responders, the average estimate thetas were low, which would lead one to incorrectly conclude that these responders were low on the construct. In addition, the valid responders misclassified as random responders had an average true theta value that was also lower than the mean, which would lead one to incorrectly conclude that these responders are random responders, when they are actually truly low on the construct. The results indicate that the RRM-GRM has difficulty distinguishing valid responders that are low on the construct from random responders, which is not surprising.

The third goal of Study 1 was to evaluate the accuracy of the RRM-GRM in estimating the proportion of random responders present. To explore this, the average estimated proportion for each of the conditions and classification accuracy were examined. The RRM-GRM estimated the proportion of random responders for the 1%, 5% and 10% conditions to be the true proportion, and the 20% condition was only off by 0.001. As for classification accuracy for each of the conditions, the entropy statistic was

higher for the conditions with lower proportions of random responders than for the conditions with higher proportions, but we still considered it to be sufficiently high, with all values above 0.90.

The last aim of Study 1 was to determine if the GRM or RRM-GRM was preferable for use with datasets containing random responders. To explore this, loglikelihood based model fit indices were compared, and it was found that each index improved with the RRM-GRM. Furthermore, as the proportion of random responders increased, the difference between the model fit indices increased as well. As a reminder, the RRM-GRM only requires one additional parameter to be estimated than the GRM.

Magnitude and direction of bias observed.

GRM. For loadings, it was found that the amount of bias and RMSE present depends on the value of the loadings. That is, the higher the factor loading value, the worse the negative bias and RMSE. A dependency was also found with thresholds and theta estimates. Particularly, thresholds at the extremes (low or high values) have weaker estimation accuracy than thresholds that are average, with negative thresholds being overestimated and positive thresholds underestimated. For theta estimates, true thetas at extreme values were more biased than thetas near the average, with low thetas being positively biased and high thetas being negatively biased. Again, the pattern of bias in the theta estimates is more a function of the estimation procedure than the use of the GRM with data including random responders.

RRM-GRM. Since loadings and thresholds were estimated with little to no bias and RMSE, it was difficult to discern if bias and RMSE were related to the true values of the parameters. However, for factor loadings, it did appear that as the population values

increased, so did RMSE, indicating that the presence of random responders contributes to sampling error more when factor loadings are high. For thresholds, it appeared that RMSE increases for population values at the extremes. That is, the presence of random responders contributed to sampling error more when thresholds were really low or really high. However, the values of bias and RMSE were so low for item parameters when the RRM-GRM was used that their accurate estimation with this model does not appear to be an issue. As for theta, only bias was evaluated and it appeared that the magnitude and direction of the bias for valid responders classified as valid depends on true theta level. Specifically, true thetas at extreme values were more biased than thetas near the average and the direction of bias was found to differ depending on whether theta was low or high. That is, low thetas were positively biased and high thetas were negatively biased. Again, the pattern of bias in the theta estimates is more a function of the estimation procedure than the use of the RRM-GRM.

Implications. The results from Study 1 help provide some understanding of the consequences associated with fitting the GRM to a data set where random responders are present and the benefits of using a model that attempts to account for such respondents, the RRM-GRM. If the GRM is used, both item parameter and theta estimates will be biased, especially when the proportion of random responders in the dataset is greater than 0.01. On average, factor loadings will be underestimated, and thresholds and theta estimates for valid responders will be overestimated with increasing bias and RMSE (for thresholds) as the proportion of random responders goes up. These negative implications are especially a concern for the 10% and 20% conditions because practitioners often use item parameters to evaluate how well a test is working. That is, loadings and thresholds

could be a factor in deciding whether to keep an item on an assessment and in making conclusions regarding reliability. For example, in this study when the GRM was fit to the data set with 20% random responders included, large loadings were underestimated by more than 0.7 units (e.g., true value was 2.35, but the estimated value in the 20% condition was 1.61). A problem with this underestimation is that a practitioner could decide to drop or modify items because of low loadings, when in fact their low loadings are only due to the presence of random responders.

Another issue with biased parameters is the fact that the value of the loadings influences how peaked item information functions are. Since item information functions are added together to get a test information function, if loadings are too small, then the test information function will be too low. Thus, this might lead one to conclude that the scale being evaluated is not as reliable as it really is when random responders are present in the data set. For example, it can be seen in Figure 12 how the test information function (TIF) changes, and thus IRT reliability changes, when the GRM is used and the proportion of random responders in the data set increases. That is, information is reduced when more random responders are present, but the information peaks do not seem to be impacted. The same sort of issue occurs with thresholds. For example, if a practitioner is attempting to create a scale that will reliably measure respondents with certain theta values and thresholds are estimated incorrectly because of the presence of random responders, items might end up being thrown out or revised because it is concluded that the item(s) are not suitable for the targeted theta range at hand.

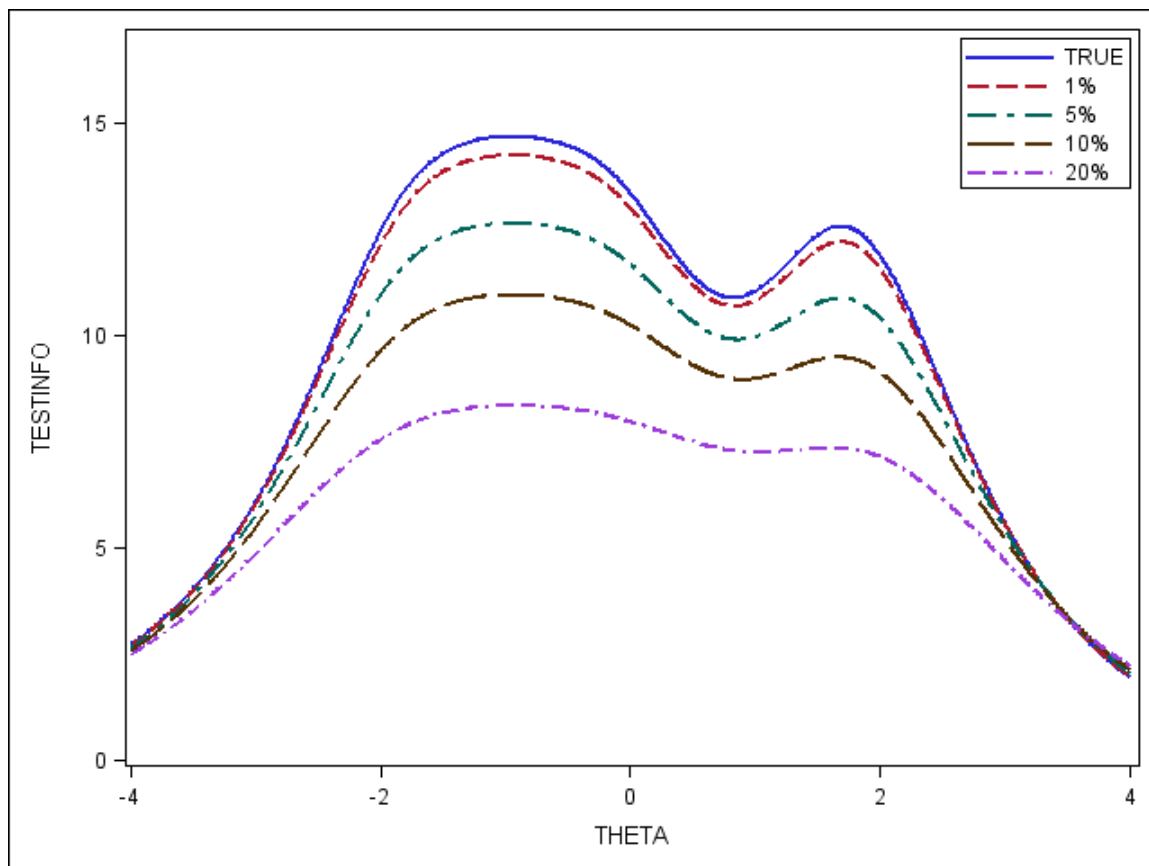


Figure 12. Test information function (TIF) for all conditions in the GRM.

Use of the RRM-GRM for situations in which random responders are present looked promising in this study. Item parameter estimates and theta estimates for valid responders assigned to the correct class were estimated with minimal to no bias for all proportions of random responders, but it did appear that as the proportion of random responders increased, sampling error appeared to become more of a factor. However, it does appear that practitioners may still be apt to make incorrect decisions in some instances. For example, when true random responders are misclassified as valid responders, practitioners may erroneously conclude that these responders were low on the construct. In addition, when valid responders are misclassified as random responders,

they are actually truly low on the construct, but the incorrect conclusion that they are random responders could be made.

Limitations. There are four limitations of this design that should be considered when examining the results that include the number of replications, sample size, length of the test, and the patterns of responses studied. First, since this was the initial study using the RRM-GRM, only 100 replications of each condition were conducted. Because the study did not contain a higher number of replications, the empirical standard error may be inaccurate, leading to an inability to confidently draw conclusions regarding estimate variability. Future studies should include a larger number of replications to better understand variability of estimates.

A second limitation of the study was the sample size. For all conditions, a sample size of 5,000 was used. The use of a set number of simulees can inhibit generalizability for instances with much different sample sizes. For example, what if the number of responders was 400 or 7,000? Future studies should explore similar proportions of simulees with various sample sizes for better understanding of how bias, RMSE, model fit and simulee classification are affected.

A third limitation of Study 1 had to do with the length of the test. As with the limitations pertaining to sample size, this study utilized only 20 items. Again, the use of a set number of items can inhibit generalizability for instances with tests that are longer or shorter. For example, what if the number of items on a test was 10 or 60? Future studies should explore similar proportions of simulees with various test lengths in order to develop a better understanding of how bias, RMSE, model fit and simulee classification are affected.

A final limitation has to do with the response styles studied. Study 1 focuses exclusively on the random responder response pattern. This response pattern is an extreme case and is characterized by the tendency of the participant to respond to items carelessly or arbitrarily. With this study, respondents are only considered to be random responders *on the entire scale* or not random responders. That is, if a responder tried on at least the first several items on the scale, then they may not be identified as a random responder.

Other response patterns exist that result from careless responding that are not random; for instance, a pattern in which the same response option is provided to all items. Unfortunately, response patterns provided by amotivated responders that are anything other than a random response pattern are not captured by this model. The ways in which respondents complete non-cognitive assessments differ; thus, other response styles, such as acquiescence, neutral, or disacquiescence, should be incorporated in future studies.

Study 2

The purpose of Study 2 was to apply the RRM-GRM to an authentic low-stakes dataset to capture and account for random responders. As anticipated, a small proportion of respondents, approximately 5.6%, were identified to be in the “random responder” class with a high degree of certainty. While classification accuracy was high, the RRM-GRM identified valid responders with more certainty than random responders. Regarding model fit, evidence from relative fit indices supports the use of the RRM-GRM over the GRM. Because the true parameter values are not known in Study 2, only the difference in parameter estimates across models were examined and results showed that on average, when the RRM-GRM was fit to the data, loadings for the UMUM-15

increased in value and thresholds decreased in value relative to when the GRM was used. These results align with the change in parameter estimates across the two models in Study 1. Thus, the RRM-GRM appeared to be performing as expected and purifying the item parameter estimates.

When working with authentic test data, the true population parameters are unknown and thus additional hypotheses regarding the make-up of the classes should be considered. In Study 2, it was hypothesized that one of the classes captured random responders who did not actively attempt to answer the items on the scale starting with the first item. However, evidence is needed to support the idea that respondents in the random responding class are actually randomly responding. That is, another hypothesis is that participants classified in the random responder class are actually not random responders, but actually are actually low to moderate on the construct of university mattering. Theoretically, the hypothesis that those in the random responder class are actually randomly responding on the UMUM-15 is championed because the test is administered in a low-stakes setting with no individual consequences to the participant.

To investigate this competing hypothesis, predictors of class membership, including gender and total scores on the effort and importance scales of the SOS, were examined. Both sex and importance were found to significantly predict membership in the random responding class, but effort was found not to be a significant predictor. The results provided evidence that it was more likely for a male to be classified as a random responder than a female, and that as importance score increased, the likelihood of being classified as a random responder decreased. One could contend that the counter argument is supported by these results if university mattering is lower for males than

females and also lower for those who think it is less important to do well on the university assessments.

Additionally, an outcome, total UMUM-15 score, was investigated for supporting evidence. For total UMUM-15 score, participants in the random responding class had a lower average score than the participants in the valid responder class. This finding was expected. Specifically, the average total UMUM-15 score for those in the random responding class was 50.72, which is encouraging because this value is close to 52.5, the average that would be expected under random responding. Furthermore, a significant difference was found between classes on total score on the UMUM-15. However, it could be argued that the participants captured in the random responding class are those with moderate levels of university mattering.

Future research. Study 2 included only one authentic dataset collected from a non-cognitive low-stakes test administered in a university setting. Replication studies that include different test types, lengths, sample sizes, and data from low-stakes settings outside of the university are desirable. For example, the UMUM-15 that was administered in Study 2 is a 15-item assessment on the topic of university mattering. A test containing more or less than 15-items and pertaining to a different construct than university mattering should be used in a replication study. Moreover, future samples should be taken from more diverse populations.

Another area for future exploration concerns the variables used in an attempt to provide validity evidence for class membership. Specifically, a greater number and wider range of external variables for the validity studies are needed. In Study 2, only four variables were used in providing validity evidence, but the results were not overly

convincing because the construct selected (university mattering) was related to the external variables in the same way as motivation. Future research may also want to select different external variables than those used in this study. For example, two scales from the SOS (effort and importance) were used as validity coefficients, but the SOS itself has a few limitations including the fact that it is a self-report measure and that it is administered after a battery of tests in a low-stakes setting. Thus, the SOS may not be indicative of a respondent's motivation level if participants respond randomly or untruthfully to the measure. Additionally, fatigue may have set in for examinees and/or because the conditions are low-stakes, examinees may provide thoughtless responses. Thus, if examinees carelessly complete the measure, the estimate of the average difference in motivation between the classes may appear to be lower than in actuality.

Comparing Study 1 and Study 2 Results

The results from Study 2 were expected to be similar to the Results found in Study 1. However, it is recognized that the length of the test and number of response options differed slightly between the two studies. That is, Study 1 used a 20-item measure with five response options, whereas the measure in Study 2 was 15-items with six response options. The findings regarding model fit were consistent across studies in that both provided evidence of better model fit for the RRM-GRM as opposed to the GRM. Both studies also provided evidence of the ability to distinguish two classes of respondents with high certainty. Like with Study 1, Study 2 found that the factor loadings under GRM were negatively biased. That is, the factor loadings were too low under the GRM, but were estimated to be higher values (or closer to true values in Study 1) with the GRM-RRM. Both studies also found similar patterns with thresholds. That

is, Study 1 found that negative thresholds were too high and positive thresholds were too low with the GRM and Study 2 found that negative thresholds appears higher and positive thresholds lower with the GRM.

Conclusions

The administration of tests for assessment and accountability purposes are a current requisite for higher education institutions in today's society. Data collected from these assessments are not only reported to external stakeholders, but also used to aid in augmentation of curriculum and facilitate decision making in academic and student affairs programs. With little to no personal consequences tied to these assessments, low motivation will remain a barrier for practitioners aiming to making valid inferences from the results. However, many modeling techniques to assist with purifying parameter estimates have been developed in an attempt to combat this problem.

In this study, an IRT mixture modeling technique was extended and applied to simulated and authentic non-cognitive polytomously scored data to examine its functioning. The results of the study are promising, but further research is necessary. Specifically, it appears that the RRM-GRM was able to classify respondents into separate classes under four different data conditions with the addition of only one extra estimated parameter. With both simulated and authentic data, the RRM-GRM had improved model fit over the GRM and less biased and more accurate parameter estimates, especially when the proportion of random responders is large. That is, by estimating only one extra parameter, the RRM-GRM provides a plethora of additional information than the GRM, which is a huge benefit of the model. Despite the fact that more validity evidence is needed to support the characteristics of the emerging classes, if the presence of random

responders in a data set is a concern, the RRM-GRM would be worth estimating for data used in the aggregate as it would provide the proportion of random responders, offer theta estimates only for those in valid responding class, and provide “purified” item parameter estimates. That is, extreme caution should be taken if the purpose of the model’s use is to identify specific examinees, as further external validity evidence is required to support the classes of examinees as “valid” and “random”. Another attraction to the model is that it appears to perform well from the simulation study and is easy to estimate in Mplus. Although use of the RRM-GRM might result in misclassification of a very small proportion of those low on the construct as random responders and vice versa, if there are no severe consequences in doing so, the RRM-GRM may be a better model for use.

Appendix A

SAS Syntax for Generating Datasets

```

FILENAME X 'C:\ ';

%INCLUDE IO(IRTGEN);

DATA paras;
input a cb1-cb4;
cards;
0.95 -4.26 -2.90 -1.25 2.01
1.48 -2.45 -1.44 -0.60 1.45
1.46 -2.07 -1.27 0.16 2.11
1.49 -1.75 -0.76 0.13 2.02
1.38 -2.19 -1.27 -0.35 1.52
1.35 -2.88 -1.97 -0.51 1.87
0.96 -3.77 -2.23 -1.27 1.34
1.32 -3.24 -2.29 -0.49 1.93
1.08 -3.28 -2.09 0.49 3.09
2.00 -1.57 -0.75 -0.13 1.68
1.22 -1.39 0.08 1.07 2.99
0.89 -2.97 -1.50 -0.41 2.44
2.05 -2.05 -1.19 -0.15 1.87
1.59 -1.20 -0.24 0.61 2.48
2.31 -1.68 -0.95 -0.25 1.69
2.07 -1.90 -1.08 -0.39 1.59
1.55 -1.80 -0.80 0.10 1.96
0.92 -3.82 -2.63 -1.20 1.67
1.64 -1.40 -0.50 0.35 2.10
2.35 -1.70 -0.90 -0.06 1.81

;
run;

%macro simulate;
%global numex numrr;
%do cond=1 %to 4;
  %do rep=1 %to 100;
    %if &cond=1 %then %do; %let numex=4950; %let numrr=4951; %end;
    %if &cond=2 %then %do; %let numex=4750; %let numrr=4751;%end;
    %if &cond=3 %then %do; %let numex=4500; %let numrr=4501;%end;
    %if &cond=4 %then %do; %let numex=4000; %let numrr=4001;%end;

    %IRTGEN(MODEL=GR, DATA=paras, OUT=OUT&cond&rep, NI=20, NE=&numex);
    proc means data=OUT&cond&rep; var theta r1-r20; title "Condition
    &cond with &numex examinees - rep &rep"; run;

    data OUT2&cond&rep; set OUT&cond&rep;
    randomresp=0;
    id=_n_;
    run;

    proc means data=OUT2&cond&rep; var r1-r20 theta; run;

    data randomresp&cond&rep;

```



```
        randomresp=1;
        array r(20) r1-r20;
        do id=&numrr to 5000;
            do item=1 to 20;
                r(item)=rantbl(0,0.20,0.20,0.20,0.20,0.20);
            end;
            output;
        end;
    run;
    proc means data=randomresp&cond&rep; var r1-r20; run;
    data both&cond&rep; set OUT2&cond&rep randomresp&cond&rep;
    total=sum(of r1-r20);

file "C:\.dat" dlm=' ';
    put id randomresp theta r1-r20;
run;

    proc means data=both&cond&rep; var total; class randomresp;
run;
    %end;
    %end;
    %mend;

%simulate;
```

Appendix B

SAS Syntax for Generating Mplus Syntax

GRM

```

%let path=C:\; *simulation computer;
%let path2=C:\; *output computer;
%let path3=C:\; *location to store data for plots;

*OPTIONS nonumber nodate nocenter formdlim=' ' pagesize=MAX
linesize=MAX;
  TITLE; ODS TRACE OFF;
%macro createsyn;
%do cond=1 %to 4;
  %do rep=1 %to 100;
data _null_;
file "&path\GRMsyn&cond&rep..inp" PRINT;
  PUT @1 "TITLE: GRM&cond&rep;";
  PUT @1 "DATA: FILE='&path\Data Sets\"
/ @1 "out&cond&rep..dat';"
/
/ ;
  PUT @1 "VARIABLE:"
/ @5 "NAMES ARE id randomresp theta r1-r20;"
/ @5 "USEVARIABLES r1-r20 ;"
/ @5 "CATEGORICAL ARE r1-r20;"
/ @5 "MISSING ARE .; "
/ @5 "IDVARIABLE IS id;"
/ @5 "CLASSES=c(1);"
/
/ ;
  PUT @1 "ANALYSIS:"
/ @5 "ESTIMATOR IS ML;"
/ @5 "LINK IS LOGIT;"
/ @5 "ALGORITHM=INTEGRATION;"
/ @5 "TYPE=mixture;"
/ @5 "STARTS=200 50;"
/ @5 "PROCESSORS=4 4 ;"
/
/ ;
  PUT @1 "MODEL:"
/
/ ;
  PUT @5 "%Overall%"
/ @5 "F by r1-r20* (rr1-rr20);"
/ @5 "[F@0];"
/
/ ;

  PUT @5 "%C#1%"
/ @5 "[r1$1-r20$1*];"
/ @5 "[r1$2-r20$2*];"
/ @5 "[r1$3-r20$3*];"
/ @5 "[r1$4-r20$4*];"
/ @5 "F@1;"
/
/ ;

  PUT @5 "MODEL CONSTRAINT:"
/ @5 "DO (1,20) rr#>0;"

```

```

/          ;

PUT          @1 "SAVEDATA:"
/           @5 "RESULTS ARE '&path\"
/           @5 "GRM_out&cond&rep..dat'";
/          ;

run;
      %end;
%end;
%mend;
%createsyn;

%macro runmplus;
%do cond=1 %to 4;
  %do rep=1 %to 100;
option noxwait xsync;
      X CALL "C:\Program Files\Mplus\mplus.exe"
          "&path\GRMsyn&cond&rep..inp"
          "&path\GRMresults&cond&rep..out";

%end;
%end;
%mend;
%runmplus;

```

RRM-GRM

```

%let path=C; *simulation computer;
%let path2=C:\; *output computer;
%let path3=C:\; *location to store data for plots;

*OPTIONS nonumber nodate nocenter formdlim=' ' pagesize=MAX
linesize=MAX;
  TITLE; ODS TRACE OFF;
%macro createsyn;
%do cond=1 %to 4; *# of conditions;
  %do rep=1 %to 100; *# of replications for each condition;
data _null_;
file "&path\GRMRRMsyn&cond&rep..inp" PRINT;
  PUT @1 "TITLE: GRMRRM&cond&rep;";
  PUT @1 "DATA: FILE='&path\Data Sets\"
/ @1 "out&cond&rep..dat'";
/
/          ;
  PUT @1 "VARIABLE:"
/ @5 "NAMES ARE id randomresp theta r1-r20;"
/ @5 "USEVARIABLES r1-r20 ;"
/ @5 "CATEGORICAL ARE r1-r20;"
/ @5 "MISSING ARE .; "
/ @5 "CLASSES=c (2);"
/
/          ;
  PUT @1 "ANALYSIS:"
/ @5 "ESTIMATOR IS ML;"
/ @5 "LINK IS LOGIT;"

```

```

/          @5 "ALGORITHM=INTEGRATION;"
/          @5 "TYPE=mixture;"
/          @5 "STARTS=200 50;"
/          @5 "PROCESSORS=4 4 ;"
/          ;
PUT       @1 "MODEL:"
/          ;
PUT       @5 "%Overall%"
/          @5 "F by r1-r20*;"
/          @5 "[F@0];"
/          ;

PUT       @5 "%C#1%"
/          @5 "F by r1-r20@0;"
/          @5 "[r1$1-r20$1@-1.386294361];"
/          @5 "[r1$2-r20$2@-0.405465108];"
/          @5 "[r1$3-r20$3@0.405465108];"
/          @5 "[r1$4-r20$4@1.386294361];"
/          @5 "F@0;"
/          ;

PUT       @5 "%C#2%"
/          @5 "F by r1-r20* (rr1-rr20);"
/          @5 "[r1$1-r20$1*];"
/          @5 "[r1$2-r20$2*];"
/          @5 "[r1$3-r20$3*];"
/          @5 "[r1$4-r20$4*];"
/          @5 "F@1;"
/          ;

PUT       @5 "MODEL CONSTRAINT:"
/          @5 "DO (1,20) rr#>0;"
/          ;

PUT       @1 "SAVEDATA:"
/          @5 "RESULTS ARE '&path\"
/          @5 "GRMRRM_out&cond&rep..dat'";
/          ;
run;
    %end;
%end;
%mend;
%createsyn;

%macro runmplus;
%do cond=1 %to 4; *# of conditions;
    %do rep=1 %to 100; *# of replications for each condition;
option noxwait xsync;
        X CALL "C:\Program Files\Mplus\mplus.exe"
                "&path\GRMRRMsyn&cond&rep..inp"
                "&path\GRMRRMresults&cond&rep..out";
    %end;
%end;
%mend;
%runmplus;

```

Appendix C

SAS Syntax for Reading Datasets into SAS

GRM

```

%macro readin;
%do cond=1 %to 4;
  %do rep=1 %to 100;
    %macro heynow;
      data mandy;
        infile "&path2\GRM_out&cond&rep..dat";
        input load1-load20
              %do i=1 %to 20;
                t1_&i t2_&i t3_&i t4_&i
              %end;
        loadSE1-loadSE20
              %do i=1 %to 20;
                t1SE_&i t2SE_&i t3SE_&i t4SE_&i
              %end;
        LL numpara AIC BIC SSABIC entropy;
        cond=&cond; rep=&rep;
        run;
        proc transpose data=mandy
          out=mandytr&cond&rep(rename=(coll=cond&cond.rep&rep)); run;
        proc sort data=mandytr&cond&rep; by _NAME_; run;
        %mend;

    %heynow;

  %end;
%end;
%mend;

%readin;

data true;
input _NAME_ $ true;
cards;
load1 0.95
load2 1.48
load3 1.46
load4 1.49
load5 1.38
load6 1.35
load7 0.96
load8 1.32
load9 1.08
load10 2
load11 1.22
load12 0.89
load13 2.05
load14 1.59
load15 2.31
load16 2.07

```

load17	1.55
load18	0.92
load19	1.64
load20	2.35
t1_1	-4.05
t2_1	-2.76
t3_1	-1.19
t4_1	1.91
t1_2	-3.63
t2_2	-2.13
t3_2	-0.89
t4_2	2.15
t1_3	-3.02
t2_3	-1.85
t3_3	0.23
t4_3	3.08
t1_4	-2.61
t2_4	-1.13
t3_4	0.19
t4_4	3.01
t1_5	-3.02
t2_5	-1.75
t3_5	-0.48
t4_5	2.1
t1_6	-3.89
t2_6	-2.66
t3_6	-0.69
t4_6	2.52
t1_7	-3.62
t2_7	-2.14
t3_7	-1.22
t4_7	1.29
t1_8	-4.28
t2_8	-3.02
t3_8	-0.65
t4_8	2.55
t1_9	-3.54
t2_9	-2.26
t3_9	0.53
t4_9	3.34
t1_10	-3.14
t2_10	-1.5
t3_10	-0.26
t4_10	3.36
t1_11	-1.7
t2_11	0.1
t3_11	1.31
t4_11	3.65
t1_12	-2.64
t2_12	-1.34
t3_12	-0.36
t4_12	2.17
t1_13	-4.2
t2_13	-2.44
t3_13	-0.31
t4_13	3.83
t1_14	-1.91

```

t2_14 -0.38
t3_14 0.97
t4_14 3.94
t1_15 -3.88
t2_15 -2.19
t3_15 -0.58
t4_15 3.9
t1_16 -3.93
t2_16 -2.24
t3_16 -0.81
t4_16 3.29
t1_17 -2.79
t2_17 -1.24
t3_17 0.16
t4_17 3.04
t1_18 -3.51
t2_18 -2.42
t3_18 -1.1
t4_18 1.54
t1_19 -2.3
t2_19 -0.82
t3_19 0.57
t4_19 3.44
t1_20 -4
t2_20 -2.12
t3_20 -0.14
t4_20 4.25
;
run;

proc sort data=true; by _NAME_; run;

%macro alltog;
data all;
merge
  %do cond=1 %to 4;
    %do rep=1 %to 100;
      mandytr&cond&rep
    %end;
  %end;
;
by _NAME_;
run;
%mend;

%alltog;

proc sort data=all; by _NAME_; run;

data final; merge all true; by _NAME_; run;
/*average estimate across replications for each condition*/

%let rep=100;
data all; set final;
  avg_cond1 = mean(of cond1rep1 - cond1rep&rep);
  avg_cond2 = mean(of cond2rep1 - cond2rep&rep);

```

```

avg_cond3 = mean(of cond3rep1 - cond3rep&rep);
avg_cond4 = mean(of cond4rep1 - cond4rep&rep);

    bias_cond1= (avg_cond1-true);
    bias_cond2= (avg_cond2-true);
    bias_cond3= (avg_cond3-true);
    bias_cond4= (avg_cond4-true);

    pctbias_cond1= (bias_cond1/true);
    pctbias_cond2= (bias_cond2/true);
    pctbias_cond3= (bias_cond3/true);
    pctbias_cond4= (bias_cond4/true);

samplingvar_cond1 = ((var(of cond1rep1 - cond1rep&rep))*(&rep-1))/&rep;
samplingvar_cond2 = ((var(of cond2rep1 - cond2rep&rep))*(&rep-1))/&rep;
samplingvar_cond3 = ((var(of cond3rep1 - cond3rep&rep))*(&rep-1))/&rep;
samplingvar_cond4 = ((var(of cond4rep1 - cond4rep&rep))*(&rep-1))/&rep;

MSE_cond1 = ((bias_cond1**2)+samplingvar_cond1);
MSE_cond2 = ((bias_cond2**2)+samplingvar_cond2);
MSE_cond3 = ((bias_cond3**2)+samplingvar_cond3);
MSE_cond4 = ((bias_cond4**2)+samplingvar_cond4);

RMSE_cond1 = (MSE_cond1**.5);
RMSE_cond2 = (MSE_cond2**.5);
RMSE_cond3 = (MSE_cond3**.5);
RMSE_cond4 = (MSE_cond4**.5);
run;

/**** Generating data for plots ****/

/*Loadings*/
data loading; set all;
if index(_NAME_, "load")=1;
if index(_NAME_, "loadSE")=0;
run;

proc means data=loading;
var bias_cond1-bias_cond4 pctbias_cond1-pctbias_cond4 RMSE_cond1-
RMSE_cond4;
run;

data loadingBIAS; set loading;
KEEP _NAME_ TRUE bias_cond1-BIAS_COND4;
run;
PROC SORT DATA=LOADINGBIAS; BY TRUE; RUN;

data loadingRMSE; set loading;
KEEP _NAME_ TRUE RMSE_cond1-RMSE_cond4;
run;

PROC SORT DATA=LOADINGRMSE; BY TRUE; RUN;

data loadingAVG; set loading;
KEEP _NAME_ TRUE AVG_COND1-AVG_COND4;
run;

```



```

PROC SORT DATA=LOADINGAVG; BY TRUE; RUN;

proc export data=loadingBIAS
  outfile="&path3\GRMstudy1plots.xls"
  replace
dbms=excel2002;
  sheet='loading_BIAS';
run;

proc export data=loadingrmse
  outfile="&path3\GRMstudy1plots.xls"
  replace
dbms=excel2002;
  sheet='loading_RMSE';
run;

proc export data=loadingAVG
  outfile="&path3\GRMstudy1plots.xls"
  replace
dbms=excel2002;
  sheet='loading_AVG';
run;

/*Thresholds*/
data thresholds; set all;
if substr(_NAME_,1,1)="t";
run;

proc means data=thresholds;
var bias_cond1-bias_cond4 pctbias_cond1-pctbias_cond4 RMSE_cond1-
RMSE_cond4;
run;

data threshBIAS; set thresholds;
KEEP _NAME_ TRUE bias_cond1-BIAS_COND4;
run;
PROC SORT DATA=threshBIAS; BY TRUE; RUN;

data threshRMSE; set thresholds;
KEEP _NAME_ TRUE RMSE_cond1-RMSE_cond4;
run;

PROC SORT DATA=threshRMSE; BY TRUE; RUN;

data threshAVG; set thresholds;
KEEP _NAME_ TRUE AVG_COND1-AVG_COND4;
run;

PROC SORT DATA=threshAVG; BY TRUE; RUN;

proc export data=threshBIAS
  outfile="&path3\GRMstudy1plots.xls"
  replace
dbms=excel2002;
  sheet='thresh_BIAS';

```

```

run;

proc export data=threshrmse
  outfile="&path3\GRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='thresh_RMSE';
run;

proc export data=threshAVG
  outfile="&path3\GRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='thresh_AVG';
run;

/*Fit Indices*/
data fit; set all;
if _NAME_ in ("AIC", "BIC", "SSABIC", "LL") ;
keep _NAME_ avg_cond1-avg_cond4;
run;
proc print data=fit; run;

proc export data=fit
  outfile= "&path3\GRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='fit';
run;

```

RRM-GRM

```

%macro readin;
%do cond=1 %to 4;
  %do rep=1 %to 100;
    %macro heyknow;
      data mandy;
      infile "&path2\GRMRRM_out&cond&rep..dat";
      input load1-load20
          %do i=1 %to 20;
            t1_&i t2_&i t3_&i t4_&i
          %end;
      mixprop
      loadSE1-loadSE20
          %do i=1 %to 20;
            t1SE_&i t2SE_&i t3SE_&i t4SE_&i
          %end;
      mixpropSE
      LL numpara AIC BIC SSABIC entropy;
      cond=&cond; rep=&rep;
    %mend heyknow;
  %mend rep;
%end cond;
%mend readin;

```

```

run;
proc transpose data=mandy
out=mandytr&cond&rep(rename=(coll=cond&cond.rep&rep)); run;
proc sort data=mandytr&cond&rep; by _NAME_; run;
%mend;

%heynow;

%end;
%end;
%mend;

%readin;

data true;
input _NAME_ $ true;
cards;
load1 0.95
load2 1.48
load3 1.46
load4 1.49
load5 1.38
load6 1.35
load7 0.96
load8 1.32
load9 1.08
load10 2
load11 1.22
load12 0.89
load13 2.05
load14 1.59
load15 2.31
load16 2.07
load17 1.55
load18 0.92
load19 1.64
load20 2.35
t1_1 -4.05
t2_1 -2.76
t3_1 -1.19
t4_1 1.91
t1_2 -3.63
t2_2 -2.13
t3_2 -0.89
t4_2 2.15
t1_3 -3.02
t2_3 -1.85
t3_3 0.23
t4_3 3.08
t1_4 -2.61
t2_4 -1.13
t3_4 0.19
t4_4 3.01
t1_5 -3.02
t2_5 -1.75
t3_5 -0.48
t4_5 2.1

```

t1_6 -3.89
t2_6 -2.66
t3_6 -0.69
t4_6 2.52
t1_7 -3.62
t2_7 -2.14
t3_7 -1.22
t4_7 1.29
t1_8 -4.28
t2_8 -3.02
t3_8 -0.65
t4_8 2.55
t1_9 -3.54
t2_9 -2.26
t3_9 0.53
t4_9 3.34
t1_10 -3.14
t2_10 -1.5
t3_10 -0.26
t4_10 3.36
t1_11 -1.7
t2_11 0.1
t3_11 1.31
t4_11 3.65
t1_12 -2.64
t2_12 -1.34
t3_12 -0.36
t4_12 2.17
t1_13 -4.2
t2_13 -2.44
t3_13 -0.31
t4_13 3.83
t1_14 -1.91
t2_14 -0.38
t3_14 0.97
t4_14 3.94
t1_15 -3.88
t2_15 -2.19
t3_15 -0.58
t4_15 3.9
t1_16 -3.93
t2_16 -2.24
t3_16 -0.81
t4_16 3.29
t1_17 -2.79
t2_17 -1.24
t3_17 0.16
t4_17 3.04
t1_18 -3.51
t2_18 -2.42
t3_18 -1.1
t4_18 1.54
t1_19 -2.3
t2_19 -0.82
t3_19 0.57
t4_19 3.44
t1_20 -4

```

t2_20 -2.12
t3_20 -0.14
t4_20 4.25
;
run;

proc sort data=true; by _NAME_; run;

%macro alltog;
data all;
merge
  %do cond=1 %to 4;
    %do rep=1 %to 100;
      mandytr&cond&rep
    %end;
  %end;
;
by _NAME_;
run;
%mend;

%alltog;

proc sort data=all; by _NAME_; run;

data final; merge all true; by _NAME_; run;
/*average estimate across replications for each condition*/

%let rep=100;
data all; set final;
  avg_cond1 = mean(of cond1rep1 - cond1rep&rep);
  avg_cond2 = mean(of cond2rep1 - cond2rep&rep);
  avg_cond3 = mean(of cond3rep1 - cond3rep&rep);
  avg_cond4 = mean(of cond4rep1 - cond4rep&rep);

  bias_cond1= (avg_cond1-true);
  bias_cond2= (avg_cond2-true);
  bias_cond3= (avg_cond3-true);
  bias_cond4= (avg_cond4-true);

  pctbias_cond1= (bias_cond1/true);
  pctbias_cond2= (bias_cond2/true);
  pctbias_cond3= (bias_cond3/true);
  pctbias_cond4= (bias_cond4/true);

samplingvar_cond1 = ((var(of cond1rep1 - cond1rep&rep))*(&rep-1))/&rep;
samplingvar_cond2 = ((var(of cond2rep1 - cond2rep&rep))*(&rep-1))/&rep;
samplingvar_cond3 = ((var(of cond3rep1 - cond3rep&rep))*(&rep-1))/&rep;
samplingvar_cond4 = ((var(of cond4rep1 - cond4rep&rep))*(&rep-1))/&rep;

MSE_cond1 = ((bias_cond1**2)+samplingvar_cond1);
MSE_cond2 = ((bias_cond2**2)+samplingvar_cond2);
MSE_cond3 = ((bias_cond3**2)+samplingvar_cond3);
MSE_cond4 = ((bias_cond4**2)+samplingvar_cond4);

RMSE_cond1 = (MSE_cond1**.5);

```

```

RMSE_cond2 = (MSE_cond2**.5);
RMSE_cond3 = (MSE_cond3**.5);
RMSE_cond4 = (MSE_cond4**.5);
run;

/**** Generating data for plots****/

data entropy; set all;
if _NAME_="entropy";
keep avg_cond1-avg_cond4;
run;

proc print data=entropy; run;

proc export data=entropy
  outfile="&path3\GRMRRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='entropy';
run;

data estclmean; set all;
if _NAME_="mixprop";
estpie_cond1 = (exp(avg_cond1)/(1+exp(avg_cond1)));
estpie_cond2 = (exp(avg_cond2)/(1+exp(avg_cond2)));
estpie_cond3 = (exp(avg_cond3)/(1+exp(avg_cond3)));
estpie_cond4 = (exp(avg_cond4)/(1+exp(avg_cond4)));
estlminuspie_cond1 = (1-estpie_cond1);
estlminuspie_cond2 = (1-estpie_cond2);
estlminuspie_cond3 = (1-estpie_cond3);
estlminuspie_cond4 = (1-estpie_cond4);
diff_cond1 = (.01-estpie_cond1);
diff_cond2 = (.05-estpie_cond2);
diff_cond3 = (.10-estpie_cond3);
diff_cond4 = (.20-estpie_cond4);
keep avg_cond1-avg_cond4 estpie_cond1 estpie_cond2 estpie_cond3
estpie_cond4
estlminuspie_cond1 estlminuspie_cond2 estlminuspie_cond3
estlminuspie_cond4
diff_cond1 diff_cond2 diff_cond3 diff_cond4;
run;

proc print data=estclmean; run;

proc export data=estclmean
  outfile="&path3\GRMRRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='estclmean';
run;

/*Loadings*/

data loading; set all;
if index(_NAME_, "load")=1;
if index(_NAME_, "loadSE")=0;
run;

```

```

proc print data=loading; run;

proc means data=loading;
var bias_cond1-bias_cond4 pctbias_cond1-pctbias_cond4 RMSE_cond1-
RMSE_cond4;
run;

data loadingBIAS; set loading;
KEEP _NAME_ TRUE bias_cond1-BIAS_COND4;
run;
PROC SORT DATA=LOADINGBIAS; BY TRUE; RUN;

data loadingRMSE; set loading;
KEEP _NAME_ TRUE RMSE_cond1-RMSE_cond4;
run;

PROC SORT DATA=LOADINGRMSE; BY TRUE; RUN;

data loadingAVG; set loading;
KEEP _NAME_ TRUE AVG_COND1-AVG_COND4;
run;

PROC SORT DATA=LOADINGAVG; BY TRUE; RUN;

proc export data=loadingBIAS
  outfile="&path3\GRMRRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='loading_BIAS';
run;

proc export data=loadingrmse
  outfile="&path3\GRMRRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='loading_RMSE';
run;

proc export data=loadingAVG
  outfile="&path3\GRMRRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='loading_AVG';
run;

/*Thresholds*/
data thresholds; set all;
if substr(_NAME_,1,1)="t";
run;

proc means data=thresholds;
var bias_cond1-bias_cond4 pctbias_cond1-pctbias_cond4 RMSE_cond1-
RMSE_cond4;
run;

data threshBIAS; set thresholds;

```

```

KEEP _NAME_ TRUE bias_cond1-BIAS_COND4;
run;
PROC SORT DATA=threshBIAS; BY TRUE; RUN;

data threshRMSE; set thresholds;
KEEP _NAME_ TRUE RMSE_cond1-RMSE_cond4;
run;

PROC SORT DATA=threshRMSE; BY TRUE; RUN;

data threshAVG; set thresholds;
KEEP _NAME_ TRUE AVG_COND1-AVG_COND4;
run;

PROC SORT DATA=threshAVG; BY TRUE; RUN;

proc export data=threshBIAS
  outfile="&path3\GRMRRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='thresh_BIAS';
run;

proc export data=threshrmse
  outfile="&path3\GRMRRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='thresh_RMSE';
run;

proc export data=threshAVG
  outfile="&path3\GRMRRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='thresh_AVG';
run;

/*Fit Indices*/
data fit; set all;
if _NAME_ in ("AIC", "BIC", "SSABIC", "LL") ;
keep _NAME_ avg_cond1-avg_cond4;
run;
proc print data=fit; run;

proc export data=fit
  outfile= "&path3\GRMRRMstudy1plots.xls"
  replace
  dbms=excel2002;
  sheet='fit';
run;

```


Appendix D

Datasets Used to Construct Plots

Table D1

Average Bias for the GRM – Loadings

Item	true	1%	5%	10%	20%
1	0.95	-0.008	-0.051	-0.102	-0.180
2	1.48	-0.016	-0.079	-0.163	-0.307
3	1.46	-0.022	-0.104	-0.186	-0.346
4	1.49	-0.008	-0.094	-0.177	-0.333
5	1.38	-0.015	-0.067	-0.133	-0.253
6	1.35	-0.017	-0.091	-0.176	-0.314
7	0.96	-0.012	-0.038	-0.068	-0.138
8	1.32	-0.021	-0.087	-0.183	-0.320
9	1.08	-0.015	-0.081	-0.159	-0.288
10	2	-0.027	-0.147	-0.276	-0.517
11	1.22	-0.020	-0.102	-0.184	-0.341
12	0.89	-0.013	-0.037	-0.074	-0.141
13	2.05	-0.044	-0.197	-0.357	-0.619
14	1.59	-0.028	-0.126	-0.244	-0.449
15	2.31	-0.054	-0.218	-0.411	-0.695
16	2.07	-0.042	-0.174	-0.323	-0.566
17	1.55	-0.019	-0.101	-0.189	-0.355
18	0.92	-0.013	-0.035	-0.070	-0.145
19	1.64	-0.021	-0.123	-0.224	-0.411
20	2.35	-0.046	-0.239	-0.436	-0.740

Table D2
Average RMSE for the GRM – Loadings

Item	true	1%	5%	10%	20%
1	0.95	0.032	0.061	0.108	0.182
2	1.48	0.043	0.088	0.167	0.309
3	1.46	0.044	0.110	0.189	0.348
4	1.49	0.043	0.102	0.181	0.335
5	1.38	0.042	0.076	0.139	0.255
6	1.35	0.041	0.097	0.179	0.316
7	0.96	0.034	0.051	0.074	0.142
8	1.32	0.041	0.094	0.186	0.321
9	1.08	0.037	0.089	0.162	0.289
10	2	0.058	0.154	0.280	0.518
11	1.22	0.044	0.106	0.187	0.342
12	0.89	0.033	0.047	0.080	0.144
13	2.05	0.065	0.202	0.359	0.620
14	1.59	0.048	0.132	0.246	0.450
15	2.31	0.078	0.222	0.413	0.696
16	2.07	0.066	0.180	0.327	0.568
17	1.55	0.045	0.107	0.192	0.357
18	0.92	0.037	0.047	0.078	0.149
19	1.64	0.043	0.128	0.228	0.412
20	2.35	0.077	0.244	0.439	0.742

Table D3
Average Bias for the GRM – Thresholds

Threshold	Item	true	1%	5%	10%	20%
1	1	-4.05	0.130	0.361	0.643	1.061
	2	-3.63	0.084	0.243	0.429	0.769
	3	-3.02	0.073	0.169	0.313	0.557
	4	-2.61	0.064	0.134	0.220	0.398
	5	-3.02	0.065	0.145	0.270	0.496
	6	-3.89	0.101	0.304	0.535	0.939
	7	-3.62	0.092	0.254	0.438	0.775
	8	-4.28	0.127	0.393	0.683	1.173
	9	-3.54	0.089	0.244	0.463	0.800
	10	-3.14	0.077	0.203	0.367	0.645
	11	-1.7	0.046	0.069	0.105	0.175
	12	-2.64	0.059	0.106	0.190	0.347
	13	-4.2	0.117	0.375	0.681	1.158
	14	-1.91	0.050	0.092	0.152	0.262
	15	-3.88	0.117	0.331	0.603	1.014
	16	-3.93	0.100	0.310	0.558	0.969
	17	-2.79	0.069	0.153	0.258	0.469
	18	-3.51	0.089	0.222	0.404	0.724
	19	-2.3	0.063	0.119	0.208	0.349
	20	-4	0.121	0.359	0.661	1.089
2	1	-2.76	0.088	0.248	0.457	0.792
	2	-2.13	0.062	0.150	0.284	0.524
	3	-1.85	0.052	0.140	0.255	0.464
	4	-1.13	0.046	0.075	0.126	0.230
	5	-1.75	0.047	0.114	0.209	0.383
	6	-2.66	0.067	0.234	0.423	0.761
	7	-2.14	0.056	0.145	0.263	0.494
	8	-3.02	0.090	0.289	0.530	0.936
	9	-2.26	0.064	0.182	0.341	0.605
	10	-1.5	0.054	0.128	0.225	0.390
	11	0.1	0.038	0.042	0.054	0.096
	12	-1.34	0.046	0.082	0.129	0.241
	13	-2.44	0.079	0.246	0.440	0.766
	14	-0.38	0.039	0.043	0.051	0.060
	15	-2.19	0.087	0.220	0.395	0.677
	16	-2.24	0.077	0.214	0.375	0.656
	17	-1.24	0.047	0.090	0.148	0.270
	18	-2.42	0.066	0.184	0.336	0.614
	19	-0.82	0.039	0.067	0.101	0.170
	20	-2.12	0.083	0.227	0.399	0.677

Table D3 (continued)
Average Bias for the GRM – Thresholds

Threshold	Item	true	1%	5%	10%	20%
3	1	-1.19	0.044	0.118	0.223	0.410
	2	-0.89	0.050	0.102	0.195	0.359
	3	0.23	0.040	0.050	0.052	0.073
	4	0.19	0.042	0.047	0.060	0.090
	5	-0.48	0.040	0.073	0.127	0.230
	6	-0.69	0.041	0.095	0.158	0.298
	7	-1.22	0.047	0.118	0.210	0.410
	8	-0.65	0.036	0.090	0.153	0.287
	9	0.53	0.029	0.039	0.036	0.045
	10	-0.26	0.045	0.088	0.138	0.238
	11	1.31	0.041	0.078	0.133	0.242
	12	-0.36	0.036	0.059	0.095	0.182
	13	-0.31	0.052	0.089	0.144	0.252
	14	0.97	0.047	0.052	0.086	0.152
	15	-0.58	0.056	0.126	0.208	0.358
	16	-0.81	0.055	0.134	0.233	0.402
	17	0.16	0.038	0.045	0.058	0.085
	18	-1.1	0.038	0.101	0.195	0.373
	19	0.57	0.037	0.038	0.043	0.045
	20	-0.14	0.055	0.090	0.129	0.219
4	1	1.91	0.042	0.055	0.089	0.138
	2	2.15	0.055	0.081	0.131	0.232
	3	3.08	0.070	0.167	0.300	0.532
	4	3.01	0.064	0.149	0.276	0.504
	5	2.1	0.051	0.071	0.121	0.191
	6	2.52	0.056	0.099	0.200	0.343
	7	1.29	0.038	0.036	0.044	0.039
	8	2.55	0.056	0.116	0.203	0.363
	9	3.34	0.081	0.207	0.376	0.675
	10	3.36	0.077	0.203	0.370	0.675
	11	3.65	0.083	0.263	0.466	0.832
	12	2.17	0.044	0.064	0.105	0.195
	13	3.83	0.104	0.299	0.537	0.931
	14	3.94	0.104	0.297	0.537	0.957
	15	3.9	0.117	0.314	0.581	0.967
	16	3.29	0.087	0.227	0.407	0.677
	17	3.04	0.074	0.156	0.284	0.515
	18	1.54	0.041	0.040	0.043	0.053
	19	3.44	0.079	0.204	0.380	0.692
	20	4.25	0.122	0.381	0.686	1.144

Table D4
Average Bias for the RRM-GRM – Loadings

Item	true	1%	5%	10%	20%
1	0.95	-0.008	-0.051	-0.102	-0.180
2	1.48	-0.016	-0.079	-0.163	-0.307
3	1.46	-0.022	-0.104	-0.186	-0.346
4	1.49	-0.008	-0.094	-0.177	-0.333
5	1.38	-0.015	-0.067	-0.133	-0.253
6	1.35	-0.017	-0.091	-0.176	-0.314
7	0.96	-0.012	-0.038	-0.068	-0.138
8	1.32	-0.021	-0.087	-0.183	-0.320
9	1.08	-0.015	-0.081	-0.159	-0.288
10	2	-0.027	-0.147	-0.276	-0.517
11	1.22	-0.020	-0.102	-0.184	-0.341
12	0.89	-0.013	-0.037	-0.074	-0.141
13	2.05	-0.044	-0.197	-0.357	-0.619
14	1.59	-0.028	-0.126	-0.244	-0.449
15	2.31	-0.054	-0.218	-0.411	-0.695
16	2.07	-0.042	-0.174	-0.323	-0.566
17	1.55	-0.019	-0.101	-0.189	-0.355
18	0.92	-0.013	-0.035	-0.070	-0.145
19	1.64	-0.021	-0.123	-0.224	-0.411
20	2.35	-0.046	-0.239	-0.436	-0.740

Table D5
Average RMSE for the RRM-GRM – Loadings

Item	true	1%	5%	10%	20%
1	0.95	0.032	0.038	0.039	0.035
2	1.48	0.041	0.043	0.042	0.045
3	1.46	0.039	0.039	0.040	0.042
4	1.49	0.043	0.044	0.040	0.043
5	1.38	0.039	0.041	0.041	0.043
6	1.35	0.038	0.036	0.037	0.046
7	0.96	0.033	0.036	0.032	0.038
8	1.32	0.038	0.040	0.043	0.037
9	1.08	0.034	0.038	0.035	0.045
10	2	0.051	0.049	0.051	0.054
11	1.22	0.039	0.031	0.038	0.037
12	0.89	0.031	0.030	0.033	0.039
13	2.05	0.049	0.053	0.051	0.065
14	1.59	0.039	0.042	0.040	0.047
15	2.31	0.058	0.050	0.062	0.057
16	2.07	0.051	0.049	0.054	0.061
17	1.55	0.042	0.040	0.039	0.049
18	0.92	0.035	0.031	0.040	0.042
19	1.64	0.038	0.041	0.051	0.045
20	2.35	0.063	0.057	0.060	0.066

Table D6
Average Bias for the RRM- GRM – Thresholds

Threshold	Item	true	1%	5%	10%	20%
1	1	-4.050	0.015	0.002	0.023	0.007
	2	-3.630	0.007	0.001	0.000	0.015
	3	-3.020	0.005	-0.001	-0.005	0.003
	4	-2.610	0.000	0.010	-0.011	-0.006
	5	-3.020	-0.009	-0.009	-0.006	0.001
	6	-3.890	0.000	0.007	-0.006	-0.002
	7	-3.620	0.002	0.011	-0.004	0.000
	8	-4.280	-0.001	0.002	-0.020	0.004
	9	-3.540	0.003	-0.008	0.003	-0.001
	10	-3.140	-0.006	0.001	0.003	0.000
	11	-1.700	0.004	0.006	0.000	0.005
	12	-2.640	0.002	-0.007	-0.003	-0.013
	13	-4.200	-0.005	-0.004	-0.007	0.005
	14	-1.910	-0.002	0.000	0.000	0.010
	15	-3.880	0.008	-0.002	0.011	-0.002
	16	-3.930	0.003	-0.005	-0.007	-0.024
	17	-2.790	0.003	0.012	-0.002	0.002
	18	-3.510	-0.004	-0.011	-0.016	-0.001
	19	-2.300	0.007	0.009	0.014	0.004
	20	-4.000	0.005	-0.001	0.015	-0.005
2	1	-2.760	0.012	0.008	0.011	0.002
	2	-2.130	-0.002	-0.007	-0.004	0.005
	3	-1.850	0.002	0.000	-0.004	0.003
	4	-1.130	-0.001	-0.003	-0.010	-0.003
	5	-1.750	-0.003	-0.002	-0.002	-0.002
	6	-2.660	-0.006	0.005	0.001	0.001
	7	-2.140	0.000	0.001	-0.001	0.005
	8	-3.020	-0.006	-0.006	-0.011	-0.009
	9	-2.260	0.003	-0.001	0.007	0.000
	10	-1.500	-0.005	0.002	0.003	0.003
	11	0.100	-0.004	-0.002	-0.002	0.002
	12	-1.340	0.008	0.007	0.001	0.001
	13	-2.440	0.001	0.004	0.000	0.006
	14	-0.380	-0.003	0.002	-0.001	0.001
	15	-2.190	0.000	-0.002	-0.004	-0.011
	16	-2.240	0.010	0.010	0.003	-0.005
	17	-1.240	0.001	0.006	-0.002	-0.001
	18	-2.420	0.002	0.001	-0.004	0.000
	19	-0.820	-0.006	0.002	-0.002	-0.002
	20	-2.120	0.009	0.005	-0.001	0.003

Table D6 (continued)

Average Bias for the RRM- GRM – Thresholds

Threshold	Item	true	1%	5%	10%	20%
3	1	-1.190	0.000	0.003	0.007	0.002
	2	-0.890	0.006	-0.005	0.002	0.004
	3	0.230	0.012	0.011	0.006	0.003
	4	0.190	0.006	0.004	0.001	0.005
	5	-0.480	0.003	-0.001	-0.002	-0.010
	6	-0.690	-0.003	0.005	-0.002	0.001
	7	-1.220	0.003	0.004	-0.004	0.006
	8	-0.650	-0.001	0.005	-0.001	0.004
	9	0.530	0.000	-0.005	0.005	-0.005
	10	-0.260	0.000	0.009	0.003	0.007
	11	1.310	-0.005	-0.003	0.000	0.006
	12	-0.360	-0.002	-0.002	-0.005	-0.004
	13	-0.310	0.008	0.003	-0.003	0.006
	14	0.970	-0.006	0.003	-0.002	0.000
	15	-0.580	0.005	0.003	-0.003	-0.001
	16	-0.810	0.003	0.006	0.001	-0.007
	17	0.160	-0.002	-0.002	-0.007	-0.013
	18	-1.100	-0.006	-0.005	-0.001	-0.001
	19	0.570	0.001	0.003	0.002	-0.008
	20	-0.140	-0.003	0.007	-0.005	0.003
4	1	1.910	0.006	0.007	-0.004	0.002
	2	2.150	-0.007	0.000	0.003	-0.005
	3	3.080	-0.001	-0.005	0.006	0.002
	4	3.010	0.021	0.007	0.005	0.005
	5	2.100	0.006	0.001	-0.010	-0.003
	6	2.520	0.003	0.015	-0.003	0.005
	7	1.290	-0.005	-0.001	-0.011	-0.008
	8	2.550	-0.006	-0.001	0.001	0.004
	9	3.340	-0.002	0.001	0.002	-0.002
	10	3.360	0.010	0.012	0.019	0.002
	11	3.650	-0.004	-0.004	0.012	0.001
	12	2.170	-0.001	0.004	0.003	-0.004
	13	3.830	0.005	0.014	0.011	0.009
	14	3.940	0.001	0.007	0.018	0.010
	15	3.900	-0.003	0.012	-0.010	0.016
	16	3.290	-0.011	-0.004	-0.015	0.016
	17	3.040	-0.006	-0.001	0.003	0.001
	18	1.540	0.000	0.001	0.000	-0.002
	19	3.440	0.006	0.002	0.005	-0.002
	20	4.250	0.005	0.015	0.005	0.018

References

- Abedi, J., & O'Neil, H.F. (2005). Assessment of noncognitive influences on learning. *Educational Assessment*, 10(3), 147–151.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csake (Eds.). *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.
- American College Testing (2013). *Cognitive Skills*. Retrieved from <http://www.act.org/workkeys/briefs/files/CognitiveSkills.pdf>
- American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Asparouhov, T. & Muthén, B. (2014a). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3), 329-341.
- Asparouhov, T. & Muthén, B. (2014b). Auxiliary variables in mixture modeling: Using the BCH method in Mplus to estimate a distal outcome model and an arbitrary secondary model. *Mplus Web Notes*, 21(2).
- Bovaird, J. A. (2002). New applications in testing: Using response time to increase the construct validity of a latent trait estimate. Doctoral dissertation, University of Kansas. Dissertation Abstracts International, 64, 998.
- Baumgartner, H., & Steenkamp, J. E. M., (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38, 143-156.

- Cao, J. & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, 73(2), 209-230. <http://dx.doi.org/10.1007/s11336-007-9045-9>
- Cloud, J., & Vaughn, G. (1970). Using balanced scales to control acquiescence. *Sociometry*, 33, 193-202.
- Cohen A.S., & Bolt, D.M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42, 133-148.
- Coleman, C. M. (2013). Effects of negative keying and wording in attitude measures: A mixed-methods study. (Doctoral dissertation). ProQuest. (UMI: 3560664).
- Cronbach, L. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6, 475-494.
- DeAyala, R. J. Kim, S-H., Stapleton, L. M. & Dayton, C. M. Differential item functioning: a mixture distribution conceptualization. *International Journal of Testing*, 2, 243-276, 2003.
- DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, 12, 23- 45. doi: 10.1080/10627190709336946
- Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, 66, 643–656.
- France, M. K. (2011). Introducing the unified measure of university mattering: Instrument development and evidence of the structural integrity of scores for transfer and native students. (Doctoral dissertation). ProQuest. (UMI: 3453711).

- Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, 73(1), 65-87.
<http://dx.doi.org/10.1007/s11336-007-9031-2>
- Gove, W. R., & Geerken, M. R. (1977). Responde bias in surveys of mental health: An empirical investigation. *American Journal of Sociology*, 82(6), 1289-1317.
<http://www.jstor.org/stable/2777936> .
- Hambleton, R. K. (2012). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (89-116). Mahwah, NJ: Lawrence Erlbaum Publishers.
- Heckman, J. J., & Rubinstein, Y. (2001). The Importance of noncognitive skills: Lessons from the GED testing program. *The Benefits of Skill*, 91(2), 145-151.
- Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 202–226.
- Hoffman, L. (2014). *Latent trait measurement models for other (not binary) responses* [PowerPoint slides]. Retrieved from <http://siri.uvm.edu/ppt/40hrenv/index.html>
- Hoogland, J.J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, 26, 329–367.
- Jin, K. & Wang, W. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement*, 51(2), 178-200.

- Kamata, A. & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, 15, 136-153.
[http://dx.doi.org/ 10.1080/10705510701758406](http://dx.doi.org/10.1080/10705510701758406)
- Kane, M. (2012). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (53-88). Mahwah, NJ: Lawrence Erlbaum Publishers.
- Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Psychological Measurement*, 67, 606-619. doi: 10.1177/0013164406294779
- Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213–236.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50, 537-567.
- Krosnick, J. A. (2011). Experiments for evaluating survey questions . In K. Miller, J. Madans, G. Willis, & A. Maitland (Eds.), *Question evaluation methods*. New York, NY: Wiley.
- Lau, A. (2009). Using a mixture IRT model to improve parameter estimates when some examinees are amotivated. (Doctoral dissertation). ProQuest. (UMI: 3366561).
- Lau, A., & Pastor, D. (2010). Application of a mixture IRT model to improve parameter estimates when some examinees are amotivated. Unpublished manuscript.
- Lautenschlager, G. J., Meade, A. W., & Kim, S.-H. (2006, April). Cautions regarding sample characteristics when using the graded response model. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

- Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, 41 (9), 352-362. DOI: 10.3102/0013189X12459679
- Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88, 767-778.
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19-40.
- Marsh, K. R. (2013). The effects of item and respondent characteristics on midpoint response option endorsement: A mixed-methods study. (Doctoral dissertation). ProQuest. (UMI: 3560681).
- McPherson, J., & Mohr, P. (2005). The role of item extremity in the emergence of keying-related factors: An exploration with the Life Orientation Test. *Psychological Methods*, 10, 120-131.
- Meyer, J. P. (2010). A mixture rasch model with item response time componets. *Applied Psychological Measurement*, 34(7), 521-538. <http://dx.doi.org/10.1177/0146621609355451>
- Mislevy, R., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195-215.
- Moustaki, I. and Knott, M. (2014) Latent variable models that account for atypical responses. *Journal of the Royal Statistical Society*, C(63), 343-360.
- Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

- Pastor, D.A. & Gagné, P. (2013) Mean and covariance structure mixture models. In Hancock, G.R. & Mueller, R.O. (Ed.), *Structural equation modeling: A second course* (343-393). Charlotte, NC: Information Age Publishing.
- Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993).) Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement*, 53, 801-813. <http://dx.doi.org/10.1177/0013164493053003024>
- Rios, J. A., Liu, O. L., Bridgeman, B. (2014). Identifying unmotivated examinees on student learning outcomes assessment: A comparison of two approaches. *The National Council on Measurement in Education Annual Meeting*. Philadelphia, PA.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.
- Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED383742)
- Schnipke, D.L. & Scrams, D.J. (1997). Modeling item response times with a two-stage mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34(3), 213-232.

- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Schmitt, N., Billington, A., Keeney, J., Reeder, M., Pleskac, T. J., Sinha, R. & Zorzie, M. (2011). *Development and validation of measures of noncognitive college student potential*. The College Board: Research Report, 1.
- Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333–343.
- Spellings, M. (2006). *A test of leadership: Charting the future of U.S. higher education*. A report of the commission appointed by the Secretary of Education. Washington DC: U.S. Department of Education.
- Steedle, J. T. (2014). Motivation filtering on a multi-institution assessment of general college outcomes. *Applied Measurement in Education*, 21(1), 58-76.
[http://dx.doi.org/ 10.1080/08957347.2013.853072](http://dx.doi.org/10.1080/08957347.2013.853072)
- Sundre, D. L. & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update*, 14 (1), 8-9.
- Sundre, D. L. & Wise, S. L. (2003). 'Motivation filtering': *An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the National Council on Measurement in Education Annual Conference, Chicago, Illinois.
- Suskie, L. (2009). *Assessing student learning: A common sense guide* (2nd ed.). San Francisco, CA: Jossey-Bass.
- Swanson, M. (2013). An introduction to and application of amotivation modeling using a 2PL two-class IRT mixture model. Unpublished manuscript.

- Swanson, M. & Pastor, D. (2014). *Detecting amotivated examinees in low-stakes testing using a 2PL IRT mixture model*. Presented at the Annual Symposium in Research and Practice. JMU Department of Graduate Psychology. Harrisonburg, VA.
- 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*, 162-188. doi: 10.1080/08957347.2011.555217
- Thelk, A., Sundre, D.L., Horst, J. S., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale (SOS) to make valid inferences about student performance. *Journal of General Education, 58*, 131-151.
- Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Greenwich, CT: Information Age Publishing, Inc.
- Tong, Y. & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227-253.
- Tourangeau, R. (1984). Cognitive science and survey methods. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington, DC: National Academy Press.
- Vermunt, J. K. (2010). Latent class modeling with covariate: Two improved three-step approaches. *Political Analysis, 18*, 450-469.

- von Davier, M., & Rost, J. (1995). Polytomous mixed rasch models. In G. H. Fisher & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371-379). New York: Springer.
- Yamamoto, K. (1989) *Hybrid model of IRT and latent class models*. (ETS Research Rep. No. RR-89-41). Princeton, NJ: Educational Testing Service.
- Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (TOEFL Tech. Rep. No. TR-10). Princeton, NJ: Educational Testing Service.
- Yang, C.C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistical & Data Analysis*, 50, 1090–1104.
- Yang, X. (2007). Methods of identifying individual guessers from item response data. *Educational and Psychological Measurement*, 67(5), 745-764.
- Whittaker T.A., Fitzpatrick S. J., Williams, N. J. & Dodd B. G., (2003). IRTGEN: A SAS Macro Program to Generate Known Trait Scores and Item Responses for Commonly Used Item Response Theory Models. *Applied Psychological Measurement*, 27, 299-300. doi: 10.1177/0146621603027004005
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1-17.
- Wise, S. L. & DeMars, C. E. (2006). An application of item response time: The effort-moderated model. *Journal of Educational Measurement*, 43, 19-38.
- Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, 15, 27–41. doi: 10.1080/10627191003673216

- Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *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.
- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18, 163-183.
- Wise, V. L., Wise, S. L., & Bholra, D.S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, 11, 65-83.
- Zerpa, C., Hachey, K., van Barnfield, C., & Simon, M. (2011). Modeling student motivation and students' ability estimates from a large-scale assessment of mathematics. *SAGE Open*, 1-9. DOI: 10.1177/2158244011421803