5-2022

Aberrant Responding with Underlying Dominance and Unfolding Response Processes: Examining Model Fit and Performance of Person-Fit Statistics

Jennifer A. Reimers
University of Arkansas, Fayetteville

Aberrant Responding with Underlying Dominance and Unfolding Response Processes: Examining Model Fit and Performance of Person-Fit Statistics

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Educational Statistics and Research Methods

by

Jennifer A. Reimers
Oklahoma State University
Bachelor of Arts in Spanish, 2009
University of Arkansas
Master of Education in Elementary Education, 2011

May 2022
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

_____
Ronna C. Turner, Ph.D.
Dissertation Chair

_____
Wen-Juo Lo, Ph.D.
Committee Member

_____
Jorge Tendeiro, Ph.D.
Ex-Officio Committee Member

_____
Elizabeth Keiffer, Ph.D.
Committee Member

ABSTRACT

Researchers have recognized that respondents may not answer items in a way that accurately reflects their attitude or trait level being measured. The resulting response data that deviates from what would be expected has been shown to have significant effects on the psychometric properties of a scale and analytical results. However, many studies that have investigated the detection of aberrant data and its effects have done so using dominance item response theory (IRT) models. It is unknown whether the impacts of aberrant data and the methodology used to identify aberrant responding when using dominance IRT models apply similarly when scales fit an unfolding IRT model. This dissertation is aimed at contributing to the literature with unfolding IRT models (specifically the generalized graded unfolding model [GGUM]) in three main ways: 1) by providing insight on GGUM model-data fit when various types of aberrant data are systematically entered, 2) by investigating how nonparametric person-fit statistics (H^T, $U3^P$, G_N^P, and G^P) perform under the unfolding framework of GGUM compared to the dominance framework of the generalized partial credit model (GPCM), and 3) by examining how the performance of parametric person-fit statistics ($l_{z(p)}$ and $l_{z(p)}^*$) is impacted by misspecifying a dominance model (GPCM) to unfolding model data (GGUM) and conversely, GGUM to GPCM data. As unfolding models have many advantages and are becoming more widely used by researchers, the rise of questions regarding the effects of data quality and the performance of person-fit statistics under this context is expected. It is essential to gain a better understanding on how underlying response processes affect data-model fit, and how effectively different types of aberrant data are identified using multiple data model frameworks.

The dissertation is organized into three studies based on a simulation design that investigates the impacts of type of aberrant responding, proportion of aberrant responders in the sample, proportion of aberrant responses within a response vector, test length, model-data generation and application on model-fit and person-fit statistic performance. In the first study, the impact of aberrant data on model fit for GGUM and GPCM data is investigated and found to be severe in some cases. However, the GGUM was found to effectively fit both dominance and unfolding data, even with 10% aberrant data in many cases. It is suggested that researchers carefully examine data quality before making conclusions about model-data fit or misfit. The second study investigates the application of popular nonparametric person-fit statistics used with dominance data to data that fit an unfolding model. Given their poor performance, further research is recommended to identify or develop person-fit statistics effective for different types of aberrant behavior exhibited in ideal point response data. Study 3 compares type I error and power rates for parametric person-fit statistics with GGUM and GPCM data that are correctly and incorrectly specified, compared to nonparametric person-fit performance. No person-fit statistic was robust against model misspecification when GPCM was fit to GGUM data. Conversely, results were comparable for GPCM data, regardless of fitting the GPCM or GGUM to the data.

ACKNOWLEDGEMENTS

Throughout the process of completing this dissertation I have received an immense amount of support. First, I would like to thank my family, which doubled in size throughout the project. Having two children while completing this dissertation made for some interesting stories and very long nights, but of course, I have no regrets. I thank my husband for all his support, and I congratulate him for surviving my Ph.D. endeavor.

Second, I would like to thank my academic advisor, mentor, and friend, Dr. Ronna Turner, whose insightful feedback challenged me to learn to conduct research on a higher level. As I progress in my professional and academic career, I strive to make her proud and will be forever thankful for the endless support she has given me throughout this journey.

I would also like to thank several other professors from whom I learned more than I ever imagined I would in the program. First, I want to recognize the amazing support I have received from Dr. Wen-Juo Lo. He has mentored me throughout my classes and various graduate student roles including my first assistantship in the nursing department when I thought I was in over my head, but ended up learning a lot with his help. I also want to thank Dr. Beth Keiffer, who has mentored me throughout the Business Analytics program and grant work. I learned so much in Dr. Keiffer's classes and she is a large reason I was offered my current position. I would also like to thank Dr. Xinya Liang, who helped me in multiple graduate classes with her well-prepared course layouts, facilitating teaching style, and ability to communicate the material.

Last, but far from the least, I would also like to thank Dr. Jorge Tendeiro, whose expertise was invaluable in conducting this research. His genuine efforts to share knowledge via feedback and discussion in the spirit and pursuit of advancing research in the field was truly inspiring.

DEDICATIONS

This work is dedicated to my two miracles, Ava Grace and Dawson Dane. God willing, this work will be one of many things in my life I plan to dedicate to you both.

To my husband who is my forever best friend.

And to my big brother and parents who raised me to believe that anything is possible if you set your mind to it.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

"Between stimulus and response there is a space. In that space, is our power to choose our response."

– Viktor E. Frankl

Underlying response processes are often overlooked when researchers apply statistical models to their response data. As it suggests in the name, item response theory (IRT) uses both items and responses to estimate latent trait levels for respondents. The majority of psychological measurement models in IRT literature involves dominance IRT models (Harris-Watson et al., 2020; Tay & Ng, 2018; e.g., 1-, 2-, 3-parameter logistic models for dichotomous data and the graded response model or the generalized partial credit model for polytomous data) to obtain this goal, which assume the underlying response process is cumulative and the probability of a response is monotonically increasing along with the latent trait level that is being assessed. However, many researchers utilize dominance IRT models without giving much thought to alternative underlying response processes that may be used when examinees are answering the items. Many complex, non-cognitive constructs may be difficult to measure with strictly dominance IRT (Chernyshenko et al., 2001; Drasgow et al., 2010; Meijer & Baneke, 2004; Stark et al., 2006). That is, items may elicit an ideal point response process whereby respondents choose to endorse (or not) the item according to how well the item-level matches their trait-level on the underlying latent construct being assessed. Tay and Ng (2018) give an example of this with a questionnaire meant to assess people's political views, where a political moderate would

be more likely to strongly endorse a more moderate political item than more extreme items at either end of the construct continuum. Conversely, a respondent may disagree with the moderate political item for one of two reasons: either their political views are located below or above the item on the underlying continuum. Thus, the moderate item may demonstrate a non-monotonic item response function, violating an assumption of dominance IRT.

Researchers have clearly warned of the potential adverse consequences for mis-specifying a dominance IRT model for ideal point response data (Chernyshenko et al., 2001; Drasgow et al., 2010; Meijer & Baneke, 2004; Stark et al., 2006). As a result, the use of unfolding models that reflect an ideal point response process has become somewhat more widespread in the last few decades (e.g., Javaras & Ripley, 2007; Joo et al., 2017, 2019; Kartal & Dirlik, 2021; Santos et al., 2021; Sgammato, 2009; Tendeiro, 2017; Weekers & Meijer, 2008; Weiss et al., 2018; Williams, 2015; Zampetakis, 2010). Even still, the use of unfolding models in practice is sparse in comparison to the use of dominance IRT models.

When investigating data for appropriate model fit, such as comparing dominance and ideal point response models, a researcher may also need to consider the quality of the data provided by the participants. If a dataset contains response strings from participants who do not use sufficient effort in answering the questions, it can impact the selection of an appropriate model for the valid data. The use of person-fit statistics to detect insufficient effort or aberrant responders has gained popularity in the last two decades (e.g., Cizek & Wollack, 2016; Conijn et al., 2014; Niessen et al., 2016; Sinharay, 2021; Turner, 2018). However, there is a substantial gap in the literature merging these two concepts (aberrant data and unfolding models). Many studies have investigated the performance of person-fit statistics in dominance IRT settings, but only one study has applied person-fit statistics in an unfolding model context (Tendeiro, 2017).

It is still uncertain how many popular person-fit statistics perform when an underlying ideal point response process is present. Similarly, it is unclear how aberrant data affects model fit for unfolding versus dominance IRT models.

Purpose of the Study

In three studies, I explore relationships between model fit, presence of aberrant responding, and detection of aberrant responding under dominance and unfolding model contexts. The investigation begins with Study 1, where the differential impacts of aberrant responding on model fit for two parametric IRT models based on different response processes are examined. Determining how model fit may be affected by aberrant responding to items reflecting a dominance or ideal point response process in Study 1 will lay a foundation for Studies 2 and 3. In Study 2, the performance of various nonparametric person-fit statistics is examined under the dominance generalized partial credit model as well as the generalized graded unfolding model. With information about how aberrant responding affects model fit from Study 1, and with the results from Study 2 informing on how nonparametric person-fit statistics perform under both dominance and unfolding model contexts, Study 3 extends Studies 1 and 2 by examining how the parametric person-fit statistics, $l_{z(p)}$ and $l^*_{z(p)}$, perform in both dominance and unfolding settings and how misspecification of the underlying response process may affect detection of aberrant responding. The overarching goal for this dissertation is to fill the gaps in current literature regarding aberrant responding and person-fit analyses with unfolding models and its relation with dominance models.

Study 1: *An Investigation of the Effects of Aberrant Responding on Model-Fit for Unfolding and Dominance Response Models*

Aberrant responding has been an issue recognized by researchers for decades, with recent studies focusing on the detrimental impacts on various survey characteristics (DeSimone et al., 2018). Previous research has demonstrated that aberrant responding can influence the psychometric properties of a scale (e.g., reliability estimates, factor structure, interitem correlations) and alter the results of statistical analyses (DeSimone et al., 2018; Turner, 2018). As item characteristics have been shown to change with the presence of aberrant responding, models that fit different types of items such as dominance and unfolding models, may fit the data differently when aberrant responding occurs. Liu and Wang (2019) showed in an empirical study that parameters estimated by the General Unfolding Model (GUM) may be biased when response styles are ignored. However, no studies to date have examined how other types of aberrant responding may affect how an unfolding model fits the data, compared to a dominance model. In reality, researchers will not know if model-data misfit is due to the aforementioned changes resulting from the presence of aberrant responding, or if it is due to true model misspecification. One of the advantages of using a simulation design for the current study is the control over these factors. Aberrant responding and model misspecification (based on the underlying response process) will be manipulated to provide insight on the confounding effects that may arise with real data. For example, it could be that aberrant responding affects model fit more severely for certain data types and conditions, which would then point the researcher to investigate data quality before making conclusions about model misfit.

Researchers have emphasized the importance of assessing model-data fit before interpreting results (Chernyshenko et al., 2001). Study 1 focuses on the differential impacts of

aberrant responding on model-data fit using unfolding and dominance models, GGUM and GPCM. This study will contribute to growing literature on unfolding versus dominance models as well as the effects of data quality. For the purpose of this study, response styles that impact participant response strings, resulting in answers that would be different than expected based on latent trait, are included as forms of aberrant responding. Thus, four types of aberrant responding will be investigated in the study, including two common insufficient effort response types (longstrings and random responding), and two types of response styles (midpoint responding and extreme responding). Other factors in the study include the number of items, number of response categories, proportion of aberrant respondees in the sample and proportion of aberrant responses in a response vector. If the types and proportion of aberrant responding have different impacts on model-fit, implications for researchers may include the importance of the approach used for identifying aberrant responding when creating a data screening plan.

Additionally, the cross-fit of the GGUM model to GPCM data and the GPCM model to GGUM data is compared in order to assess the flexibility of both models with different types of data. Previous research has suggested more flexibility exists with GGUM (i.e., the GGUM fits dominance data reasonably well; Stark, 2006) but this has not been investigated with aberrant data as a factor. It would be helpful to know if either of the two models is flexible in fitting data types, especially in situations where there is aberrant responding (of different types and proportions).

If model-data misfit is shown to increase the estimated parameter bias in Study 1, then the detection of the aberrant responses may in turn be affected. Tendeiro (2017) found that bias of estimated parameters affected the detection rates of extreme responding more so than for midpoint responding. Thus, if aberrant responding affects the model-data fit, and the model-data

fit affects the detection of aberrant responding, the researcher faces a difficult decision while attempting to maintain the integrity of the data quality. More research in this area is warranted to help guide researchers in data cleaning when fitting unfolding or dominance models.

Study 2: ***Performance of Nonparametric Person-Fit Statistics with Unfolding versus Dominance Response Models***

The primary purpose of Study 2 is to examine the performance of several polytomous person-fit statistics under an unfolding model context. Although a large body of literature exists covering the performance of person-fit statistics, all studies but one assume an underlying dominance response process. However, in the last fifteen years, an ideal point response process has been recognized as more appropriate than a dominance response process for several types of non-cognitive data such as assessment for creativity using the Gough's Creative Personality Scale (Zampetakis, 2010), conscientiousness (Carter et al., 2014), personality inventory of self-judgement on the order-facet (a feature of conscientiousness; Chernyshenko et al., 2007; Weekers & Meijer, 2008), 16 personality factor subscales (Stark et al., 2006), and job satisfaction (Carter & Dalal, 2010). Because the ordering of persons based on latent trait scores may be severely affected by the underlying item response process (Stark et al., 2006), it is reasonable to question the applicability of the findings from previous person-fit studies using dominance IRT models to an ideal point model context.

Tendeiro (2017) studied the $l_{z(p)}$ and $l_{z(p)}^{*}$ person-fit statistic under the generalized graded unfolding model (GGUM) which assumes an ideal point response process. The study suggested that the detection rates for midpoint response style patterns using the $l_{z(p)}^{*}$ person-fit statistic were promising in many conditions. Detection rates for extreme response style patterns were lower, but the author believes this may be due to specific data generation conditions in the

study. As this was the first person-fit study using an unfolding model, many questions still remain regarding how person-fit statistics perform assuming an underlying ideal point response process.

Study 2 extends the work of Tendeiro (2017) in two fundamental ways. First, several other person-fit statistics' performance under the unfolding framework are investigated. While $l_{z(p)}$ and $l^*_{z(p)}$ are popular parametric person-fit statistics, nonparametric person-fit statistics have been found to perform just as well, if not better than parametric person-fit statistics under various conditions when applied to data using dominance response models (Emons, 2008; Karabatsos, 2003; Tendeiro & Meijer, 2014). However, their application to data that more appropriately fit an unfolding model has not been studied. Study 2 includes four nonparametric person fit statistics (H^T, $U3^P$, G_N^P, and G^P) that have been shown to perform relatively well in comparison with other person fit statistics (Emons, 2008; Karabatsos, 2003; Tendeiro & Meijer, 2014; Turner, 2018). Second, in addition to midpoint and extreme responding, aberrant responding due to longstring responses and random responding is also considered in Study 2. These types of aberrant responding have been a concern in a variety of contexts such as employee surveys, customer surveys, training evaluations, personality inventories and attitudinal surveys.

The goal of Study 2 is to compare person-fit results obtained from fitting dominance and unfolding IRT models to polytomous data simulated to reflect different response processes. The aforementioned extensions will add to the data quality literature in attempt to facilitate decisions made for analytical procedures in settings where unfolding models are appropriate.

Study 3: ***Impacts of Misspecification of Underlying Response Processes on the Performance of Nonparametric and Parametric Person-Fit Statistics***

Selecting the most appropriate IRT model that reflects the response process for the items on a test may be a crucial step in person-fit analyses, as misspecification of the item response process has been shown to significantly alter the rank ordering of persons based on latent trait scores (Roberts et al., 1999; Stark et al., 2006). When items are nonmonotonic and reflect an ideal point response process, using a dominance IRT model to fit the data may incorrectly suggest that people with the most extreme attitudes or opinions will have more moderate positions on the underlying latent attitude continuum (Roberts et al., 1999). If the ordering of people on the underlying latent continuum is distorted, then it is likely that person-fit statistics anchored to this order will misclassify aberrant responders. For example, if a person with an extremely high true attitude is incorrectly identified as having a more moderate attitude due to using a dominance IRT model for items reflecting an ideal point response process, then when this person gives an extreme answer to an item, the response may be flagged as aberrant due to the expectation of a moderate response. Similarly, if the person with an extremely high true attitude gives a moderate answer to an item, they may not be flagged even though this may be an aberrant response. If misclassification rates of true attitudes are high, this could have severely detrimental effects on scales intended to serve as admission criteria or classification status. Furthermore, using person-fit statistics based on the erroneous ordering of persons in the data cleaning process could make matters worse. The previous Study 2 informs on this issue for nonparametric person-fit statistics. However, no studies have explored the potential effects of misspecification of dominance models for unfolding data (and vice/versa) on the performance of parametric person-fit statistics.

In study 3, parametric person-fit statistics, $l_{z(p)}$ and $l^*_{z(p)}$, are included to assess their performance with the aforementioned GGUM and GPCM datasets with varying conditions (test length, proportion of aberrant responders and proportion of aberrant responses) for midpoint and extreme responding, longstring responses, and random responding. Model-data fit results from Study 1 will inform decisions made for Study 3. It is anticipated that in the cases of poor model fit, the person-fit statistics will not perform as well. This study aims to provide insight on the point at which these declines in performance are most apparent.

Because parametric person-fit statistics depend on parameters estimated by the applied IRT model, misspecification could potentially lead to inaccurate results in flagging aberrant responses. In this case, data cleaning could do more harm than good. Therefore, an additional condition is included in Study 3 where the cross-fitting of the GGUM model to GPCM data and the GPCM model to the GGUM data is implemented. The primary goal of Study 3 is to reveal the impacts of model misspecification (GPCM to GGUM data and the GGUM to GPCM data) on the person-fit statistics. The nonparametric person-fit statistic(s) selected from the results of Study 2 are hypothesized to be less affected by model misspecification since nonparametric person-fit statistics do not rely on parameter estimates. Thus, it will be informative to compare the performance of the parametric and nonparametric person-fit statistics under these conditions.

Research Questions

The research questions driving each of the three studies include:

1) How does model fit compare for dominance and unfolding models (GPCM and GGUM) applied to both dominance and ideal point response data simulated with no aberrant responses? [Study 1]

a) Is the GGUM able to fit data generated under a dominance response process reasonably well in comparison to the GPCM?

2) How is model fit impacted for both dominance and unfolding models applied to both dominance and ideal point response datasets (GGUM fit to GGUM data, GGUM fit to GPCM data, GPCM fit to GGUM data, GPCM fit to GPCM data) when different types and proportions of aberrant response strings are included? [Study 1]
    a) Do certain conditions (type of aberrant response, proportion of aberrant responders and proportion of aberrant responses within an aberrant response vector) have different results for model-data fit?
3) How do the selected nonparametric person-fit statistics (H^T, $U3^P$, G_N^P, and G^P) perform under an unfolding model versus a dominance model framework? [Study 2]
    a) Are the trends and magnitudes for detection and type I error rates (e.g., higher detection rates with longer tests) the same under unfolding and dominance frameworks?
4) What kinds of aberrant behavior are most/least easily detectable via nonparametric person-fit analyses when using unfolding vs dominance response frameworks? [Study 2]
5) How does the performance of the parametric person-fit statistics $l_{z(p)}$ and $l_{z(p)}^*$ compare under an unfolding and dominance model context? [Study 3]
    a) Are the trends and magnitudes for detection and type I error rates (e.g., higher detection rates with longer tests) the same under unfolding and dominance frameworks?

6) What kinds of aberrant behavior are most/least easily detectable via parametric person-fit analyses when using unfolding vs dominance response frameworks, and under what conditions? [Study 3]
7) How accurately do the selected person-fit statistics identify aberrant responding when a(an):
    a) Dominance model is applied to dominance data?
    b) Unfolding model is applied to unfolding data?
    c) Dominance model is applied unfolding data?
    d) Unfolding model is applied to dominance data?
8) What are the effects (if any) of model misspecification of dominance models for unfolding data and unfolding models for dominance data on the performance of person-fit statistics? [Study 3]
    a) If $l_{z(p)}$ and $l^*_{z(p)}$ do not perform well when models are misspecified, do the optimal nonparametric person-fit statistics for dominance and ideal point data (identified in Study 2) perform better for the types of aberrant responding studied?

Summary

Application of person-fit analyses has gained prominence in several fields including education, personality assessment, psychological assessment, and attitudinal assessment (Rupp, 2013). Although the importance of addressing person-misfit in data has gained relative awareness in the last few decades, a paucity remains in the literature regarding person-fit under unfolding model frameworks. In order to effectively investigate these joint concepts, it is necessary to first understand any differential impacts of aberrant responding on model-fit for unfolding and dominance IRT models (Study 1). Once model fit with aberrant data is better understood for these two IRT approaches, the performance of nonparametric person-fit statistics

is further investigated under dominance and unfolding model contexts (Study 2). Lastly, the impacts of model misspecification on the performance of parametric person-fit statistics, that use model-specific parameter estimates for evaluating aberrant responding, are investigated (Study 3).

When an ideal point response process is responsible for observed empirical data, the application of an unfolding IRT model is practical, and many times advised. Unfolding models are considered flexible in that they are able to scale items from both extreme ends of a continuum as well as scaling neutral items in the middle of a continuum. Recoding reverse worded items is also unnecessary in an unfolding model context, due to the nature of how the probabilities are computed. As unfolding models have many advantages and are becoming more widely used by many researchers, the rise of questions regarding how data quality measures perform in the data cleaning process is expected. It is essential to gain a better understanding on how underlying response processes affect data-model fit, and how effectively aberrant data are identified when screening for higher quality data when using an unfolding model framework versus the more popular dominance models.

The three main novel contributions of this dissertation include: 1) insight on GGUM model-data fit when aberrant data is systematically entered, 2) how nonparametric person-fit statistics (H^T, $U3^P$, G_N^P, and G^P) perform under the framework of GGUM, and 3) how the performance of parametric person-fit statistics ($l_{z(p)}$ and $l_{z(p)}^*$) is impacted by misspecifying the GPCM to GGUM data and the GGUM to GPCM data. Additionally, the conditions of existing research in the field (e.g., how $l_{z(p)}$ and $l_{z(p)}^*$ perform in GGUM context) are extended to complement the literature.

CHAPTER 2

LITERATURE REVIEW

The literature review is organized into three main parts. Part I focuses on building an understanding of what aberrant responding is and provides a brief background on its detection with person fit statistics. Part II describes dominance and unfolding models. This section first covers the underlying response processes (ideal point and dominance) and how they are reflected in the respective unfolding and dominance IRT models. The primary differences and similarities between the two model frameworks are outlined and previous research is reviewed. Lastly, research involving the detection of aberrant responses under an unfolding model context is recognized and factors relevant to the current study are presented in Part III.

Part I: Detection of Aberrant Responding

Aberrant Responding

Several cognitive models for answer processes exist that help explain how and why people respond the way they do. Understanding these processes is imperative in recognizing their potential relationship with data quality. For example, Tourangeau and Rasinski (1988) theorized that the process of answering attitude questions begins with interpreting the question and retrieving information from the brain, followed by using the retrieved information to form judgement, and finally mapping the participant's judgement on to one of the available answer choices (additionally, participants may edit their choices by checking their consistency with other answered questions). Ideally all participants would go through an honest cognitive process similar to this. In reality however, participants may lack motivation, or the cognitive effort required for such a process, yielding response data that is aberrant, or deviant from what would be expected if the participants were accurately recording their perspectives. Researchers have

long been aware that responders may be distracted or lack the cognitive effort to provide accurate and meaningful responses (Cronbach, 1946). Further, some respondents may possess the cognitive effort, but willfully provide false responses for social desirability or some ulterior motive (Paulhus, 1984). Both cases would be considered aberrant data, and the inclusion of these types of responses can lead to misleading conclusions by researchers.

Types of Aberrant Responding. Aberrant responses stem from numerous possible behaviors and characteristics of the respondent. It is important for researchers to be aware of the sources of these behaviors and knowledgeable about the different forms they can take. Many of these behaviors are linked to carelessness (or inattentiveness), random responding, straight-lining (or long-string responses), and creative responding (Curran, 2016; Huang et al., 2012; Johnson, 2005; Karabatsos, 2003; Meijer et al., 1996).

Types of Response Styles. Additionally, various types of response styles may result in misfitting item scores. Extreme response style (ERS) refers to people tending to choose the upper or lower extreme categories, regardless of the item content (Greenleaf, 1992). People who tend to choose the middle response option regardless of item content may be deemed as exhibiting what is known as middle response style (MRS; Baumgartner & Steenkamp, 2001), or mid-lining. Another type of response style includes acquiescent response style (ARS; Baumgartner & Steenkamp, 2001), where respondents tend to agree with items or select positive responses regardless of item content. Researchers have warned that when a sample has large disparities in exhibiting response styles (such as ERS), comparing participants' test scores becomes very difficult and the contamination may threaten the validity of conclusions drawn from the data (Baumgartner & Steenkamp, 2001; de Jong et al., 2008; van Herk et al., 2004). Person-fit analyses, discussed later, may help identify participants that contribute to this lack of

comparability due to response styles or other types of aberrant responding present in the data (Emons, 2008, 2009; Karabatsos, 2003; Meijer & Sijtsma, 2001; Tendeiro, 2017).

Impact of Aberrant Responding. When the proportion of aberrant responses to items on an instrument is at least moderate (a common scenario), these aberrant response vectors have the potential to impact the reliability of the measure, its validity, and ultimately lead to misleading conclusions made from the data. DeSimone et al. (2018) found that even 10 to 15 percent contamination (of insufficient effort response vectors) in a dataset should be a cause for concern. Random responding, for example, may lead to lower interitem correlations, lower reliability estimates, and mask the real factor structure. Thus, the researcher is at higher risk for making Type II errors and failing to reveal relationships between variables that may actually exist (McGrath et al., 2010). Straight-lining (or long-string responses) on the other hand, may artificially inflate reliability if the items are worded in only one direction.

In a study investigating the impact of simulated aberrant response vectors, researchers confirmed that response vectors mimicking random responding have a different impact than those that are invariant (i.e., straight-lining; DeSimone et al., 2018). More specifically, random responding seemed to decrease inter-item correlations, resulting in flatter eigenvalue distributions in the PCA, and lower coefficient alpha estimates. The samples in the study that contained straight-line vectors showed an increase in inter-item correlations, increased reliability, and more skewed PCA results.

The impacts of aberrant responding on the performance of person fit statistics has also been studied (Emons et al., 2003; Glas & Meijer, 2003; Karabatsos, 2003; Rudner, 1983; Tendeiro, 2017; Tendeiro & Meijer, 2014). Many studies have found that the accuracy of person fit statistics increases as the amount of aberrant responses within a response string increases

(conditions studied up to 41%; Emons et al., 2003, 2004; Glas & Meijer, 2003; Karabatsos, 2003; Tendeiro, 2017). St-Onge and colleagues (2011) found that this relationship may not be linear, but rather increase until a peak is reached. Specific peaks ranged from approximately 30% to 55% depending on the type of person fit statistic and the type of aberrant response.

Overall, research suggests a somewhat more complex answer to how aberrant responders are going to impact data and analyses. The answer depends on several factors including, but not limited to, how items are written (negatively versus positively), inter-item correlations, types of aberrant responding, proportion of aberrant responses, proportion of aberrant respondents in the sample, and the number of items. What is certain, is that researchers should be cognizant of the potential detriment low quality data may have on their conclusions (e.g., increased Type I and Type II error and obscured factor structure).

Detecting Aberrant Responses

Several methods for detecting aberrant responses have been studied and utilized to minimize threats to data quality. Response time is one of the most common methods used to identify "speeders" who do not meet the criteria based on a minimum time necessary for sufficient cognitive effort (Cyr, 2000; Huang et al., 2012; Wood et al., 2017; Zhang & Conrad, 2014). Instructed items may also serve as an attention check (DeSimone et al., 2015; Kung et al., 2018; Meade & Craig, 2012a). Other data quality measures include, but are certainly not limited to, psychometric and semantic synonyms and antonyms, Mahalanobis Distance, a long-string index, and an individual reliability measure (Bowling et al., 2016; Cyr, 2000; DeSimone et al., 2015, 2018; Huang et al., 2012; Jackson, 1976; as cited in Johnson, 2005; Meade & Craig, 2012a; Turner, 2018). For the purpose of this dissertation, methods used to detect aberrant responses are focused on person-fit statistics which are described below.

Person-Fit. As mentioned above, person-fit statistics exist among many methods for examining response behaviors on cognitive and non-cognitive assessments. "Person-fit," also referred to as "appropriateness measurement," refers to the degree to which a person's item response pattern departs from what is expected based on an item response theory (IRT) model or the response patterns of other non-aberrant responding persons in the group. Researchers have been concerned with person-fit since the early 1900's, involving theories and methods for estimating reliability and recognizing measurement error (Cronbach, 1946; Lord & Novick, 2008; Spearman, 1910; Thurstone, 1927). Rupp (2013) notes that applications of person-fit have gained popularity in several assessment areas including educational settings, psychological assessment, personality assessment, attitudinal assessment, and health outcomes assessment. The establishment of item response theory initiated an escalation in person-fit research, resulting in over forty statistics available to test person-fit (Meijer & Sijtsma, 2001).

Two general types of person-fit statistics include parametric and nonparametric statistics. Parametric statistics involve some form of measuring the disparity between the observed data and the estimated response predictions resulting from an IRT model. In contrast, nonparametric person-fit statistics are based on a more general model framework and less strict assumptions than parametric IRT (Sijtsma & Molenaar, 2002). Additionally, person-fit statistics may be considered global or local. Global person-fit refers to assessing misfit of persons using all items in a response vector. Conversely, local person-fit is evaluated using subsets of items in a response vector (Emons, 2009; Rupp, 2013). These distinctions are important in reviewing person-fit statistics literature and prompt deep consideration in comparing results across studies.

Person-Fit Comparison Studies. Table 1 outlines several person fit simulation studies and contains information on the generating model(s), sample size(s), percent aberrant

respondents, percent aberrant responses, test length, type of aberrant behavior(s), and person fit statistic(s) involved. The table has been adapted from Rupp's (2013) review of methodology for person fit analysis research spanning the timeframe of 2000 to 2010, with the addition of more current studies up to 2020. The additional studies were found by searching key words such as "person fit," "fit," "aberrant response," "longstring," "random response," "insufficient effort," and related words. Sources included *Google Scholar* and the University of Arkansas library and dissertation database. Various studies from the table, as well as empirical person fit studies, are discussed in more detail below.

Table 1. Person Fit Simulation Studies (modified and updated from Rupp, 2013)

Author(s)	Model	Sample Size	% AbN	Test Length	% AbI	Type of Aberrant Response	Person Fit Statistic(s)
Armstrong et al. (2007)	3PL	10,000	50	121	15, 20, 30	Spuriously high/low	l_z
Armstrong & Shi (2009a)	3PL	10,000	1,3	100	8, 10, 12	Spuriously high/low/mixed	$CUSUM_{LR}, CUSU$ $CUSUM_{IRT}, U3, C$
Armstrong & Shi (2009b)	3PL	10,000	1,3	100	8, 10, 12	Aberrant	$CUSUM_{LR}, CUSU$ l_z, U, W, UB
Artner (2016)	1PL	100, 500	5, 30	25, 50	variou s	Careless, Cheating, Guessing, Distorting, Fatigue	$C^*, U3, H^T$
Choi & Cohen (2008)	3PL-T	1,000	5, 10, 20	35	20	Guessing	l, U, W
Clark (2010)	GRM	1,000	1, 5, 10, 25	25	10, 30, 50	Cheating	l_{c0}, l_{cz}, M $- l_{c0}, M - l_{cz}$

Table 1 (Cont.)

Conijn et al. (2014)	GRM	10,000	10, 30	30, 60	20, 40, 50, 60, 80, 100	Random responding	$l^p_{z(uni)}, l^p_{z(sub)}, l^p_{zm}$ $l^p_{z(com)}, l^p_{z(sel)}$
Cui & Leighton (2009)	AHM	4,000	50	14, 28, 42	100	Creative, Structural Misspecification, Random	HCI
de la Torre & Deng (2008)	3PL	5,000	100	10, 30, 50	10, 30	Cheating, Speeding, Lack of Motivation	l_z
Dimitrov & Smith (2006)	1PL	9,000	27	10, 20, 30	20, 40	Guessing, Cheating	t, t^*, Z_3, Z_3^*, H^T
Emons et al. (2003)	4PL	1,000	100	20, 40	12, 20, 25, 50	Cheating, Inattentive	$G^*, U3, l_0, \zeta$
Emons et al. (2004)	4PL	1,000	NP	20, 40	12.5, 25, 40	Answer Copying, Test Anxiety	$LR-\beta, ZU3,$ $G^2_\beta, G^2_{\gamma,\mathrm{SL}}, G^2_{\gamma,\mathrm{SH}}$
Emons (2008)	2PL, GRM	6,000	50	12, 24	25, 50, 75, 100	Careless, Extreme options, Reverse wording	$U3^p, G^p_N, l^p_z$
Emons (2009)	GRM	1,000; 3,000	5, 10	12, 24	17, 25, 33, 50, 67, 100	Careless, Extreme options	l^p_z, p_{x_v+}
Ferrando (2009)	LFA	5,000	6	10, 18, 24	20, 25	Random	$M-l_{c0}, M$ $-l_{cz}$
Ferrando (2010)	CRM	500	6	10, 30	20	Random	l_{c0}, l_{cz}
Glas & Meijer (2003)	3PL	400; 1,000	10	30, 60	17, 33, 50	Local dependence, Guessing	$l, W, UB, T_1, T_2,$ $T_{lag}, \gamma_1, \gamma_2$
Glas & Dagohoy (2007)	2PL, GRM, SM, GPCM	400; 1,000	10	40, 60	25, 50	Ability increase, Guessing	LM test
Hendrawan et al. (2005)	3PL	400; 1,000	10	30, 60	17, 33, 50	Item disclosure, Guessing	$l, UB, W, \zeta_1, \zeta_2$

Table 1 (Cont.)

Karabatsos (2003)	1PL	500	5, 10, 25, 50	17, 33, 65	18, 41, 100	Cheating, Guessing, Careless, Creative, Random	36 statistics
Liu et al. (2009)	DINA	1,000; 2,200	18, 100	60, 90	100	Spuriously high/low, Strategy switching	$LR\ test(T_1, T_2)$
Raiche & Blais (2003)	1PL	1,000	100	various	10, 20	Incorrect, Random	$l_z, W, \zeta, I_{ran}, I_{inv},$
Sijtsma & Meijer (2001)	1PL, 4PL	3,000	100	40, 80	12.5	Careless	$P, ZU3$
St-Onge et al. (2009)	1PL, 2PL, 3PL	100; 1,000	5	40	20	Spuriously high	$ECI2_z, ECI4_z, l_z$
St-Onge et al. (2011)	2PL	1,000	5	20, 40, 60, 80	10 to 60	Spuriously high/low	$l_z, U3, ECI2, H^T$
Tendeiro (2017)	GGUM *	1,000	5, 10, 20	10, 20, 40, 100	10, 20, 25	Extreme and midpoint responding	$l_{z(p)}, l^*_{z(p)}$
Tendeiro & Meijer (2014)	3PL	1,000	5, 10, 25	15, 25, 40	20, 40, 50	Spuriously high/low/mixed	$C^*, U1, U3, H^T, PE, CUSUM_l, CUSUM_u, CUSUM_{2-sided}$
Turner (2018)	RSM	1,000	5, 10, 15, 20, 25, 30	36	50, 100	Random, careless, fatigued, semantically driven, straightlined	U^{3+}, Z_h^+, l_z^*
Wang et al. (2008)	1PL	6,000	18	60	20	Cheating	$l_z, ECI4_z, \chi^2_D$
Zhang & Walker (2008)	2PL	1,000	10	10, 20, 40	20	Cheating	$H^T, D(\theta)$

Note. % AbN= percent of simulated respondents with aberrant responses. % AbI= percent of items with aberrant responses in a single aberrant response vector. U= Unfolding model. D = dominance model. 1PL, 2PL, 3PL, 4PL= 1-, 2-, 3-, and 4-parameter logistic models respectively. 3PL-T = 3-parameter logistic testlet. GRM = graded response model. AHM = attribute hierarchy method. LFA = Linear factor analysis. CRM = continuous response model. SM = sequential

model. DINA = deterministic inputs noisy and-gate. GGUM = Generalized graded unfolding model. RSM = Rating Scale Model. Table adapted from Rupp (2013) and modified with relevant person fit studies.

Karabatsos' (2003) simulation study is one of the most cited person-fit comparison studies. In the study, the performance of 36 person-fit statistics in detecting aberrant responding examinees (cheaters, creative respondents, lucky guessers, careless respondents, and random respondents) is compared under the context of the Rasch model. The five types of aberrant-responding examinees are crossed with two other factors: 1) Proportion of aberrant-responding examinees (5%, 10%, 25%, 50%) and 2) test length (17 items, 33 items, and 65 items). Results were organized by factor (type of aberrant-responding examinee, proportion of aberrant-responding examinee, and test length). With respect to the type of aberrant-responding examinee, the results suggested that creative and cheating respondents are the most difficult to detect, while careless and random respondents are the easiest to detect. For cheaters, creative respondents, and careless respondents, the person-fit statistics H^T and $D(\theta)$ performed the best. In addition to H^T and $D(\theta)$, E_i also performed the best at detecting lucky-guessing respondents. Several person-fit statistics (H^T, $D(\theta)$, $E_i, r_{pbis}, C, MCI, U3, ECI3, ECI5, and\ M$) were considered the most effective at identifying random responding examinees. In reference to the proportion of aberrant-responding examinees, detection rates typically decreased as the proportion of aberrant responders increased. While there were negligible differences in the person-fit statistics' performances under the 5%, 10%, and 25% aberrant-responding examinee conditions, $E_i, ECI1, ECI2,$ and $ECI6$ performed the best when 50% of the examinees had aberrant response vectors. Finally, with regards to the test length conditions, results illustrated how detection rates increased as test lengths increased. For all three test lengths, H^T, $D(\theta)$, and l were the most effective person-fit statistics. Furthermore, when all simulees from all conditions were

combined, H^T was the most effective at detecting aberrant responding examinees, with $D(\theta), C, MCI$, and $U3$ tying for second-best. Karabatsos (2003) concludes by suggesting critical values for these top five performing person-fit statistics which maximizes the sensitivity and specificity rate ($H^T \leq .22,\ D(\theta) \geq .55,\ C \geq .53,\ MCI \geq .26$, and $U3 \geq .25$).

Rudner (1983) compared the ability of nine person-fit statistics ($r_{pbis}, r_{bis}, NCI, C_i, U1, U3, W1, W3, and\ L3$) to detect spuriously low and high respondents. Two datasets were independently generated using a 3PL model based on the parameterization of 80 verbal Scholastic Aptitude Test items and 45 teacher-developed biology exam items. Response outcomes were changed for 5, 10, 15, or 20 percent of items to create spuriously high and low performing respondents. Critical values were set to the cut-off point where the lowest 5% of the most extreme values lied for the control group. As the number of aberrant item responses increased, detection rates also increased, indicating the severity of aberrant response vectors does make a difference in misfit detection. Spuriously high scores were usually easier to detect than the spuriously low scores. The $U3$ seemed to work well on the longer test but not on the shorter test.

In 2014, Tendeiro and Meijer conducted a simulation study similar to Rudner (1983) in that they examined person-fit statistic performance for detecting spuriously high and spuriously low responding examinees with dichotomous items generated using a 3PL model (Tendeiro & Meijer, 2014). Spuriously high responding examinees were generated by taking a proportion (.05, .10, or .25) of the low ability examinees (theta value below -.5) and replacing 20%, 40% or 50% of the failed items (responses of 0) with scores drawn from a Bernoulli distribution with a .80 probability. Thus, incorrect responses (values of 0) had a .80 probability of being replaced with correct responses (values of 1). Spuriously low response vectors were similarly created

using a Bernoulli distribution with a .20 probability, resulting in a .80 chance for 1s to be changed to 0s. Data was analyzed using three types of datasets according to the type of aberrant response behavior involved: 1) datasets with proportions of only spuriously low response vectors, 2) datasets with proportions of only spuriously high response vectors, and 3) datasets with equal proportions of both spuriously low and high response vectors (mixed). Datasets were simulated such that aberrant responding was spread throughout the test as well as constrained to be local (consecutive). Nonparametric person-fit statistics $C^*, U1, U3, H^T, PE$ were compared. Additionally, they examined the performance of the lower, upper and two-sided cumulative sum (CUSUM) indices proposed by Krimpen-Stoop and Meijer (2001). Furthermore, the corrected version of the popular l_z statistic (Drasgow et al., 1985), l_z^* (Snijders, 2001), was estimated in order to compare the nonparametric statistics with a well-known parametric statistic. The true-positive rate, false-positive rates, and correlations between the person-fit statistic and total scores were used for evaluation criteria. Four-way ANOVAs were conducted for the person-fit statistics to investigate the impacts of item discrimination, test length, proportion of aberrant respondents, and proportion of aberrant responses. For the general, non-consecutive aberrant data, the proportion of aberrant responders had a negligible effect for all indices except l_z^*, where detection rates decreased as the proportion of aberrant responders increased. The proportion of item responses in a response vector had no practical effect with spuriously low respondents, but for spuriously high respondents, detection rates increased as the proportion of aberrant items increased from 20% to 40%, but no additional increase from 40% to 50%. Increasing the proportion of aberrant responses in a response vector also had a moderate effect, increasing detection rates for the sample with mixed spuriously high and low respondents. For all indices, increasing the discrimination parameters increased detection rates. Detection rates also increased

as test lengths increased, especially for spuriously low and mixed respondents. The results from this study suggested that the H^T statistic was the most efficient in detecting spuriously low, high, and mixed response vectors with general-type data ($U3, C^*, and\ U1$ were a close second-best). This corresponds with Karabatsos' (2003) finding that H^T was the best performing person-fit statistic.

Emons (2008) conducted nonparametric person-fit analyses with polytomous items using both simulated and empirical, dominance response-type data to compare three person-fit methods: 1) Number of Guttman errors for polytomous items (G^P), 2) normed Guttman errors (G_N^P), and 3) the generalized $U3$ statistic (Van der Flier, 1982). The parametric l_z^p person-fit statistic was also included in the study for comparison purposes. In the simulation study, data were generated under the graded response model (GRM) and misfitting item score vectors were created to mimic carelessness and inattention, extreme response style, and reverse scoring effects. By simulating two datasets (one clean, normal behavior and one with aberrant responses), the critical value for each Type I error rate was obtained. For the normal behavior ("clean") dataset, the distribution of the G^P statistic conditional on the sum score varied across sum scores with the distributions for middle range sum scores having a higher mean and larger variance. The conditional distributions of G_N^P and $U3$ across sum scores were fairly consistent except for very high and very low sum scores. Emons notes that this finding supports the notion that person-fit indices should not be used with extreme scores. Therefore, extreme scores were removed from the datasets for subsequent analyses. For the careless and inattentive response condition, G^P showed the best overall performance out of the three nonparametric statistics. However, the parametric l_z^p statistic revealed slightly higher detection rates (differences ranged from .01 to .11) for the aberrant response behavior than the nonparametric indices. Increasing the

number of response options only slightly increased detection rates, with a somewhat larger effect when the number of mis-fitting items in a response vector was larger. Detection rates were lower for extreme response style and showed low power (for test length of 12 and the number of misfit items in a vector equaling 6, the power was less than .50 for conventional Type I error rates), indicating this type of aberrant response behavior is difficult to detect. However, increasing the number of response options from 3 to 5 yielded acceptable detection rates when the number of aberrant responses in a response vector was high enough. The detection rates for extreme response behavior were actually higher for lower item discrimination. Emons posits that this can be explained by the fact that higher discrimination for easy and difficult items will lead to higher probabilities of choosing the highest and lowest categories respectively. Thus, the difference between normal and aberrant behavior (due to extreme responding) diminishes with higher item discrimination. G_N^P and $U3$ were recommended over G^P and l_z^p for extreme response behavior, however G^P generally performed better than the other two nonparametric person-fit statistics for the other types of aberrant responding, and even slightly better than l_z^p in some conditions (though differences were small). In the empirical study, Emons (2008) used data from a study using 17 items to assess people's coping behavior with industrial malodor (Cavalini, 1992). Items were on a 4-point scale and a subset of eight items were selected that demonstrated a unidimensional scale. The three nonparametric person-fit statistics used in the simulation study were applied to the empirical dataset. Correlations between the person-fit statistic ranged from .88 to .89. A total of 44 response vectors (out of the sample of 675) were flagged by at least one person-fit statistic. G^P flagged 10 respondents that were not flagged by either G_N^P nor $U3$, and G_N^P flagged 10 respondents that were not flagged by either G^P nor $U3$. The statistic $U3$ did not flag any response vectors that were not also flagged by either of the other two statistics.

Dimitrov and Smith (2006) simulated datasets using the dichotomous Rasch model to compare four parametric person-fit statistics (t, t^*, Z_3, Z_3^*) and one nonparametric person-fit statistic (H^T). They used three different shorter test lengths (10, 20, and 30 items) and two levels of aberrant response severity (20% or 40% of the most difficult items within a response vector were considered aberrant) to test the performance of the five person-fit statistics in detecting two types of aberrant response behavior (cheating or guessing). Results suggested that the adjusted parametric statistics (t^* and Z_3^*) consistently outperformed the respective unadjusted parametric statistics (t and Z_3), but not by a large amount. Additionally, the nonparamtric person-fit statistic (H^T) once again outperformed the parametric person-fit statistics in most conditions. Specifically, for the longer test lengths of 20 and 30 items, H^T seemed to consistently detect cheating and guessing more efficiently. However, for the shorter test length of 10 items, *t* and t^* slightly outperformed the other statistics in detecting guessing, while the nonparametric statistics all had very similar results for detecting cheating (with H^T somewhat behind).

Overall, the general consensus seems to be that the comparative performance of person-fit statistics depends on the conditions in place, such as test length, aberrant response type, proportion of aberrant responding examinees, proportion of aberrant responses within each response vector, distributions of item parameters, item characteristics, and more. Regarding methodology, Rupp (2013) notes four principal steps taken by almost all person fit simulation studies: 1) data generation according to a specified model, 2) generated response vectors are altered to reflect a type and severity of aberrant responding, 3) the statistical model(s) is (are) fit to the updated data from step 2, and 4) the person fit statistics are computed and performance is evaluated. Results have been vast and somewhat varied across studies. However, disregarding a

few exceptions, trends have emerged within certain conditions. These trends are noted in Table 2 and described in the section below.

Aberrant Responding Detection Under Various Conditions. The performance of person fit statistics in the detection of aberrant responding has been shown to depend on several conditions including how many people respond aberrantly in a sample, how many of their item responses are aberrant, test length, and number of response options. For example, as the proportion of aberrant responders in a sample increases, detection rates may decrease due to over-contamination and blurring the lines between what response patterns are normal and which are abnormal (Armstrong & Shi, 2009a, 2009b; Emons, 2009; Karabatsos, 2003). On the other hand, as the proportion of aberrant responses in a response vector for a specific individual increases, detection rates have generally increased (Emons, 2009; Rupp, 2013). Additionally, detection rates of several person-fit statistics seem to increase as the test length increases (Dimitrov & Smith, 2006; Emons, 2009; Karabatsos, 2003; Meijer et al., 1996; Tendeiro & Meijer, 2014). Increasing the number of response options has also been shown to increase detection rates (Emons, 2008). The performance of person fit statistics has also been shown to depend on the type of aberrant responding (cheating, guessing, random responding, etc.). Table 2 provides four examples of studies with regards to the type of aberrant responding and person fit performance. The diversity in results of the studies highlights the complexity in determining which person fit statistic is the optimal choice. Determining which person-fit statistic to use requires careful consideration of all of these conditions.

Table 2. Examples of Person-Fit Performance Under Various Test Lengths, Types of Aberrant Responding, and Proportions of Aberrant Respondents

Condition: Aberrant Response Types		
Author and Person-Fit Statistics Used	Type	Results
Karabatsos (2003) 36 Person-fit statistics	1) cheaters 2) lucky guessers 3) creative 4) careless 5) random	Creative and cheating respondents were the most difficult to detect, while careless and random respondents were the easiest to detect. For cheaters, creative respondents, and careless respondents, the person-fit statistics, H^T and $D(\theta)$, performed the best. In addition to H^T and $D(\theta)$, E_i also performed the best at detecting lucky-guessing respondents. Several person-fit statistics (H^T, $D(\theta)$, $E_i, r_{pbis}, C, MCI, U3, ECI3, ECI5, and\ M$) were considered the most effective at identifying random responding examinees.
Rudner (1983) $r_{pbis}, r_{bis}, NCI, C_i,$ $U1, U3, W1,$ $W3, and\ L3$	Spuriously low and high respondents	Spuriously high scores were usually easier to detect than the spuriously low scores. The weighted Birnbaum model, $W3$, seemed to be the most consistent at efficiently identifying spurious respondents over all condition.
Tendeiro & Meijer (2014) $C^*, U1, U3, H^T, PE, l_z^*$	Spuriously low, high, and mixed respondents	H^T statistic was the most efficient in detecting spuriously low, high, and mixed response vectors ($U3, C^*, and\ U1$ were a close second-best)
Emons(2008) G^P, $G_N^P, U3$ and l_z^p	Carelessness and inattention, extreme response style, and reverse scoring effects	Detection rates were lower for extreme response style and show low power (For this response behavior, $G_N^P$and $U3$ were recommended over G^P and l_z^p) G^P and l_z^p performed better for reverse wording effect.

Based on results from previous research, popularity in practice, and prior use with ideal point response data, the following person-fit statistics were chosen for inclusion in this

dissertation: $l_{z(p)}, l^*_{z(p)}, H^T, U3^P, G_N^P$,and G^P. Further details on the choice for inclusion are given for each statistic in the following section.

Person-Fit Statistics

The body of literature involving person-fit statistics is immense. Therefore, the following descriptions cover the statistics that are most commonly used and relevant to the current study. As previously mentioned, there are two general types of person-fit statistics: parametric and nonparametric. The person-fit statistics discussed below are organized according to which of the two classification methods they belong (parametric: $l_{z(p)}$ and $l^*_{z(p)}$, nonparametric: $H^T, U3^P, G_N^P$,and G^P). Throughout the chapter, mathematical notation is used to help demonstrate how each person fit statistic is computed. Respondents are indexed by *n* (*n*=1, ..., *N*) and items are indexed by *j* (*j*=1, ..., *J*). In IRT, theta (θ) is generally used to represent each respondent's trait or ability level, and the probability of observing a particular response to an item can be estimated using an IRT probability function (P) which incorporates theta and certain item characteristics (e.g., discrimination, difficulty, guessing). Observed responses of respondent *n* to item *j* are represented by X_{nj} and the probabilities of a response to an item, given by an IRT model, are represented by P_{nj}. Further, the probability of endorsement (or correct response) of an item by a person is represented by P_{nj1}, where the probability of non-endorsement (or incorrect response) of an item by a person is represented by P_{nj0}.

Parametric statistics. Person-fit statistics that involve some form of measuring the disparity between the observed data and the estimated response predictions resulting from an IRT model's parameter estimates are considered parametric (Karabatsos, 2003).

$l^*_{z(p)}$. The $l^*_{z(p)}$ statistic (Drasgow et al., 1985) stems from possibly the most well-known parametric person-fit statistic, l (Levine & Rubin, 1979). To clearly demonstrate how $l^*_{z(p)}$ is

computed, it is beneficial to illustrate its transformation from the earlier l statistic which measures the log-likelihood fit of a response to an item with the prediction based on an IRT model (Karabatsos, 2003). The binomial loglikelihood statistic is computed as follows:

$$l = \sum_{j=1}^{J}[\, X_{nj}(\ln P_{nj1}) + \big(1 - X_{nj}\big)\big(lnP_{nj0}\big)]. \quad (1)$$

To better understand this formula, consider all four possible scenarios for item *j*, given that $\ln(1) = 0$ and $\ln(0) \rightarrow -\infty$:

Table 3. Possible Scenarios in Illustrating the Nature Of *l*

Scenario	Probability of correct answer P_{n1}	Response: correct (1) or incorrect (0) (X_n)	$l = X_n(\ln P_{n1}) + (1 - X_n)(lnP_{n0})$
A	High → 1	1	$l \rightarrow 0$
B	Low → 0	0	$l \rightarrow 0$
C	High → 1	0	$l \rightarrow -\infty$
D	Low → 0	1	$l \rightarrow -\infty$

Notice how if the examinee's response matches what is expected from the probability (scenarios A and B), the likelihood statistic approaches zero. However, if the response does not match what is expected from the probability, the statistic approaches negative infinity (scenarios C and D). This demonstrates how a smaller l-statistic indicates larger misfit.

The limitation of the likelihood-statistic is that it is not standardized and it is unknown what the distribution under a fitting IRT model is. Drasgow et al. (1985) proposed a standardized normal version of the likelihood statistic, l_z, in the following way:

$$l_z = \frac{l - E(l)}{\sqrt{var(l)}}, \quad (2)$$

where $E(l) = \sum_{j=1}^{J}\{P_{nj1}(\theta)\ln\,(P_{nj1}(\theta)) + P_{nj0}(\theta)\ln\,(P_{nj0}(\theta))$ and

$var(l) = \sum_{j=1}^{J} P_j(\theta)Q_j(\theta)\left(log\,\frac{P_j(\theta)}{Q_j(\theta)}\right)^2$.

For the polytomous case where the number of response categories minus 1 is *C*, the mean and variance for equation 2 are defined as follows (Sinharay, 2016):

$$E(l) = \sum_{j=1}^{J} \sum_{k=1}^{C} [\,[\, P_{jk}(\theta)\big(\ln P_{jk}(\theta)\big)]] \text{ and}$$

$$var(l) = \sum_{j=1}^{J} \sum_{k_1=1}^{C_j} \sum_{k_2=1}^{C_j} P_{jk_1}(\theta) P_{jk_2}(\theta) \ln\Big(P_{jk_1}(\theta)\Big) \ln\left(\frac{P_{jk_1}(\theta)}{P_{jk_2}(\theta)}\right).$$

Sinharay (2016, p. 996) further describes the mathematical reasoning behind the above equation.

The interpretation of l_z is similar to the interpretation of l (i.e. lower (more negative) values of the statistic indicates greater misfit). Still, it is only when true theta values are used, that this statistic can be assumed to have an asymptotically standard normal distribution (Molenaar & Hoijtink, 1990). In practice, it is unrealistic to assume true theta values are available. Consequently, a modified version of l_z, l_z^* , was proposed that addresses this concern by accounting for the sampling variability of the estimated theta parameters (Sinharay, 2016).

A thorough, and helpful explanation of the computational formulas involved in calculating l_z^* can be found in Magis et al. (2012). Sinharay (2016) further extended this corrected version for polytomous cases, $l_{z(p)}^*$. Although Tendeiro (2017) found very similar results for the $l_{z(p)}$ and $l_{z(p)}^*$ statistics in an unfolding context, the current study will include both statistics to provide a second study to test whether they perform similarly under different conditions.

Nonparametric statistics. In contrast to the parametric person-fit statistics, nonparametric person-fit statistics are based on a more general model framework and less strict assumptions than parametric IRT (Sijtsma & Molenaar, 2002).

***H^T*.** The H^T statistic (Sijtsma, 1986; Sijtsma & Meijer, 1992) is an adapted version of Mokken's (1971) H_j index, which permits an item to be scaled to the Guttman (1944) model. In a

Guttman scale, items are arranged hierarchically so that the endorsement of one item suggests the endorsement of items below it. Further explanation of the Guttman scale and errors can be found in the Guttman person fit statistic section (G^p) of this chapter. In order to focus this procedure on persons rather than items, Sijtsma (1986) simply transposed the item by person matrix. This transposition results in a statistic that could then detect respondents that do not conform to the Guttman model. Below, the H^T statistic is generalized for polytomous items.

The data matrix for the H^T statistic is composed of *N* rows of participants and *J* columns of items, with each element in the matrix representing an item score. Suppose participant n_1 has an item-score vector, $\boldsymbol{X}_{n_1}$, composed of *j = 1 . . . J* item-scores. The total score for item *j*, $\boldsymbol{T}_j$, is then computed as the sum of all participants' scores for that particular item. The vector $\boldsymbol{T}$ is composed of the total scores for all items (T_1 *to* T_J) and the vector $\boldsymbol{T}_n = \boldsymbol{T} - \boldsymbol{X}_n$. That is, $\boldsymbol{T}_n$ is the vector of all item-score totals excluding participant *n*. Finally, the scalability coefficient for participant *n* is computed as follows:

$$H_n^T = \frac{Cov(\boldsymbol{X}_n, \boldsymbol{T}_n)}{Cov_{max}(\boldsymbol{X}_n, \boldsymbol{T}_n)}, \tag{3}$$

where $Cov(\boldsymbol{X}_n, \boldsymbol{T}_n)$ is the covariance between the participant *n*'s item-scores and the item-score totals for the remaining participants in the sample (excluding that participant). $Cov_{max}(\boldsymbol{X}_n, \boldsymbol{T}_n)$ is the maximum covariance possible between $\boldsymbol{X}_n$ and $\boldsymbol{T}_n$, given the marginal distributions.

The H_n^T person fit statistic represents the degree to which a respondent's item responses match the same ordering as the item-score totals. This statistic is included in the study because it has been found to have improved detection efficiency in several studies (Beck et al., 2019; Karabatsos, 2003; Tendeiro & Meijer, 2014). The coefficient H^T can be used to summarize the individual H^T statistics for all participants in a sample (Ligtvoet et al., 2010):

$$H^T = \frac{\sum_n^N Cov(\boldsymbol{X}_n, \boldsymbol{T}_n)}{\sum_n^N Cov_{max}(\boldsymbol{X}_n, \boldsymbol{T}_n)}. \quad (4)$$

If a clear ordering exists among items and item response functions are spread apart, the H^T coefficient should be high. However, if item response functions overlap, the H_n^T statistics will be less stable and result in lower values for a dataset. If H^T for the overall sample is low, it may not be an appropriate indicator to use for the dataset. H^T values less than or equal to 0.22 have been generally considered unreliable or a result of aberrant data (Karabatsos, 2003).

G^p. The G^p statistic summarizes the number of Guttman errors for polytomous items (Molenaar, 1991). To do this, the item step difficulties (π_{jx_j}) are computed as the proportion of respondents who scored x_j or higher on item *j*. Using an example similar to Emons (2008), suppose a four-category item (item 1) with response options 0 = strongly disagree, 1 = disagree, 2 = agree, and 3= strongly agree, had step difficulties of $\pi_{11} = .85$, $\pi_{12} = .40$, and $\pi_{13} = .20$. This means that 85% of the respondents passed the first step (chose a category higher than the first option), 40% passed the second step (chose an option higher than or equal to the third category), and 20% passed the third step (chose the fourth category). Now suppose item 2, with the same response options, had step difficulties equal to $\pi_{21} = .60$, $\pi_{22} = .30$, and $\pi_{23} = .10$. If a respondent scored $x_1 = 1$ on the first item and $x_2 = 3$ on the second item, the ordered vector (based on item step difficulties from least difficult to most difficult) of item step scores for this respondent would be (***Y***= ordered vector; y_k = element *k* of vector ***Y***):

Table 4. Example of Item Step Scores Vector

Ordered Item Steps	$\pi_{11} = .85$	$\pi_{21} = .60$	$\pi_{12} = .40$	$\pi_{22} = .30$	$\pi_{13} = .20$	$\pi_{23} = .10$
Y	1	1	0	1	0	1

This respondent passed the second step of item 2 but failed to pass the easier second step of item 1. Additionally, this respondent passed the third step of item 2 but failed to pass the easier third

step of item 1. Because the G^p statistic counts all pairwise Guttmann ordering errors of all possible item-step pairs, the G^p for this respondent would be equal to three. Formally, the G^p statistic is given by the following equation:

$$G^p = \sum_{l<k}^{JC} y_k(1 - y_l), \quad (5)$$

where *J* is the number of items which here are assumed to all have the same number of response categories (*V*), and thus the same number of steps (*C*). Thus, if *C*= 1, the G^p statistic is specified for dichotomous items. In equation 6, y_l represents all elements of vector $\boldsymbol{Y}$ that are prior to y_k. The more Guttmann errors a respondent has, the greater the G^p statistic, indicating greater person misfit.

G_N^P. The normed number of Guttmann errors, G_N^P, normalizes the G^p statistic to have a range of [0, 1]. Emons (2008) explains that while the minimum possible G^p value is equal to 0 (no misfit), the maximum depends on the sum score (X_+) of the respondent and the ordering of the item steps. In order to compare the G^p statistics across different X_+ scores, the G^p statistic was normed using the following equation (Emons, 2008):

$$G_N^P = \frac{G^p}{\max(G^p|X_+)}. \quad (6)$$

Because the denominator of Equation 6 cannot be expressed in closed form (the item-step scores are structurally dependent), a recursion algorithm can be used to determine the maximum of G^p conditional on the item step ordering. Both the G^p and G_N^P statistics are fairly simple, and thus popular. The statistics are also included in the PerFit package in R. Due to their popularity and effectiveness with certain conditions such as careless and inattentive responders (e.g., Emons, 2008), they are included in the current study.

$U3^P$. Emons (2008) generalized the $U3$ person-fit statistic (Van der Flier, 1980) for application to polytomous items, resulting in the creation of the $U3^P$ statistic. It is included in the

study because several studies have found this polytomous version of $U3$ to have comparative if not better performance than other parametric and nonparametric person-fit statistics (Emons, 2008; Karabatsos, 2003; Tendeiro & Meijer, 2014; Turner, 2018). The statistic is defined in a few steps. First, the sum of the log-odds of the item step difficulties of the steps that were passed, $W(\boldsymbol{y})$, is computed as follows:

$$W(\boldsymbol{y}) = \sum_{k=1}^{JC} y_k \log\left(\frac{\hat{\pi}_k}{1-\hat{\pi}_k}\right), \tag{7}$$

where $\boldsymbol{Y}$ is an observed response vector for J items with C+1 response categories, y_k is the item-step score for item-step k (taking a value of 1 if step k is passed or 0 if step k is not passed), and $\hat{\pi}_k$ is the item-step difficulty for item-step k where π_k is the population proportion of respondents who passed item-step k. To clarify, recall that the vector $\boldsymbol{Y}$ contains elements (y_k) for each step of each item, ordered by difficulty. Next, norming $W(\boldsymbol{y})$ results in the $U3^P$ person-fit statistic as follows:

$$U3^P = \frac{\max(W|X_+) - W(\boldsymbol{y})}{\max(W|X_+) - \min(W|X_+)}, \tag{8}$$

where X_+ is the sum score computed as $X_+ = \sum_{k=1}^{JC} y_k$. The $\max(W|X_+)$ can only be obtained if the following holds:

$$\max(W|X_+) = \sum_{k=1}^{X_+} logit(\hat{\pi}_k). \tag{9}$$

For example, if $X_+ = 10$, then the $\max(W|10) = \sum_{k=1}^{10} log(\frac{\hat{\pi}_k}{1-\hat{\pi}_k})$. The $\min(W|X_+)$ cannot be expressed in closed form due to the structural dependencies between the item-step scores. That is, based on Guttman scaling principals it is assumed that passing a step for an individual item means that all easier steps of that same item are also passed. Thus, Emons (2008) proposed to compute $\min(W|X_+)$ using a recursion algorithm. For more details on the recursion algorithm, see the Appendix from Emons (2008).

Methods for Evaluating Person-Fit Statistic Performance

The true positive rate and true negative rates, reflecting sensitivity and specificity, are commonly used in the literature for evaluation of aberrant response behavior detection (Huang et al., 2012; Meade & Craig, 2012b; Niessen et al., 2016; Turner, 2018). A 2x2 contingency table as demonstrated in Table 3 can be used for classification purposes. As illustrated by the table, if an examinee has aberrant response behavior in reality and is correctly flagged by the detection method, the flagged vector is considered a "true positive" (TP). Similarly, if the examinee has normal response behavior in reality and is correctly not flagged, then the vector is considered a "true negative" (TN). However, if an examinee has normal response behavior in reality, but is incorrectly flagged for aberrant response behavior by the detection method, the flagged vector is considered to be a "false positive" (FP). Finally, if the examinee's response behavior is actually aberrant but they are not flagged then the response vector is considered a "false negative" (FN).

Table 5. Contingency Table for Aberrant Response Classification

	Aberrant Response Behavior in Reality	Normal Response Behavior in Reality	Marginal
Flagged	True Positive (TP)	False Positive (FP)	TP+FP
Not Flagged	False Negative (FN)	True Negative (TN)	FN+TN
Marginal	TP+FN	FP+TN	

$$Accuracy = \frac{TN+TP}{Total\ sample\ size} \quad (10)$$

As noted in Turner's (2018) dissertation on "The Detection and Impact of Low Cognitive Effort Survey Responses," some researchers use accuracy (see Equation 10) to summarize detection efficiency (Meade & Craig, 2012b; Turner, 2018). However, the proportion of the sample that is contaminated (with aberrant responses) must be considered when interpreting results this way. That is, if one is testing the detection performance of a method for a sample containing 8% aberrant response vectors, and the method flags zero aberrant response

vectors, the accuracy rate may still be as high as 92%. Thus, it is important to report multiple measures of performance assessment (e.g., false positives, false negatives, etc.).

Part II: Dominance and Unfolding Models

Response Processes: Dominance vs Ideal Point Process

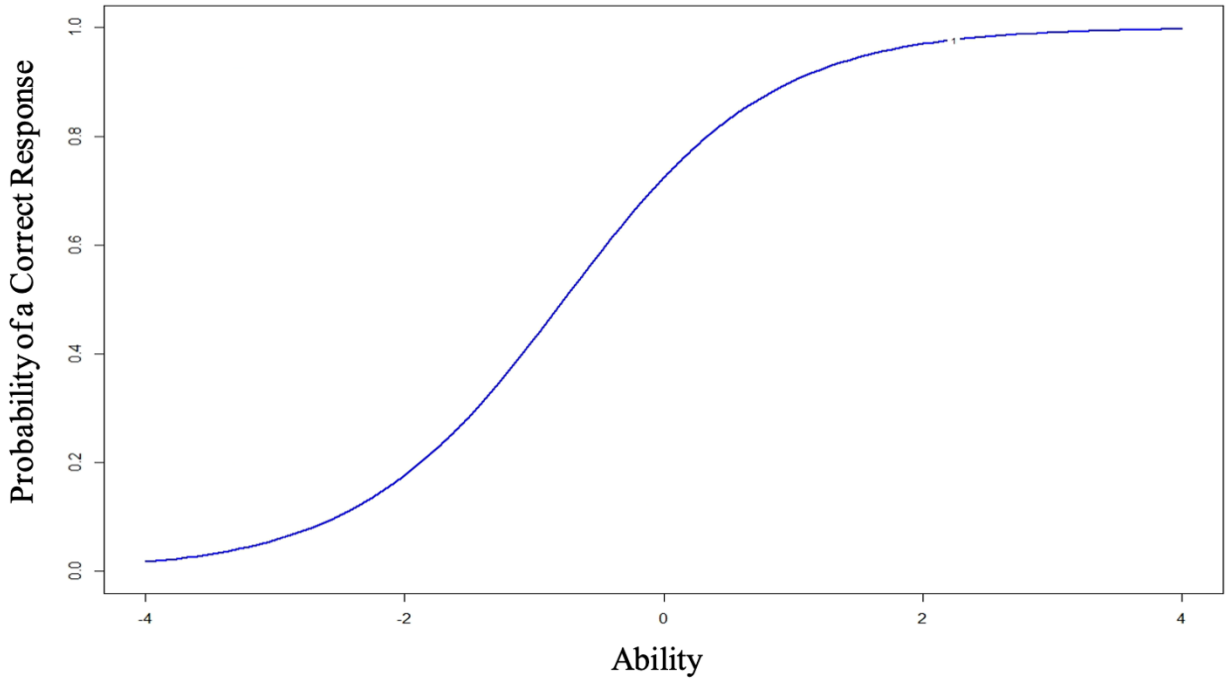
The way test-takers respond to items may differ based on the psychological process used to address the item. The approach to measuring the relationship between these two locations differs depending on the type of model conceptualized. In the dominance response process, as the test-taker's ability or trait level increases, the probability of endorsing an item increases regardless of the item's location on the continuum. For example, when measuring the extravert nature of respondents with items like "I enjoy interacting with people," it may be reasonable to assume that the more extraverted the respondent is, the more likely he or she will endorse this item. The dominance approach has been conceptualized as early as the 1930's (Likert, 1932) and gained more recognition by Coombs (1964), who described the approach asserting that someone who has a trait level higher than the item's standing on that trait will endorse the item.

Evidence from several self-report inventories has indicated that response processes may not always align with what would be expected under the dominance approach. Some items on the 16 Personality Factor Questionnaire (16PF; Conn & Rieke, 1994) have been shown to be non-monotonic (Chernyshenko et al., 2001; Stark et al., 2006). Similarly, the probability of endorsement of some items on the depression content scale for the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher et al., 1989) has been shown to decrease at the higher end of the trait continuum (Meijer & Baneke, 2004). Researchers have suggested the use of an ideal point model in place of a dominance model in cases like these measuring personal preferences and attitudes (Drasgow et al., 2010). The ideal-point process is based on a notion

conceptualized by Thurstone (1928) and termed by Coombs (1964), that assumes a person will endorse an item to the degree that the item reflects the person's own standing on the construct being measured. With this approach, the likelihood of endorsing an item increases as the difference between the test taker's and item's location on the continuum representing the construct of interest decreases.

IRT Models. Item response theory (IRT) refers to a framework of mathematical models that aim to link the observed performance of a test taker (item scores) and the latent (unobservable) ability or trait level of the test taker (Hambleton & Swaminathan, 2013). IRT models are widely used for the development and scaling of assessments in a variety of fields. With IRT, one can model the probability of a response as a function of the latent trait of interest in what is called an item response function (IRF). Figure 1 provides a visual representation of an IRF for a dichotomous item. Note that only the IRF for the probability of a correct response is depicted in Figure 1. Technically, one could show the IRF for an incorrect response as well, however, for a dichotomous item, this would yield redundant information.

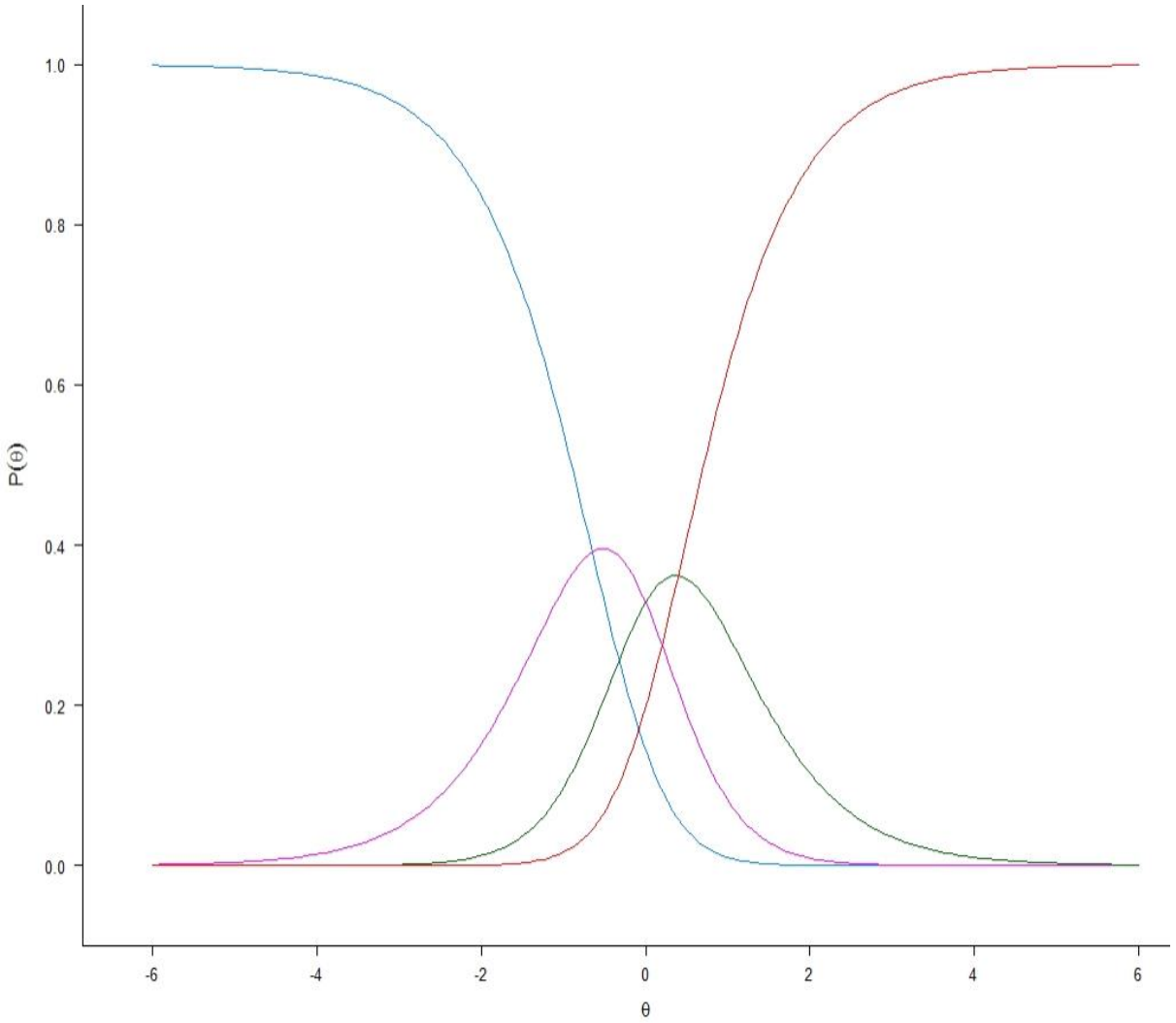
Figure 1. Example dominance item response function for a dichotomous item (a = 1.25, b = -0.77)



Unidimensionality, local independence, invariance, and a specific form of the IRF underlie the assumptions of most IRT models (Meijer & Baneke, 2004; Weekers & Meijer, 2008). Unidimensionality refers to the measurement of only one latent trait (i.e., responses do not depend on more than one underlying latent construct). When the observed item response is not dependent on other item responses on the test, for participants of equal trait levels, local independence holds, and suggests that the latent trait is what relates items to one another. Additionally, IRT models assume that the IRFs produce specific shapes depending on the IRT model. When fitting IRT models to the data, it is critical to distinguish which IRF best characterizes the relationship between the probability of a response and the latent trait level (Weekers & Meijer, 2008). *One of the observable differences between dominance and unfolding IRT models is the shape of the IRF functions for each.*

Dominance IRT Models. Dominance IRT models assume a cumulative or dominance response process is utilized, therefore as the trait level increases, so does the probability of item endorsement. With this characteristic, item response functions are monotonically increasing, as depicted in Figure 1 for a dichotomous item. For polytomous items, unique IRFs for each possible item outcome are specified. A pictorial representation of the IRFs under dominance IRT models for items with four response options are displayed in Figures 2.

Figure 2. Dominance item response function for polytomous (4-category item)



Two types of IRT models include parametric and nonparametric models. With parametric models, parameters for both items and persons are considered. Nonparametric IRT models are more flexible in that they do not consider item parameters and do not specify an exact shape for the IRF, only that it is monotonically increasing. For example, the Mokken's nonparametric Model for Monotone Homogeneity (MH; Mokken, 1971) assumes that the IRFs are non-decreasing (along with other assumptions such as unidimensionality and local independence of items). This is less strict than the parametric IRT assumption that a strict mathematical form underlies the IRFs. With the MH model, IRFs are not parametrically defined and orderings of persons are based on the number-correct true score from classical test theory (Sijtsma & Meijer, 1992). The double montonicity model adds the assumption of having invariant item ordering with non-intersecting item response functions. Although parametric IRT models are more common, nonparametric models are starting to become more popular (Sijtsma and Ark, 2017; Meijer & Baneke, 2004; Sijtsma & Molenaar, 2002).

Popular parametric, unidimensional IRT models include the one-parameter logistic (1PL) model (Rasch, 1960), the two-parameter logistic (2PL) model (Birnbaum 1957; as cited in Hambleton & Swaminathan, 1985), and the three-parameter logistic (3PL) model (Birnbaum, 1968; as cited in Baker, 2001). The 3PL model (Birnbaum, 1968; as cited in Baker, 2001) includes three item parameters (item discrimination, item difficulty, and guessing) represented by *a*, *b*, and *c*, respectively:

$$P(\theta) = c + \frac{(1-c)}{1+e^{-a(\theta-b)}} \tag{11}$$

The *c*-parameter is the probability of getting an item correct by guessing or chance. As mentioned previously the *b*-parameter represents the difficulty of an item. For a 3PL model, *b* is equal to the ability or trait level that corresponds with a (1+*c*)/2 chance of getting the item

correct. When no guessing is present ($c = 0$), the b parameter is equivalent to the θ value where respondents have a 50% chance of getting the item correct. The 2PL model is represented when $c = 0$. The 1PL model occurs when one average discrimination parameter is estimated for all items.

When items are polytomous, intermediary step functions are modeled. Step functions are defined by modeling the transition, or "stepping" to successively higher score categories. Four popular approaches for defining step functions include adjacent category, continuation ratio, cumulative, and nominal (Penfield, 2014). Within the adjacent category approach is the partial credit model (PCM; Masters, 1982), which models step functions as the probability of success at the adjacent k^{th} step as specified by the Rasch model. The generalized partial credit model (GPCM; Muraki, 1992) also uses the adjacent category approach to defining step functions, but the 2PL model is used and an item-level discrimination parameter is estimated.

GPCM. As briefly mentioned above, the GPCM uses the adjacent category approach to defining step functions, and the 2PL model is used which allows for the estimation of an item-level discrimination parameter. The item-step response function (ISRF) using the GPCM is given by

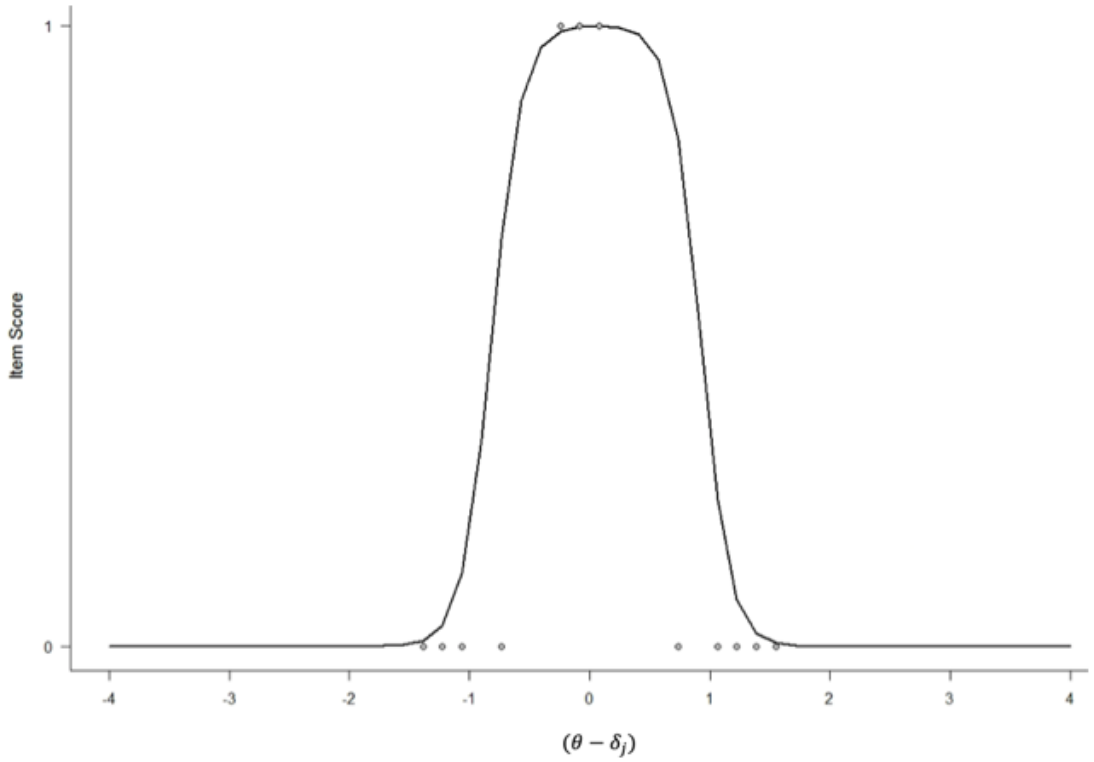
$$P(Y_j = l|\theta_n) = \frac{\exp\{\sum_{k=1}^{l}[a_j(\theta - b_{jk})]\}}{1+\sum_{r=1}^{m}\{\exp\sum_{k=1}^{r}[a_j(\theta - b_{jk})]\}} \quad (12)$$

where Y_j is the observed response of person n (with ability or trait level, θ_n) to item j, $k = 1, 2, \ldots, l, \ldots, L$ represents the steps from one response category to the next adjacent score category (e.g., if $l = 2$ for person n, this means person n was successful in passing the second step), a_j is the discrimination parameter, and b_{jk} is the difficulty parameter or location parameter of the k^{th} step, and $r = 1, 2, \ldots, m$ represents the total m exponent terms in the denominator.

Unlike dichotomous items, the denominator includes an exponent term for each step (i.e., *m* corresponds with the total number of steps for the item).

Unfolding IRT Models. Unfolding models assume an ideal-point response process, taking into account the disparity between person and item locations on the underlying continuum for the construct being measured. In contrast to dichotomous IRF's for dominance IRT models (Figure 1), unfolding models yield peaked IRFs. Figure 3 provides a visual representation of an IRF of a dichotomous item for an unfolding model. Note that instead of the probability increasing as the trait level increases (as seen in Figure 1 for dominance models), the IRF for the unfolding model peaks when the difference between the trait level (θ) and item difficulty (δ_i) is equal to zero. As the respondent's trait level gets farther away from the item's location (either higher or lower), the probability of endorsing the item decreases. In other words, respondents may not endorse the item for one of two reasons: either the respondent disagrees because they display a trait level lower than required by the item, or the respondent disagrees because their theta is higher than the item's location.

Figure 3. Item response function for dichotomous item under GGUM (unfolding)



At the root of most unfolding models, is the parallelogram model suggested by Coombs (1964):

$$Y_{nj} = 1 \; if \; \left|\theta_n - \delta_j\right| \leq \; \tau, \tag{13a}$$

$$Y_{nj} = 0 \; if \; \left|\theta_n - \delta_j\right| > \; \tau, \tag{13b}$$

where Y_{nj} is the response of person n, $n = 1, 2, \ldots, N$, to item j, j= 1, 2, …, J,

θ_n is the ability or trait level of person n,

δ_j is the location of item j on the underlying continuum for the construct, and

τ is the threshold that determines the maximum distance between θ_n and $\delta_i$where the respondent will still answer with a "positive" response ($X_{nj} = 1$). The limitation to this model is that it is deterministic in nature (the probabilities of a "positive" response can only be 0 or 1). This may be too restrictive for ordered response options of polytomous items. Johnson and

Junker (2003) give the example of having four ordered stimuli (A, B, C, and D) where, using Coomb's deterministic model, there would only be 11-item response patterns possible because the triplet 1, 0, 1, would not be possible (i.e. a person could not agree with A and C, but not B). Hoijtink (1991) suggested a probabilistic version for the IRF of Coomb's (1964) model:

$$P(X_{nj} = x_{nj}|\theta_n, \delta_j) = f(|\theta_n - \delta_j|, x_{nj}) \quad (14)$$

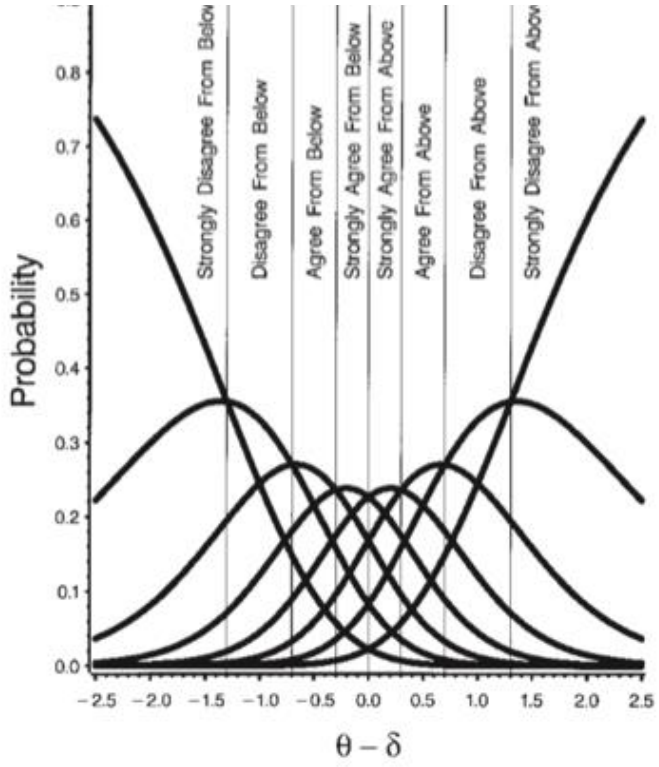
Using a probabilistic model with stochastic parameterizations of the IRF makes it possible for researchers to make statistical inferences about person and item parameters (M. S. Johnson & Junker, 2003; Sgammato, 2009).

A number of probabilistic unfolding models have been developed over the past few decades. Assumptions of these models include conditional independence, unidimensionality, and non-monotonic (unimodal) item response functions (M. S. Johnson & Junker, 2003). Some common unfolding models include the Squared Logistic Model (SLM; Andrich, 1988), PARELLA model (Andrich, 1988), Hyperbolic Cosine Model for dichotomous data (HCM; Andrich & Guanzhong Luo, 1993), General Hyperbolic Cosine Model for polytomous data (GHCM; Andrich, 1996), Graded Unfolding Model (GUM; Roberts & Laughlin, 1996), and the generalized graded unfolding model (Roberts et al., 2000). Several aspects of the GGUM make it a popular choice, and rationalize its selection for this investigation. First, it can be used with both polytomous and dichotomous data, and the discrimination parameters are allowed to vary across items. The thresholds for each response option are also allowed to vary across items. The freedom for variability among the discrimination and threshold parameters create a flexibility for the IRFs under this model to take on a wide range of shapes (Stark et al., 2006). Further, the GGUM package in R (Tendeiro & Castro-Alvarez, 2019) makes it an accessible option for researchers. Although many unfolding models have been used for ideal point data, the GGUM

has been recognized as the most popular choice for applied studies in the non-cognitive field (Joo et al., 2019).

The Generalized Graded Unfolding Model (GGUM). The GGUM uses three types of item parameters to describe each IRF. The location parameter (δ_j) describes where each item (*j*) lies on the underlying latent continuum for the construct being measured and can be depicted as the maximum point on the single-peaked IRF for a dichotomous item under the GGUM. The discrimination parameter (α_j) measures the degree to which the item discriminates between persons. The GGUM also incorporates a subjective response category threshold (τ_j) parameter. Because the GGUM assumes two possible unobserved reasons for endorsing a certain response category, each observed response category (ORC; Roberts et al., 2000) is unfolded to incorporate two subjective (unobserved or latent) response categories. Thus, a four-category item (strongly agree, agree, disagree, and strongly agree) would have seven thresholds that separate the eight subjective response categories (SRCs; Roberts et al., 2000). An illustration of the SRC probability functions for a hypothetical four-category item is given by Figure 4. In the figure, the x-axis is defined by the signed distance of the person location (θ) from the item location on the underlying continuum. The hypothetical item has a α_j parameter equal to 1.0 and the τ_{jk} for each step function, *k*, equal -1.3, -.7, -.3, .0, .3, .7, and 1.3. The seven vertical lines represent the thresholds (τ_{jk}), or points of intersection that divide the possible responses into seven subjective response categories. The most likely SRC for each interval is labeled in the figure. As the discrimination parameter increases, the probability for the corresponding most likely SRC will increasingly "dominate" within each interval (Roberts et al., 2000).

Figure 4. The subjective response curve probability functions for a hypothetical four-category item from Roberts et al., 2000, p. 5.



Under the GGUM proposed by Roberts et al. (2000), each subjective response follows the generalized partial credit model (GPCM; Muraki, 1992). As mentioned above in the dominance IRT Models section, the GPCM uses the adjacent category approach to defining step functions, and the 2PL model is used which allows for the estimation of an item-level discrimination parameter. The ISRF using the GPCM is given by Equation 12. Ultimately, the model must be defined in terms of observed responses. Because the two subjective responses associated with an observed response are mutually exclusive, the probability that a person will respond using a specific observed response category is equal to the sum of the probabilities for each of the subjective responses. This is illustrated with the following equation where Z_j is the observed response to item j, and $z = 0, 1, 2, \ldots, C$ with $z = 0$ corresponding to the strongest level

of disagreement and $z = C$ corresponds with the strongest level of agreement (C is equal to the number of ORC minus 1), Y_j is the subjective response of person n to item j, $y = 0, 1, 2, \ldots, M$ such that $y = 0$ corresponds to the strongest level of disagreement from *below* the item and $y = M$ corresponds to the strongest level of disagreement from *above* the item (M is the total number of SRCs minus 1; $M = 2C+1$),

$$P(Z_j = z|\theta_n) = P(Y_j = z|\theta_n) + P[Y_j = (M - z)|\theta_n]. \quad (15)$$

In other words, if an item had four observable response categories ($C = 3$ and z = 0, 1, 2, 3), then the item would have eight subjective response categories ($M = 7$). Using the above equation, the probability of choosing the second observed response category (e.g. "disagree"; $z = 1$), will be equal to the sum of the probabilities of choosing the subjective response "disagree from below" ($z = 1$) and the subjective response "disagree from above" ($M - z = 6$). To visualize this, see Figure 5 (SRC probability functions for a four-category item) and note that passing the first threshold yields the response "disagree from below", and passing the 6th threshold yields "disagree from above".

Under the GGUM, the threshold parameters, τ_{jk}s, are symmetric about the point $(\theta_n - \delta_j)$ which implies that participants are just as likely to agree with an item that is located h units below their personal location on the continuum as they are to agree with an item located h units above their location. Roberts et al. (2000) offer the following identity that describes this relationship:

$$\sum_{k=0}^{z} \tau_{jk} = \sum_{k=0}^{M-z} \tau_{jk} \quad (16)$$

Hence, the formal equation for the GGUM is given by incorporating the identity above (Equation 16) into Equation 15 (sum of mutually exclusive probabilities):

$$P(Z_j = z|\theta_n) = \frac{\exp\{\alpha_j[z(\theta_n - \delta_j) - \sum_{k=0}^{z} \tau_{jk}]\} + \exp\{\alpha_j[(M-z)(\theta_n - \delta_j) - \sum_{k=0}^{z} \tau_{jk}]\}}{\sum_{w=0}^{C}\{\exp\{\alpha_j[w(\theta_n - \delta_j) - \sum_{k=0}^{w} \tau_{jk}]\}\} + \sum_{w=0}^{M}\{\exp\{\alpha_j[(M-w)(\theta_n - \delta_j) - \sum_{k=0}^{w} \tau_{jk}]\}\}} \quad (17)$$

Comparison of Dominance and Ideal Point Models

Table 4 outlines a few basic differences between unfolding and dominance IRT models including underlying response process, model examples, item characteristics, and examples of items that would be expected to fit well under the corresponding model.

Table 6. Comparisons between dominance and unfolding IRT models

IRT Model	Underlying Response Process	Model Examples	Item Characteristics	Item examples
Dominance IRT	*Cumulative/ Dominance* (the higher your theta, the more likely you are to agree/get correct/ etc.)	Rasch, 1PL, 2PL, 3PL, 4PL, GRM, GPCM, PCM, NRM	Response function: Logistic/ogive/monotonic	My social skills are at least as good as those of an average person. (example from Cheryshenko, et al., 2007)
Unfolding IRT	*Ideal-point* (takes into account the disparity between person and item locations on the underlying continuum)	GUM, GGUM, SLM, HCM, GHCM	Response function: Single-peaked, non-monotonic	a) My social skills are about average. (example from Cheryshenko, et al., 2007) b) Abortion is basically immoral except when the woman's physical health is in danger. (Roberts at al., 1999) c) Although I try to keep everything in its place, it does not always work for me (Weekers & Meijer, 2008)

Weekers and Meijer (2008) (also see Stark et al., 2006) highlight three important differences between dominance and ideal point approaches. First, under the dominance approach, scales are mostly constructed with items that are either slightly negatively or positively worded, or *extremely* negatively or positively worded. This is because constructing scales under the

dominance approach hinges on having high item-total correlation, high internal consistency reliability, and are essentially unidimensional with high item factor loadings. Because dominance IRT models assume monotonically increasing IRFs, they tend to not include neutrally worded items. For example, a neutrally worded item like, "My ability to process my emotions is about average," will most likely result in a non-monotonic IRF because people with trait levels near the middle of the continuum will have the highest probability of endorsing this item. Under a dominance model, neutral-worded items will have a single-peaked item information curve, while under an unfolding model, the information curve may be double-peaked. The spread of information across a latent continuum is greater for unfolding models than dominance models when neutral items are used. If scales using a dominance model include more extreme items, they will tend to have higher precision at the extreme ends of the continuum.

The second difference noted by Weekers and Meijer (2008) is the notion that a dominance model may be thought of as a special case of ideal point models where the ideal point is located at positive or negative infinity. Unfolding models have been considered more general models than dominance models and thus may be less prone to misspecification (Weekers & Meijer, 2008; Zampetakis, 2010). Thirdly, positively and negatively worded items are commonly used in non-cognitive measures and the use of reverse scoring is not necessary when using unfolding models. This is because it is the absolute distance from theta to item location that matters for estimating ideal point responses.

The dominance approach to modeling item responses is more appropriate in many settings. The most commonly cited context suitable for dominance models is in cognitive testing. For example, for an item that asks examinees to choose the response that gives the most closely related synonym to the provided word, one would reasonably expect examinees with a higher

vocabulary ability to have a higher probability of getting the item correct. There is, however, the case where respondents have such a high ability level, they may give an "incorrect" response, sometimes termed as a "creative" response. Cognitive items more prone to this situation may result in non-monotonic IRFs, though it is not as common. Findings from a study by Zampetakis et al. (2015) suggest that a dominance response process may also underlie participants self-reports of anticipated affect (what people predict about their affective responses to events in the future; Loewenstein & Lerner, 2003, as cited in Zampetakis et al., 2015). Results from the study revealed better fit of the graded response model (dominance) than the GGUM (unfolding) for data collected from participants regarding self-reported anticipated positive and negative affect. However, this is not always the case. The following section highlights a few studies where unfolding models provided a better framework for a scale than a dominance model framework.

Research Comparing Dominance and Unfolding Models

An empirical study by Chernyshenko et al. (2007) highlighted the advantages of using ideal point methodology over the dominance response process approach for scale construction for the order facet of conscientiousness. The authors constructed three different scales using three different underlying frameworks: 1) traditional classical test theory (CTT), 2) dominance IRT, and 3) ideal point IRT. The data collected using the three scales were scored using a dominance model (2PLM) and an unfolding model (GGUM). The IRT scores for the scale constructed using the ideal point process were highly correlated with those from the dominance IRT (r = .92) and CTT (r = .88) scales, providing evidence that including neutral items did not reduce the validity for the ideal point order scale. Also highlighted in the results was the flexibility demonstrated by using the ideal point model (GGUM) to score items from the scale constructed using the dominance IRT approach. Results revealed that using the GGUM on the

data from the dominance scale yielded IRT estimates highly correlated with the estimates using the 2PL model (r = .97). Conversely, using the 2PL model to fit the data collected from the ideal point order scale resulted in theta estimates that differed from those using the GGUM, especially on the lower (theta range from -3 to -1) and upper (theta range from +1 to +3) extremes of the continuum where correlations between the sets of scores were .33 and .21. Chernyshenko et al. (2007) conclude that using the ideal point process allowed for greater flexibility (wider range of item options in scale construction, e.g., neutral items) while maintaining validity of personality test scores.

Stark et al. (2006) compared the model fit of two ideal point process IRT models and two dominance IRT models using cross validation samples of 13,059 respondents who took the Sixteen Personality Factor Questionnaire (16PF; Conn & Ricke, 1994, as cited in Stark et al., 2006). The two ideal point process IRT models involved in the study included the parametric GGUM (Roberts et al., 2000) and Levine's nonparametric maximum likelihood formula scoring (MFS) model with ideal point constraints (Levine, 1984). The two dominance IRT models were the parametric 2PL (Birnbaum, 1968; as cited in Baker, 2001) and the nonparametric MFS with dominance constraints (as Stark et al., [2006] note, the MFS allows researchers to impose different mathematical constraints on the IRFs to test various assumptions about responding processes). The graphical comparison of fit plots for the IRFs using each type of model and the direct comparison of chi square statistics for items, item pairs and item triplets were used to examine model fit. Nine of the sixteen PF subscales contained items with non-monotonic IRFs, with four of those subscales containing four or more non-monotonic items (Liveliness, Sensitivity, Abstractedness, and Privateness). The majority of folding took place at the extreme ends of the continuums. When all items were monotonic, both dominance and ideal point process

models had similar fit. This concurs with Roberts et al. (1999) who also note that ideal point process and dominance models produce similar IRFs for extreme items that also have high item-total correlations, as Stark et al. (2006) mention was the case for most of the 16PF items. However, for seven of the nine subscales with items that failed to pass the montonicity assumption, MFS with ideal point constraints provided the better fit. The authors concluded that the ideal point process models fit the data for the majority of the16PF subscales (which were constructed under the assumption of a dominance response process) as well, if not better than the dominance IRT models used in the study.

Weekers and Meijer (2008) extended the work of Stark et al. (2006) and Chernyshenko et al. (2007) by comparing the fit of unfolding and dominance models to data from two personality trait inventories constructed using the ideal-point process (Order Scale; Chernyshenko et al., 2007) or the dominance response process (NPV-J; Luteijn et al., 2005). One nonparametric and one parametric model for each approach (dominance and unfolding) was utilized to analyze the data. Item analysis revealed the presence of some single-peaked items in both types of scales. The authors note that these are usually neutrally worded items and can influence the fit of an IRT model to the data, which led to the suggestion that dominance response process scales may be useful for scales that consists of extreme statements in attempt to measure extreme behavior, as in psychopathology scales. On the other hand, the authors suggest general personality self-report inventories that intend to measure people on a greater spread of the latent continuum where neutral items are necessary, may be best described by an unfolding model. Similar to conclusions made by Stark et al. (2006), Weekers and Meijer (2008) note that misspecification of the underlying response process may have crucial effects on conclusions drawn from ordering persons by latent trait scores. They illustrated this by plotting the estimated trait scores ($\hat{\theta}$) using

the unfolding and dominance models. The correlations between the trait scores using both models were high for the scale constructed under the dominance response process (r = .988) and the scale that was constructed under the ideal point process (r = .971). However, for the ideal point process scale (where items had both monotonically increasing and single-peaked IRFs), the scatterplot showed a departure from the diagonal line for a compelling number of people at the higher end of the continuum. This indicates that people, especially those located at the upper end of the continuum for these constructs, are ordered differently using dominance versus unfolding models. Thus, if decisions or cutoffs are made using the top 5%, for example, conclusions may be inappropriate.

In the last few decades, researchers have become somewhat more aware of the applicability of unfolding models to non-cognitive data. Researchers have used unfolding models to successfully describe non-cognitive data such as assessment for creativity (Zampetakis, 2010) using the Gough's Creative Personality Scale, conscientiousness (Carter et al., 2014), personality inventory of self-judgement on the order-facet (a feature of conscientiousness; Chernyshenko et al., 2007; Weekers & Meijer, 2008), 16 personality factor subscales (Stark, 2006), control preferences in medical contexts (Control preferences Scale; Degner et al., 1997), attitude and affect constructs (LaPalme et al., 2018), censorship data (C.-W. Liu & Wang, 2019), and job satisfaction (Carter & Dalal, 2010). Nonetheless, dominance models are much more widely used than unfolding models. This may be due to the increased complexity of unfolding models or because they are still in the earlier stages of applied research (Sgammato, 2009; Stark et al., 2008). As such, further research to understand their application with non-cognitive data, and the use of other data management procedures such as data quality assessments are potentially useful. The next sections include a discussion of the assessment of model fit with unfolding and

dominance models and the application of person-fit statistics to assess data quality with data using these two frameworks.

Assessment of Dimensionality and Model Fit. The underlying assumption of unidimensionality for probabilistic unfolding and dominance IRT models can be assessed using a principal components analysis (Roberts et al., 2000). Davison (1977, as cited in Roberts et al., 2000) demonstrated that two main principal components underlie ideal point responses from unfolding models which has been supported by GGUM simulation studies that have revealed this structure as well (Roberts et al., 2000). Item-level communality estimates from the first two principal components can be used in dimensionality assessment, where Roberts et al. (2000) used cut-off of greater than or equal to 0.3 based on previous simulations indicating unidimensionality ("for analysis purposes").

Item fit for dominance polytomous IRT models can be evaluated using a likelihood ratio χ^2 statistic. This is usually computed using an $R \times C$ contingency table, where *R* is the number of groups formed by dividing respondents into equal intervals based on their θ_j values and *C* is the number of response categories for the item. The observed versus expected frequencies are then compared within each cell of the contingency table. For comparison purposes, it is suggested that the number of groups, *R*, be constant across items. For unfolding models, however, expected frequencies can be small due to the fact that respondents are only expected to endorse items that are located near their trait level on the underlying continuum. Because the distribution of the χ^2 statistic becomes suspect if the expected frequency in any cell of the contingency table is small (say, less than 5), and because correcting this problem by collapsing groups would be difficult to do consistently across all items, using the statistic as described above for dominance IRT models can be problematic for unfolding models. Item fit in unfolding

models can be evaluated using chi square statistics by dividing participants into groups of equal size based on the signed distance of $\hat{\theta}_j - \hat{\delta}_i$ for every item-person pair. The expected and observed responses for each item can be averaged for each group and plotted as a function of the group $\hat{\theta}_j - \hat{\delta}_i$ mean. Model-data misfit can be observed across the latent continuum where large differences between observed and expected frequencies exist on the plot. It may also be useful to compute correlations between observed and expected frequencies.

The formula for the χ^2 statistic is computed using the observed frequencies of item *j* response option *z* (O_{jz}) and the expected frequencies (E_{jz}) based on the estimated item parameters and distribution of abilities.

$$\chi_j^2 = \sum_{z=0}^{C} \frac{(O_{jz} - E_{jz})^2}{E_{jz}}, \quad (18)$$

with $E_{jz} = N \int P_{jz}(\theta)\phi(\theta)d\theta$, where $\phi(\theta)$ is the standard normal density function. This equation is written for single items, but is often generalized to apply to pairs of items (doublets) and triples of items (triplets) as this has been shown to be more reliable estimates of model (mis)fit that with single items (Drasgow et al., 1995). The adjusted χ^2 ratio (χ^2/df) can also be computed to adjust the sample size to 3,000 (Drasgow et al., 1995):

$$\chi^2/df = \frac{3{,}000(\chi_j^2 - df)}{N} + df \quad (19)$$

The degrees of freedom in Equation 19 depend on the number of singlets, doublets or triplets used. It has been suggested that values of χ^2/df larger than 3 be considered heuristically indicative of model misfit (Tendeiro, 2017).

If models are nested, the likelihood ratio statistic can test the incremental fit of a constrained model to a less constrained model. For example, the GUM is a constrained version of the GGUM, where the discrimination parameter is constrained to 1 and the threshold

parameters are constrained to be equal across all items. Because the dominance IRT model (GPCM) and the unfolding IRT model (GGUM) used in this study are not nested, this statistic cannot be used for their comparison. Instead, information criterion based statistics, Akaike Information Criterion (AIC; Akaike, 1974) and Bayes Information Criterion (BIC; Schwarz, 1978) are used to compare relative model fit. These indices take into account the number of parameters being estimated and are computed as follows:

$$AIC = 2k - 2\ln(\hat{L}) \tag{20}$$

$$BIC = k\ln(n) - 2\ln(\hat{L}) \tag{21}$$

Here, k is the number of parameters estimated by the model, $\hat{L}$ is the maximized value of the likelihood function for the model, and n is the sample size. Lower AIC and BIC values indicate better fit.

Part III: Detection of Aberrant Responding with Ideal Point Responses

Person-fit Analysis Under Unfolding Models

While there exists an immense body of literature regarding person-fit statistics, it is unclear how they perform under an unfolding framework. To date, only one publication has reported on assessing person-fit in the unfolding model context. Tendeiro (2017) conducted a simulation study to understand how the modified log likelihood person-fit statistic for polytomous items ($l^*_{z(p)}$) works under the generalized graded unfolding model (GGUM; Roberts et al., 2000; Roberts & Laughlin, 1996). Type I error and power were evaluated for detecting extreme and middle response style patterns under varied conditions for four factors: Scale length (*I*) with four levels (10, 20, 40, 100 items), number of observed response categories with three levels (4, 6, 8), the proportion of simulees with aberrant response vectors with three levels (*AbN* = .05, .10, .20), and the proportion of aberrant item scores (*AbI*) within the aberrant response

vectors with three levels (.10, .20, .25). Half of the aberrant responding simulees [(*N* X *AbN*)/2] had response vectors that contained a proportion of aberrant item scores (*AbI*) that reflected extreme responding while the other half reflected middle responding. Extreme response patterns were generated by replacing randomly selected middle responses from the randomly selected "aberrant simulees'" response vectors, with the closest extreme response option with probability 1 (e.g., if an aberrant responding simulee had a response of '1' to an item with four observable response categories (0, 1, 2, and 3) and the strongest level of agreement is 3, then a '1' would be changed to a '0' because that is the closest extreme response option to the middle response of '1'). Conversely, middle response patterns were generated by replacing randomly selected extreme responses with the corresponding middle response with probability 1 (e.g., a response of '0' would be changed to '1'). Results for the detection rates of extreme response patterns revealed low power with no more than 30% of the aberrant response vectors detected for 92 of the 108 conditions. Only for large numbers of items (40 or 100) and answer options (8 response options), with *AbI* = .20 or .25 did the detection rates increase beyond .30 (range .32 to .90). Similar to other studies, as the proportion of aberrant respondents increased, detection rates decreased (Karabatsos, 2003). The mean detection rate for middle responding (*M* = .45; quartile1 = .17 and quartile 3 = .72) was higher than the mean detection rate for extreme responding (*M* = .17; quartile 1 = .06 and quartile 3 = .22). The author hypothesizes that the discrepancy between performances of the statistic for extreme and middle responding may be due to the number of extreme item scores being much higher than the number of middle item scores (ratio ranged from 1.40 to 1.89). Because of this saturation, extreme responding was less likely to be detected as "unexpected" or aberrant.

Two four-way ANOVAS revealed that the main effects of *I*, the number of response options (*C)*, *AbI*, and *AbN* all had moderate to strong effects on detection rates for both middle and extreme responding. Detection rates increased with *I*, *AbI*, and *C*, and decreased, as previously mentioned, with *AbN*. The effect of increasing *C* on detection rate for extreme responding was especially strong, which the author notes may be explained by other research indicating the evidence for extreme responding is stronger when the middle and extreme options are further apart. Type I error rates for the $l_{z(p)}^{*}$ statistic across all conditions were conservative, averaging .03 (*SD* = .01) compared to the nominal rate of .05. Although power for detecting extreme responding was low using $l_{z(p)}^{*}$ under the GGUM, the author posits that the detection method showed promising results for detecting middle response style patterns in some conditions.

Emons (2008, p. 242) mentioned that a "topic for further research is applications of person-fit methods to nonmonotonic items." Tendeiro (2017, p. 56) states: "The lack of published research concerning person-fit analytical approaches suitable to unfolding models is striking." Even a few years after the publication of Tendeiro's work with $l_{z(p)}^{*}$ and the GGUM, research is still severely lacking in the area. Information on how other person-fit statistics perform, as well as the detection of other types of aberrant responding (other than middle and extreme responding) under an unfolding model context remains unknown.

CHAPTER 3

METHODS

To investigate aberrant responding with underlying dominance and unfolding response processes, three simulation studies were conducted that build upon prior dominance response model research to include conditions with the unfolding model. The components of the studies that advance the literature include 1) providing insight on GGUM model-data fit when aberrant data is systematically entered, 2) examining how nonparametric person-fit statistics (H^T, $U3^P$, G_N^P, and G^P) perform under the framework of the GGUM, and 3) investigating how the performance of parametric person-fit statistics ($l_{z(p)}$ and $l_{z(p)}^*$) is impacted by misspecifying the GPCM to GGUM data and the GGUM to GPCM data. All three studies are based on the datasets generated in Study 1.

Study 1: *An Investigation of the Effects of Aberrant Responding on Model-Fit Assuming Different Underlying Response Processes*

Simulation Factors

The first study focused on simulated IRT data for 6-point items and a fixed sample size of 1,000. Various types, proportions, and degrees of aberrant responding were incorporated. Four types of aberrant responding were considered in the study, including two types of insufficient effort: random responding and longstrings, and two types of response styles: extreme response style (ERS) and midpoint response style (MRS). Additionally, the four aberrant response types were combined in a 'mixed aberrant response type' condition to simulate realistic situations where a sample may be composed of several types of aberrant responders including those due to insufficient effort and those due to response styles. Response data were simulated using three different proportions of aberrant responders (*AbN*): .04, .10, .20. Furthermore, aberrant response

vectors were simulated using different proportions of aberrant responses to the items (*AbI*). That is, each response vector that was classified as aberrant was generated so that either 20%, 40%, or 60% of the *items* within the vector were not as expected (aberrant). The levels for both *AbN* and *AbI* are similar to the conditions used in Tendeiro (2017) for comparison purposes in studies 2 and 3 when person fit statistics are implemented. Two test lengths of 20 and 40 items were considered. Data were generated using two data response processes: a dominance response model and an ideal-point (unfolding) response model. All datasets were also analyzed using a dominance and an unfolding model. A total of 2 (data generating mechanisms: GGUM and GPCM) × 2 (applied model fit: GGUM and GPCM) × 3 (proportion of aberrant responders in the sample, *AbN*) × 3 (proportion of aberrant responses in response vectors, AbI) × 2 (test lengths) × 5 (types of aberrant responding and response styles) = 360 fully crossed conditions. Each condition was replicated 100 times. All code for generating and estimating model parameters, model fit statistics, and person-fit statistics was written in R (R Core Team, 2020) and is available on the Open Science Framework (https://osf.io/) as well as listed in the Appendices.

Models

Two models were chosen to generate and fit the data according to either an underlying dominance or ideal-point (unfolding) response process. For the unfolding model, the GGUM was selected. It incorporates a subjective response category threshold (τ_j) parameter and two subjective response categories (from above or below) for each response category. The GGUM allows the item discrimination as well as category threshold parameters to vary. The GGUM is becoming increasingly popular in applied research, and accessible software for estimating the model is readily available. This investigation used the GGUM proposed by Roberts et al. (2000),

where each subjective response follows the generalized partial credit model (GPCM; Muraki, 1992). For this reason, the GPCM was chosen to model the polytomously-scored dominance data in order to maximize comparability of results.

Item Parameter Generation

The GenData.GGUM function in R was used to generate all model (item and person) parameters as well as the item scores. Procedures for generating ideal-point data under the GGUM closely followed those taken by Tendeiro (2017). The true item discrimination parameters (α_j) were randomly sampled from a uniform distribution with the interval (0.5, 2.0). The true item location parameters (δ_j) were randomly sampled from the standard normal distribution truncated between -2.0 and 2.0, since extreme values of δ_i may sometimes lead to issues of low accuracy and variability of MML estimates under the GGUM (Roberts & Thompson, 2011 as cited in Tendeiro, 2017). The true location of the threshold parameters (τ_{jk}) (relative to the location of the *i*th item) are typically constrained so that

$$\tau_{j(C+1)} = 0 \tag{22}$$

and

$$\tau_{jz} = -\tau_{j(M-z+1)} \; for \; z = 1, \ldots, C, \tag{23}$$

where *M* represents the number of subjective response thresholds and *C* represents the number of item response steps. For example, if an item had 6 observable response categories (*C* = 5; 12 subjective response categories; 11 subjective response thresholds, *M* = 11), the first constraint illustrates that the middle subjective response threshold ($\tau_{j(C+1)} = \tau_{j6}$) would equal zero. And the second constraint illustrates that the other subjective response thresholds are symmetric about that middle threshold. So, for the score of *z* = 2 on a 6-point scale,

$$\tau_{j2} = -\tau_{j(M-z+1)} \tag{24a}$$

$$\tau_{j2} = -\tau_{j(11-2+1)} \tag{24b}$$

$$\tau_{j2} = -\tau_{j10}. \tag{24c}$$

This example demonstrates how the 2nd and 10th subjective thresholds are symmetric about the 6th subjective threshold for an item with 6 observable response categories.

Using the GenData.GGUM function, the true locations of the threshold parameters (τ_{jk}) (relative to the location of the *j*th item) were recursively generated using the following procedure also used in Roberts et al. (2002). First, the τ_{jk} parameter was randomly drawn from a uniform distribution ranging from (-1.4, -.4). Next, the locations for the thresholds were computed using the recursive equation (Roberts et al., 2002):

$$\tau_{jk-1} = \tau_{jk} - .25 + e_{jk-1}, \text{ for } k = 2, 3, \ldots, C, \tag{25}$$

where e_{jk-1} represents a random error term generated from a $N(0, .04)$ distribution.

To generate data under the GPCM, item category thresholds, d_{jk}, for step *k* of item *j* were simulated. First, uncentered d_{jk} parameters were simulated by taking the sequential cumulative sum of five numbers drawn from a random uniform distribution between .3 and 1. This interval ensured that the distance between categories will not be less than .30 because if the categories are too close, some may not be chosen as often (Chalmers, 2012). For example, say five numbers were randomly drawn from a uniform distribution between .3 and 1 (.8, .4, .5, .6, .7). Then the numbers were transformed to be the sequential cumulative sum (.8, 1.2, 1.7, 2.3, 3.0). Next, the mean for the set of sequential cumulative sums was subtracted from each number in the set [e.g., (.8-1.8), (1.2-1.8), (1.7-1.8), (2.3-1.8), (3.0-1.8); or (-1.0, -.6, -.1, .5, 1.2)]. This new set of numbers represented the d_{jk} parameters, or the item *j* category thresholds for step *k*. Additionally, an initial item category threshold, d_{j0}, was arbitrarily set to 0 in order for the model to be identified (Muraki, 1992). This is the recommended constraint for GPCMs. The item location

parameters (b_j) were randomly drawn from a standard normal distribution. To get the item location parameters for each step (b_{jk}), the item category thresholds were subtracted from the item location parameters using Equation 26:

$$b_{jk} = b_j - d_{jk} \quad (26)$$

By definition, item location parameters are the mean of all step location parameters (b_{jk}) for a particular item. Item parameters were generated so that the item discrimination and difficulty parameters were comparable to those generated for the unfolding data. Thus, the item discrimination parameters were sampled from a uniform distribution (0.5, 2.0), and item difficulty parameters for each item were sampled from the standard normal distribution $N(0,1)$. Once these parameters were generated, the sim_gpcm function from the PP package R (Reif & Steinfeld, 2021) was used to simulate the dominance data under the GPCM. The person parameters used for the GGUM and GPCM data were randomly drawn from the standard normal distribution.

Generation of Clean Data

Simulee latent ability/trait levels were generated by taking random samples from the standard normal distribution. Response data was then created by comparing a random number drawn from a uniform distribution (0,1), r, to the cumulative distribution of F_{jk}, where $F_{jk} = P(u_{jk}|\theta_n) + P(u_{jk-1}|\theta_n) + \cdots + P(u_{j0}|\theta_n)$, and j represents the item, k represents the step, and n represents the participant. For each replication and condition, response vectors for 1,000 simulees were generated according to the model specified.

Generation of Aberrant Responses

Once clean datasets were generated, a proportion (depending on the *AbN* condition) of response vectors were selected and replaced with aberrant response vectors. The aberrant

response vectors were created by replacing a proportion of the item scores (depending on the *AbI* condition) with aberrant scores stemming from either random responding, longstrings, midpoint response style (MRS), or extreme response style (ERS). A fifth aberrant type included a mixture of all four aberrant response types or styles. For this condition, an equal proportion of each type of aberrant response (or response style) was simulated to make up the total proportion of aberrant respondents. For example, if the condition called for 4% of the sample to have aberrant response vectors, those 40 respondents (4% of 1,000) would be made up of 10 random responders, 10 longstrings, 10 MRS responders, and 10 ERS responders. Procedures for creating the aberrant scores are described below.

Random Responders. Many researchers have investigated the sources for insufficient effort responding that lead to aberrant response vectors and low-quality data. One source involves participants who lack the motivation to provide thoughtful responses. This often results in random responding to complete the survey with the minimum cognitive effort. To simulate aberrant responses due to random responding, first, AbI% of items were randomly selected. Next, AbI% of values were selected from {0, 1, 2, 3, 4, 5}. Finally, original scores were replaced with the randomly generated scores.

Longstrings. Longstring responding occurs when a respondent answers the same way to a long string of consecutive items disregarding item content, identified by invariant response vectors (e.g., Costa & McCrae, 2008; Huang et al., 2012). To create aberrant responses due to longstrings, proportions of item scores (based on *AbI*) were replaced with an invariant set of responses. To do this, a random number was generated from a uniform distribution [0, C]. The specified proportion of consecutive items ($AbI \times number\ of\ items$) were set to equal the randomly generated number (DeSimone et al., 2018).

Midpoint Responding. When participants consistently tend to choose the middle response option over the adjacent categories, regardless of the construct being measured, they are said to exhibit midpoint response style (MRS). To create response vectors that demonstrate MRS for 6-point items, endpoint item scores and item scores adjacent to the endpoints were replaced with the corresponding midpoint response. For example, on the 6-point scale ranging from 0 to 5, items scores of 0 and 1 were replaced with a 2. Additionally, scores of 4 and 5 were replaced with a 3.

Extreme Responding. Extreme response style (ERS) is indicative of the tendency to choose the response options at the extreme categorical endpoints across multiple content areas. Although response styles such as MRS and ERS may not be considered as "aberrant" responding by all, for the purpose of this study, aberrant responding will be used loosely to describe tendencies that influence data quality by reducing the comparability of individual test scores and yield item scores that may not be expected based on true ability or trait level values (θ). To mimic responses reflecting ERS, item scores in the middle range were changed to the corresponding endpoint responses. For example, on the 6-point scale ranging from 0 to 5, 1s and 2s were changed to 0 and 3s and 4s were changed to 5.

Fitting the IRT Models

After response strings for all simulees were generated based on the GPCM and GGUM, the two IRT models were fit to the response data both before and after inputting aberrant scores. This resulted in each dataset being fit twice for each condition in the study. The GPCM was fit to the data using the MIRT function (itemtype = gpcm) in R (Chalmers, 2012) whereas the GGUM was fit to the data using the GGUM function in the GGUM package in R (Tendeiro & Castro-Alvarez, 2020). The convergence tolerance was set to .001. Using the GGUM function, item

parameters were estimated using marginal maximum likelihood (MML) algorithm from Roberts et al. (2000). Using MIRT, item parameters were estimated using maximum likelihood. Using both R packages GGUM and MIRT, person parameters were estimated using an estimated a posteriori (EAP) method. It is worth noting, that there is sign indeterminacy for location parameters for GGUM models (Roberts & Cui, 2004). Thus, for replications of GGUM fittings to data where the correlations between true and estimated location parameter values were highly negative (e.g., r = -.98), the location parameters were transformed to their opposite value by multiplying by negative one, resulting in opposite theta signs as well. Any time GPCM was fit to GGUM responses, reverse coding of items on the lower end of the continuum (lower 30% of items) was carried out. Further reverse coding was completed for items that still had negative loadings on the principal component (Tay et al. 2011; Tay and Drasgow, 2012).

Evaluating Model Fit

Chernyshenko et al. (2007) recommend that in IRT, model-data fit be evaluated by examining both the model assumptions and tests of goodness of fit. Therefore, investigating the relative model-data fit with respect to all conditions in the study was primarily two-fold. First, model assumptions for the data were checked. For the purpose of this study, the underlying response process was varied and thus, this assumption was purposely violated under certain conditions. Attention to the IRT assumption for unidimensionality was the primary focus of the first step in evaluating model fit. Second, predictions based on the estimated model and observed data were compared using statistical methods to test goodness of fit. Additionally, relative model data fit was examined using information based fit indices. Specifically, information criterion based statistics (AIC and BIC) and adjusted χ^2/df ratios for singles, pairs and triplets of items were reported (Drasgow et al., 1995). Further, quality of parameter estimates

using correlations, mean absolute deviations and bias were used to reveal any important differences in parameter recovery for each model and dataset reflecting the different response processes used in the study.

Testing the Assumption of Dimensionality. One of the first tasks in practical research is to test model assumptions. Both the GPCM and GGUM models assume that the probability of a response is a function of a single underlying latent trait (theta) or a unidimensional composite of skills. Literature regarding how to assess dimensionality with cumulative, dominance IRT models is immense. Various procedures exist for testing the assumption of unidimensionality including the modified parallel analysis (Drasgow & Lissak, 1983), an automated item selection procedure (AISP), confirmatory factor analysis, and DIMTEST (Stout, 1987). However, there is a lack of research and direction for how to assess dimensionality with unfolding models. Nandakumar et al. (2002) showed that linear principal component analyses do not work in the same way for dominance and unfolding models. The study revealed that two linear principal components arise for every single unfolding dimension. In William's thesis (2015), an (r + 1) rule is suggested for using results from a PCA on unfolding polytomous data, where r is the number of unfolding dimensions. In this study, parallel analysis is used to assess dimensionality derived from the observed simulated data with and without aberrant data present. For unfolding data, the r + 1 rule is used and if $r < 2$ it is concluded that the assumption of unidimensionality for the unfolding data is met.

Model-fit. Information criterion fit indices AIC and BIC were utilized in comparing relative data model fit because both of these indices not only consider the non-nested structure of the two models but they also factor in penalties for additional parameters in more complicated models like GPCM and GGUM. A lower AIC and BIC value indicates better fit.

Tay et al. (2011) recommend that when determining which model (unfolding or dominance) fit the data better, the doubles and triples adjusted χ^2 ratios may be used. Using the MODFIT function in R, the adjusted χ^2 ratios (χ^2/df) were computed for item doublets and triplets in this study. This function uses the equations for the adjusted chi-square ratio introduced by Drasgow et al. (1995).

Assessing Estimated Parameter Quality. Estimated parameters from the generated datasets including aberrant responders ($Data_{Ab}$) and the datasets free of aberrant responders ($Data_{clean}$) were compared to true parameter estimates. Procedures in Tendeiro (2017) were followed by looking at the bias, mean absolute deviation (MAD), and the correlation (COR) between true and estimated parameters. The following equations were used:

$$MAD = \sum_{t=1}^{T} |\hat{\gamma}_t - \gamma_t^{TRUE}|/T \quad (27)$$

$$BIAS = \sum_{t=1}^{T} (\hat{\gamma}_t - \gamma_t^{TRUE})/T \quad (28)$$

$$COR = cor(\hat{\gamma}_t, \gamma_t^{TRUE}), t = 1, \dots, T, \quad (29)$$

where γ_t is the parameter representing either $\alpha_i, \delta_i, \tau_{ik}, or\ \theta_j$ for the GGUM parameters or α_i, $b_{ik} or\ \theta_j$ for the GPCM, and *T* is the corresponding total number of that parameter. For example, *T* is equal to the number of items for α_i and δ_i. *T* equals *I* (the number of items) times *C* (the number of observed response categories minus 1) for τ_{ik}, and *T* is equal to the sample size for θ_j. The MAD, bias, and correlations were computed and averaged over all replications for each condition. Standard deviations for these averages are also reported to examine the variability in the results.

In the case of cross-fitting models to the opposing type of data (e.g., GPCM fit to GGUM data), several estimated parameters are not comparable. For example, the location parameter in dominance IRT models represents the theta associated with a .50 probability of choosing a

response. However, in unfolding models, the location parameter represents the theta associated with the highest probability of endorsement. Furthermore, discrimination parameters for non-monotonically increasing items may result in negative values. Thus, for cross fitting models, parameter quality was assessed for person scoring only. Worse model fit for cross-fitting conditions was anticipated. In Study 1, the goal is to examine these trends for the two model types. For example, if the GGUM is not as affected by aberrant data, or is able to fit dominance data reasonably well in comparison to the dominance model, this information could be useful for researchers and possibly expand the use of ideal point models in practice.

Study 2: *Performance of Nonparametric Person-Fit Statistics with Unfolding versus Dominance Response Models*

Simulation Factors

The second study examines 180 completely crossed conditions including 2 types of data generating mechanisms (dominance data modeled by the GPCM and unfolding data modeled by the GGUM) × 3 proportions of aberrant responders in the sample (*AbN*: .04, .10, .20) × 3 proportions of aberrant responses within response vectors (*AbI*: 20%, 40%, or 60%) × 2 test lengths (20, 40) × 5 (types of aberrant responding and response style conditions). No cross-fitting of models to data was relevant in Study 2 since the focus was on the performance of the nonparametric person fit statistics which do not rely on parameter estimates. Data were generated with 0 *AbN* and 0 *AbI* to obtain a baseline for "clean", non-aberrant data. This results in a total of 180 conditions replicated 100 times. Each dataset contained simulated response vectors for 1,000 respondents. All code for generating and estimating model parameters, model fit statistics, and person-fit statistics was written in R (R Core Team, 2020) and is available on the Open Science Framework (https://osf.io/) as well as listed in the Appendices.

Methods for the second study can be summarized in 5 fundamental steps:

1) Generate the "clean" data for the unfolding data under the GGUM and dominance data under the GPCM.
2) Generate aberrant data for simulated random responders, longstrings, ERS, MRS, and mixed aberrant responders.
3) Test how well the items are ordered and if this ordering is consistent across conditions.
4) Compute the person-fit statistics for each simulee and flag simulees who meet the decided criteria for having aberrant responses.
5) Calculate the type I error rates (falsely identifying a simulee as aberrant when their responses were "clean") and detection rates (correctly identifying a simulee as aberrant).

Methods for the generation of the data (steps 1 – 2) are described in Study 1. Steps 3 through 5 are detailed below.

Testing Item Ordering (Step 3)

It is good practice to test the ordering of items before drawing conclusions from nonparametric person-fit statistics that depend on invariant item ordering (Van der Ark, 2007). For example, if the overall H^T coefficient (sum of H_i^T for all participants) is less than 0.3, researchers suggest that invariant item ordering may be too unstable to be useful (Ligtvoet et al., 2010). For the current study, the overall H^T coefficient was computed for every condition and averaged across replications to assess how well and consistent the simulated items were ordered. It was hypothesized that the violations of monotonicity under the unfolding data would affect item ordering among simulees and thus adversely affect the performance of the nonparametric

person-fit statistics. An aim of the study was to investigate the degree to which type I error and power is impacted for the person-fit statistics assuming monotonicity with the unfolding versus dominance data.

Computing Nonparametric Person-Fit Statistics (Step 4)

The values for each of the four nonparametric person-fit statistics were computed of each simulee in step 4. For $U3^P$, G_N^P, and G^P, the PerFit package in R was utilized. The H^T statistic was not computed using the same package because the PerFit package only provides the dichotomous version of the H^T statistic. The H^T statistic is essentially a modified version of Mokken's (1971) H_i statistic, which uses a data matrix consisting of items as rows, and people as columns; it measures how well items on a scale order respondents according to a Guttman pattern. In fact, if the data matrix used to compute H_i is transposed, then the resulting H^T statistic measures how well people respond to items according to the Guttman scale. Thus, the statistic was computed using the Mokken package in R and transposing the data matrix so that simulees were represented by rows and items by columns.

Cutoff Criteria for Aberrant Identification. For all four nonparametric person-fit statistics, the distribution of each statistic value for simulees was examined for the clean datasets across all conditions and replications. A cutoff for each statistic was then determined by finding the value associated with the critical value for a 5% probability of a type I error. For example, for the G^p statistic, the more Guttmann errors a respondent has, the greater the G^p statistic, indicating greater person misfit. Thus, in the equation $P(G_p) \geq \text{value}_{\text{critical}} = .05$, the person-fit statistic critical value ($\text{value}_{\text{critical}}$) was used as the cutoff criteria. Several researchers have used the 5% quantile as the cutoff value (e.g., Magis et al., 2012; Emons, 2008; Tendeiro, 2017).

Evaluating Performance of Person-Fit Statistics (Step 5)

The true positive rate and true negative rates, reflecting sensitivity and specificity, are commonly used in the literature for evaluation of aberrant response behavior detection (Huang et al., 2012; Meade & Craig, 2012b; Niessen et al., 2016; Turner, 2018). In this study, both type I error (false positive) and detection (true positive) rates were computed to assess the performance of the person-fit statistics. Type I error was computed as the proportion of simulees with "clean" or non-aberrant responses that were flagged as aberrant by the person-fit statistic. Detection rates were computed as the proportion of simulees generated to have aberrant response vectors that were correctly flagged as aberrant by the person-fit statistic. Additionally, accuracy rates are summarized and reported (computed as the sum of true positives and true negatives, divided by the total sample size). All rates were computed for each condition and each replication and then averaged (raw data were not aggregated).

Study 3: *Impacts of Misspecification of Underlying Response Processes on the Performance of Nonparametric and Parametric Person-Fit Statistics*

Simulation Factors

The third and final study included 360 completely crossed conditions including 2 data generating mechanisms (GPCM and GGUM) × 2 models used in fitting the data (GPCM and GGUM) × 3 proportions of aberrant responders in the sample (*AbN*: .04, .10, .20) × 3 proportions of aberrant responses within response vectors (*AbI*: 20%, 40%, or 60%) × 5 types of aberrant responding and response style conditions (random responders, longstring, MRS, ERS, mixed) × 2 test lengths (20, 40). This resulted in a total of 360 conditions. A total of 100 replications were generated for each condition. All code for generating and estimating model parameters, model fit statistics, and person-fit statistics was written in R (R Core Team, 2020)

and is available on the Open Science Framework (https://osf.io/) as well as listed in the Appendices.

Methods for the third study can be summarized in 7 fundamental steps:

1) Generate the "clean" or uncontaminated unfolding data under the GGUM and dominance data under the GPCM.
2) Generate aberrant data for simulated random responders, longstrings, ERS, MRS, and mixed condition.
3) Use GPCM and GGUM to estimate item and person parameters for all conditions.
4) Assess the quality of the estimated parameters.
5) Evaluate how each model (GPCM and GGUM) fits each dataset per condition.
6) Compute the $l_{z(p)}$ and $l^*_{z(p)}$ parametric person-fit statistics for each simulee (and selected nonparametric person-fit statistics identified in study 2) and flag simulees who meet the decided criteria for having aberrant responses.
7) Calculate the type I error rates (falsely identifying a simulee as aberrant when their responses were "clean") and detection rates (correctly identifying a simulee as aberrant) for each condition.

Methods for the generation of the data along with estimating the parameters and assessing their quality (steps 1 – 4), and model-data fit (step 5) are described in Study 1. For this study, the $l_{z(p)}$ and the $l^*_{z(p)}$ statistics will use both the GPCM and the GGUM (depending on the condition) to estimate relevant parameters in its computation, thus model fit and parameter quality will include conditions where GPCM fits GPCM data, GGUM fits GGUM data, as well as the cross-fitting of the GPCM to GGUM data and the GGUM to GPCM data.

As Tendeiro (2017) notes, it is important to assess the quality of the parameters estimated and ensure that the aberrant data did not affect these estimates to a large degree because it would then be difficult to determine if the performance of the person-fit statistic was primarily due to the context of the data or the quality of the estimates. This is primarily a concern for the parametric person-fit statistics in the study ($l_{z(p)}$ and the $l^*_{z(p)}$) which use item and person parameter estimates in their calculations. However, in the cross-fitting conditions (e.g., where the statistic uses GPCM to fit GGUM), parameter quality is not expected to be relatively high, even for clean data. In these cases, it is expected for the person-fit statistics to not perform as well. The purpose for this type of condition is to evaluate the flexibility of the models and investigate the rate at which the person-fit statistic performance declines. Because the nonparametric person-fit statistic(s) selected from the results of Study 2 should not be affected by model misspecification since nonparametric person-fit statistics do not rely on parameter estimates, selected statistics from Study 2 were chosen for inclusion in Study 3 in order to compare the performance of the parametric and nonparametric person-fit statistics under misspecified model conditions. The appropriate model was also fit to the corresponding data (e.g., the GGUM was fit to the GGUM data), and parameter quality as well as model-data fit was assessed to ensure data was simulated properly for the purpose of the study. These insights are all taken from Study 1 to gain a better understanding of the context for which the person fit statistics are used in the current Study 3.

Computing $l_{z(p)}$ and the $l^*_{z(p)}$ Person-Fit Statistics (Step 6)

Code was written in R to compute two versions of the $l_{z(p)}$ and $l^*_{z(p)}$ statistics: one specifying the GPCM, and one specifying the GGUM as the model used to estimate the item parameters. When fitting the GPCM to each type of data, the $l_{z(p)}$ was computed using the

lzpoly function in the PerFit package in R, with the IRT model argument set to 'GPCM', to validate the coding was done correctly. To the author's knowledge, no packages in R exist for the $l^*_{z(p)}$ statistic and none for the $l_{z(p)}$ and $l^*_{z(p)}$ statistics with an option to specify the GGUM model. Thus, for these conditions, the results reflect the newly written code for this dissertation. After values were obtained for each simulee, the distribution was examined. To identify simulees as aberrant or non-aberrant, a flagging criterion had to be set. This cutoff was determined by finding the value associated with the critical value for a 5% probability of a type I error using the data with no aberrant responding. For example, for the $l_{z(p)}$ statistic, lower (more negative) values of the statistic indicate greater misfit. Thus, in the equation $P(l_{z(p)}) \leq \text{value}_{\text{critical}} = .05$, the person-fit statistic critical value ($\text{value}_{\text{critical}}$) was used as the cutoff criteria. This methodology for setting flagging criteria has been used in previous person-fit research (e.g., Magis et al., 2012; Emons, 2008; Tendeiro, 2017).

Evaluating Performance of $l_{z(p)}$ and $l^*_{z(p)}$ (Step 7)

To assess the performance of the $l_{z(p)}$ and and $l^*_{z(p)}$ person-fit statistics, both type I error and detection rates were computed using GPCM and GGUM data for each condition. Type I error was calculated using the subset of each sample that was not designated to be aberrant and calculating the proportion of these simulees that were incorrectly flagged as aberrant. Detection rates were relevant in each condition where datasets included aberrant response vectors and were calculated as the proportion of simulees that were generated to have aberrant response vectors that were correctly flagged as aberrant by the person-fit statistic.

Summary

In attempt to address the issue of data quality, researchers are using person-fit analyses in various fields (Rupp, 2013). Additionally, the use of unfolding models for non-cognitive

measures is growing (e.g., Harris-Watson et al., 2020; S. Kartal & Dirlik, 2021; S. K. Kartal & Kutlu, 2020; Liu & Zhang, 2020). Nonetheless, a paucity remains in the literature regarding person-fit under unfolding model frameworks. In order to effectively investigate these joint concepts, Study 1 investigates the differential impacts of aberrant responding on model-fit for unfolding models (and dominance models for comparison), Study 2 examines the performance of nonparametric person-fit statistics under unfolding model contexts, and Study 3 extends the research on parametric person fit statistic performance under an unfolding model by looking at different types of aberrant responding and the impacts of model misspecification.

The methods for all three studies are closely related. In Study 1, the methods outline how the data is generated and models are fit to the data. In Study 2, the methodology is extended to apply nonparametric person fit statistics to detect the generated aberrant respondents. The methods in Study 3 describe how the parametric person fit statistics are tested under the various conditions including model misspecification. All three studies aim to add to the literature regarding the relationships between model fit, presence of aberrant responding, and detection of aberrant responding specifically under an unfolding model framework.

CHAPTER 4

STUDY 1

An Investigation of the Effects of Aberrant Responding on Model-Fit Assuming Different Underlying Response Processes

Abstract

Aberrant responding on tests and surveys has been shown to affect the psychometric properties of scales and the statistical analyses from the use of those scales in cumulative model contexts. This study extends prior research by comparing the effects of four types of aberrant responding on model fit in both cumulative and ideal point model contexts using graded partial credit (GPCM) and generalized graded unfolding (GGUM) models. When fitting models to data, model misfit can be both a function of misspecification and aberrant responding. Results demonstrate how varying levels of aberrant data can severely impact model fit for both cumulative and ideal-point data. Specifically, longstring responses have a stronger impact on dimensionality for both ideal point and cumulative data, while random responding tends to have the most negative impact on data model fit according to information criteria (AIC, BIC). The results also indicate that ideal point data models such as GGUM may be able to fit cumulative data as well as the cumulative model itself (GPCM), whereas cumulative data models may not provide sufficient model fit for data simulated using an ideal point model.

Introduction

Ideally, all respondents in a sample respond to items using a cognitive process that accurately reflects their attitude or the trait level being measured. Tourangeau and Rasinski (1988) theorized that when people answer attitude questions, the response process begins with interpreting the question and retrieving information from the brain, followed by forming a judgement based on the retrieved information, and finally mapping the participant's judgement on to one of the available answer choices. In reality, however, a subset of participants in a sample may lack the motivation to expend the cognitive effort required for such a process (i.e. low cognitive effort, LCE; Murphy, 1996; Turner, 2018). Meijer (1996) outlines several types of item score patterns that are deviant (aberrant) from what may be expected when participants are employing sufficient cognitive effort. Two commonly listed behaviors include providing random responses or the same response (longstrings) to a block of items by carelessly or purposely disregarding item content. These response strings can impair the accuracy of a validation assessment and provide estimates of construct outcomes that are inaccurate for participants. Additionally, some participants may possess the cognitive effort, but engage in response styles that can also impact instrument validation and trait estimates. Two more commonly studied response styles include the tendency to use extreme response options, that is, extreme response style (ERS; Bachman & O'Malley, 1984; Baumgartner & Steenkamp, 2001; Chen et al., 1995; Greenleaf, 1992; Hui & Triandis, 1985; Marin et al., 1992; Weijters et al., 2010) or the tendency to use midpoint response options, that is, midpoint response style (MRS; Baumgartner & Steenkamp, 2001; Chen et al., 1995; Stening & Everett, 1984). These characteristics can result in response data that is aberrant or deviant from what would be expected given the participant's trait level and the characteristics of the item. The literature has revealed many potential adverse

effects of aberrant data and thus its detection has gained substantial attention in several fields (Rupp, 2013). Once aberrant responses are detected, the next steps for the researcher include trying to understand how that data may impact their study. This includes understanding how robust their model of choice is in the presence of low quality data.

Item response theory (IRT) is a latent class model framework that has been recognized for its many advantages in item design and analysis in various applied settings (Bortolotti et al., 2013; Drasgow & Hulin, 1990; Hambleton & Swaminathan, 1985; Reise et al., 2005).Using IRT, researchers can evaluate a latent construct for a set of individuals based on information gathered from item responses. Some of the most widely used unidimensional IRT models (e.g., 1-, 2-, and 3-parameter logistic models for dichotomous data) are based on cumulative item response functions (IRFs) that are monotonically increasing and imply an underlying *dominance* response process. This approach is derived from the work of Likert (1932), and suggests that when an individual has higher agreement to a positively worded scale item, it indicates that individual has a higher level of the measured trait (Dalal et al., 2014). However, methods derived from Thurstone (1927, 1928, 1929) have been proposed to reflect an ideal-point response process where the probability that an individual will endorse an item increases as the disparity between the location of the item and individual on the underlying latent continuum decreases. This resulting IRT model is said to be "unfolding" and allows for peaked, non-monotonic IRFs.

Many studies that have investigated the effects of aberrant data have done so using dominance IRT models. It is unknown whether impacts of aberrant data when using dominance IRT models apply similarly when scales are developed using an unfolding IRT model. The current study focuses on the differential impacts of four types of aberrant responding on model-data fit and parameter recovery using both dominance and unfolding IRT models.

Dominance and Unfolding Models

A primary difference between dominance and unfolding IRT models is the assumed underlying response process for each. In the dominance IRT response process, regardless of the item's location on the continuum, the probability of endorsing an item increases as the test-taker's ability or trait level increases. This dominance characteristic is illustrated with monotonically increasing IRFs [see Figure 5(A)]. For example, when measuring the self-reported social skills of respondents with an item like "My social skills are at least as good as those of an average person," it may be reasonable to assume that the more extraverted the respondent is, the more likely he or she will endorse this item (Chernyshenko et al., 2007). On the other hand, the ideal-point response process assumes that a person will endorse an item to the degree that the item reflects the person's own standing on the construct being measured. In other words, the likelihood of endorsing an item increases as the difference between the test taker's and item's location on the continuum representing the construct of interest decreases. For example, the item, "My social skills are about average," would most likely elicit an ideal point response process where the probability for endorsement is highest for those who feel they have average social skills (Chernyshenko et al., 2007). Respondents may disagree with this item for two reasons: either they feel their social skills are above average or because they feel their social skills are below average. The IRF for this respondent-item relationship can be seen in Figure 5B. Due to this "unfolding effect," items that elicit an ideal point response process are often measured using an unfolding model. Some common unfolding models that are available for both dichotomous and polytomous data include the Squared Logistic Model (SLM; Andrich, 1988), PARELLA model (Andrich, 1988), Hyperbolic Cosine Model for dichotomous data (HCM; Andrich & Guanzhong Luo, 1993), General Hyperbolic Cosine Model for polytomous data (GHCM;

Andrich, 1996), Graded Unfolding Model (GUM; Roberts & Laughlin, 1996), and the generalized graded unfolding model (Roberts et al., 2000). Several aspects of the GGUM make it a popular choice and rationalize its selection for this investigation. First, it can be used with both polytomous and dichotomous data, and the discrimination parameters are allowed to vary across items. The thresholds for each response option are also allowed to vary across items. The freedom for variability among the discrimination and threshold parameters create a flexibility for the IRFs under this model to take on a wide range of shapes (Stark et al., 2006). Further, the GGUM package in R (Tendeiro & Castro-Alvarez, 2019) makes it an accessible option for researchers.

The GGUM proposed by Roberts et al. (2000), defines each subjective response to follow the generalized partial credit model (GPCM; Muraki, 1992). To maximize comparability of results in this study, the GPCM was chosen to model the data reflecting a dominance response process. The formulas for both the GPCM and GGUM are given below.

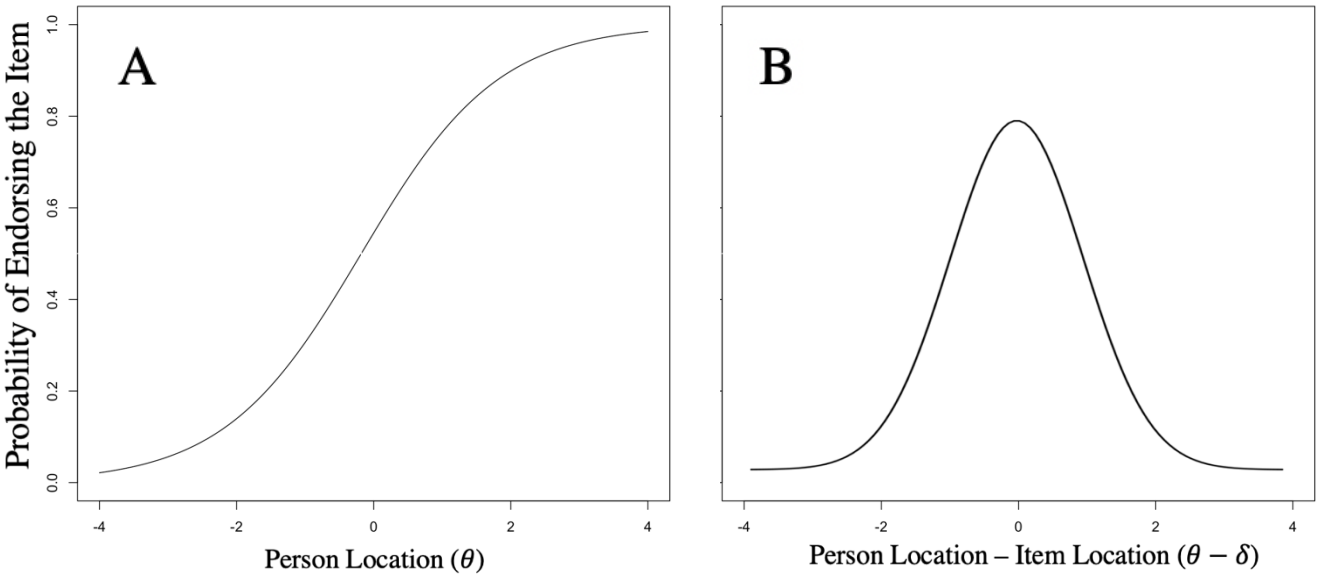
$$GPCM:\ P\left(Z_j = z|\theta\right) = \frac{\exp\{\sum_{k=0}^{z}[a_j(\theta - b_{jk})]\}}{\sum_{r=0}^{C}\{\exp\sum_{k=0}^{r}[a_j(\theta - b_{jk})]\}} \tag{30}$$

where Z_j represents the observed response (with ability or trait level, θ) to item j, and $z = 0, 1, 2, \ldots, C$ with $z = 0$ corresponding to the strongest level of disagreement and $z = C$ corresponding with the step that reflects the strongest level of agreement. Thus, C is the number of observed response categories minus 1. The discrimination parameter for item j is represented by a_j, and b_{jk} is the difficulty parameter or location parameter of the k^{th} step. In the denominator, $r = 1, 2, \ldots, C$ represents the total C exponent terms.

$$GGUM:\ P\left(Z_j = z\middle|\theta\right) = \frac{\exp\{\alpha_j[z(\theta - \delta_j) - \sum_{k=0}^{z}\tau_{jk}]\} + \exp\{\alpha_j[(M-z)(\theta - \delta_j) - \sum_{k=0}^{z}\tau_{jk}]\}}{\sum_{w=0}^{C}\{\exp\{\alpha_j[w(\theta - \delta_j) - \sum_{k=0}^{w}\tau_{jk}]\} + \exp\{\alpha_j[(M-w)(\theta - \delta_j) - \sum_{k=0}^{w}\tau_{jk}]\}\}}, \tag{31}$$

where *M* is the total number of subjective response categories minus 1 (*M = 2C +1*), α_j is the discrimination parameter for item *j*, δ_j is the item location parameter, and τ_{jk} is the location of the k^{th} threshold on latent continuum relative to the location of the j^{th} item.

Figure 5. IRFs based on Dominance (A) and Ideal Point (B) Models



Note. A illustrates a monotonically increasing item response function (IRF). B illustrates a non-monotonic IRF, violating monotonicity assumption for dominance item response theory models

Although dominance IRT models like the GPCM are more widely used in the literature (Harris-Watson et al., 2020; Tay & Ng, 2018), evidence from several self-report inventories has indicated that an unfolding approach (e.g., GGUM) may be more suitable for several types of assessments intended to measure non-cognitive constructs. These include the assessment for creativity using the Gough's Creative Personality Scale (Zampetakis, 2010), measures of conscientiousness (Carter et al., 2014), personality inventory of self-judgement on the order-facet (a feature of conscientiousness; Chernyshenko et al., 2007; Weekers & Meijer, 2008), the fifth edition of the Sixteen Personality Factor Questionnaire (Stark et al., 2006), and the Job Descriptive Index measuring job satisfaction (Carter & Dalal, 2010).

Impacts of Aberrant Responding

DeSimone et al. (2018) suggested that even 10 to 15 percent contamination of LCE response vectors in a dataset should be a cause for concern for researchers. This is consistent with Meijer's study in 1997 where 15% or higher contamination led to substantial decreases in criterion-related validity under the context of a three parameter logistic (3PL) IRT model. In the DeSimone et al. (2018) study, random responding led to lower interitem correlations, lower reliability estimates, and masked the true factor structure. Thus, when random responding is present, the researcher may be at higher risk for making Type II errors and failing to reveal relationships between variables that may actually exist (McGrath et al., 2010). Longstring responses on the other hand may artificially increase inter-item correlations and inflate reliability if the items are worded in only one direction (DeSimone et al., 2018).

Avşar (2021) investigated the impact of aberrant data on confirmatory factor analysis results and item parameter estimates under a graded response model and a Mokken Homogeneity Model (both dominance IRT models). The study used person fit statistics to trim the data of aberrant responses and compared CFA and statistical results with the untrimmed data. Overall, results were mixed depending on the person fit statistic used, but generally suggested that including the aberrant data resulted in worse goodness-of-fit results. When aberrant data were identified and set aside using the l_z^p person-fit statistic, item discrimination parameters increased and goodness-of-fit results were improved.

Very few studies have examined the issue of aberrant data under an unfolding model context. Liu and Zhang (2020) investigated the appropriateness of ideal point models for detecting faking on personality measures (as opposed to honest responding). Results indicated that fitting the GGUM to the data under conditions where faking was present resulted in shifts in

item location parameter estimates, demonstrating that faking could increase or decrease personality factor scores reflecting conscientiousness and neuroticism. Liu and Wang (2019) showed that parameters estimated by the General Unfolding Model (GUM) may be biased when response styles are ignored. The study also reported that test-retest reliability decreased with the presence of aberrant responses. However, it is unknown how other types of aberrant responding (e.g., random responding, longstrings) may affect model fit for unfolding models compared to dominance models. Tendeiro (2017) investigated the detection of MRS and ERS using a parametric person fit statistic ($l^*_{z(p)}$) in a GGUM simulation study. Model fit was investigated before and after adding aberrant data, with little impact under the conditions used in the study (most extreme aberrant condition was 25% of response strings aberrant for 20% of the sample). However, the detection rates for ERS were found to be more strongly affected by parameter bias than detection rates for MRS. If different types of aberrant data affect model fit and parameter bias differently for dominance and unfolding models, this could lead to important implications for studies using person fit statistics.

Purpose

In reality researchers will not know if model-data misfit is due to the presence of aberrant responding, or if it is due to true model misspecification. One of the advantages of using a simulation design for the current study is the full control over these factors. Aberrant responding and model misspecification (based on two underlying response processes) will be manipulated to provide insight on the confounding effects that may arise with real data. For example, it could be that aberrant responding affects model fit more severely for certain data types and conditions, which would then point the researcher to investigate data quality before making conclusions about model misfit.

If model-data misfit is shown to increase the estimated parameter bias, then the detection of the aberrant responses may in turn be affected. Tendeiro (2017) found that bias of estimated parameters affected the detection rates of extreme responding more so than for midpoint responding. Thus, if aberrant responding affects the model-data fit, and the model-data fit affects the detection of aberrant responding, the researcher faces a difficult decision while attempting to maintain the integrity of the data quality. More research in this area is warranted to provide insight for researchers in detecting aberrant responses. The purpose of this study is to contribute to growing literature on unfolding versus dominance models as well as the impact of aberrant data by providing insight on what conditions and types of data may be affected differently by aberrant data (Freund & Lohbeck, 2021; LaPalme et al., 2018; Nye et al., 2020; Wilgus & Travis, 2019). The research questions driving the study include:

1) How does model fit compare for dominance and unfolding IRT models (GPCM and GGUM) applied to both dominance and ideal point response data simulated with no aberrant responses?
2) How is model fit and parameter recovery impacted for both dominance and unfolding models applied to both dominance and ideal point response datasets (GGUM fit to GGUM data, GGUM fit to GPCM data, GPCM fit to GGUM data, GPCM fit to GPCM data) when different types and proportions of aberrant response strings are included?
    a. Do certain conditions (test length, type of aberrant response, proportion of aberrant responders and proportion of aberrant responses within an aberrant response vector) have different results for model-data fit and the quality of parameter recovery?

Methods

A series of simulations was carried out for 6-point items and a fixed sample size of 1,000. Four types of aberrant responding were included: two types of response styles (extreme response style and midpoint response style), and two forms of low cognitive effort responding (random responding and longstrings). Additionally, a 'mixed aberrant response' condition combined all four aberrant response types to simulate realistic situations where a sample may be composed of several types of aberrant responders within a dataset. Extending two published studies on aberrant data within an unfolding IRT model (Liu & Wang, 2019; Tendeiro, 2017), three proportions of aberrant responders (*AbN* = .04, .10, .20) and three proportions of aberrant responses to the items within each aberrant response string (*AbI* = .20, .40, .60) were used to simulate aberrant response data. Two test lengths of 20 and 40 items were included. This resulted in a total of 2 (data generating mechanisms: GGUM and GPCM) × 2 (applied model fit: GGUM and GPCM) × 3 (proportion of aberrant responders in the sample, *AbN*) × 3 (proportion of aberrant responses in response vectors, *AbI*) × 2 (test lengths) × 5 (types of aberrant responding and response styles) = 360 fully crossed conditions. All conditions were replicated 100 times. Under certain conditions (especially when cross-fitting an inappropriate model with the data), results did not converge. In these cases, more than 100 replications were necessary to generate new clean datasets until 100 iterations successfully completed. Each condition began with generating a new "clean" dataset and then replacing response strings with aberrant data based on the *AbI* and *AbN* condition. The clean datasets were used to obtain a baseline for non-aberrant responses, and to compare against the aberrant datasets in each condition. Seed values were used to allow for replicability. Code for the data generation, model fit, and all analyses were written in R (R Core Team, 2016). Code is attached in Appendices A through F.

Data Generation

For both the GGUM and GPCM datasets, person parameters were randomly drawn from the standard normal distribution. The GenData.GGUM function in R (Tendeiro & Castro-Alvarez, 2019) was used to generate all item and person parameters as well as the item scores for the GGUM datasets. The item discrimination parameters (α_j) were randomly sampled from a uniform distribution [0.5, 2.0]. The item location parameters (δ_j) were randomly sampled from the standard normal distribution truncated between -2.0 and 2.0 due to reports of extreme values of δ_i sometimes leading to issues of low accuracy and variability of MML estimates under the GGUM (Roberts & Thompson, 2011 as cited in Tendeiro, 2017). The locations of the threshold parameters (τ_{jk}) (relative to the location of the *j*th item) were recursively generated using procedures described in Roberts et al. (2002) using the GenData.GGUM function in R.

To generate data under the GPCM, the item discrimination parameters were also sampled from a uniform distribution [0.5, 2.0], and item difficulty parameters for each item were sampled from the standard normal distribution $N(0,1)$. Item category thresholds, d_{jk}, for step k of item j were simulated by taking the sequential cumulative sum of five numbers drawn from a random uniform distribution between .3 and 1, as described in Chalmers (2012). This interval was chosen because it ensures that the distance between categories will not be less than .30. If the categories are too close, some may not be chosen as often. Next, the mean for the set of sequential cumulative sums for each item was subtracted from each number in the set. The initial item category threshold, d_{j0}, was set to 0 in order for the model to be identified (Muraki, 1992). Once the parameters were generated, the sim_gpcm function in R (PP package; Steinfield & Reif, 2021) was used to simulate the response data.

Generation of Aberrant Responses

Aberrant datasets were created by replacing a proportion (depending on the *AbN* condition) of response vectors in the clean datasets with aberrant response vectors. To simulate aberrant responses due to random responding, first a random sample of integers were drawn from a uniform distribution [0, 5], and then the values of the randomly sampled numbers replaced either 20%, 40%, or 60% of the responses (*AbI*) randomly throughout an aberrant response vector. To create aberrant responses due to longstrings, a single random number was drawn from a uniform distribution [0, 5]. Next, a valid random starting position in the vector was generated and the specified proportion of consecutive items ($AbI \times number\ of\ items$) was set to equal the randomly generated number (DeSimone et al., 2018).

To create response vectors that demonstrate MRS for 6-point items, endpoint item scores and item scores adjacent to the endpoints were replaced with the closest midpoint response (i.e., on the 6-point scale ranging from 0 to 5, items scores of 0 and 1 were replaced with a 2, and scores of 4 and 5 were replaced with a 3). To mimic responses that reflect ERS, item scores in the middle range were changed to the corresponding endpoint responses (i.e., 1s and 2s were changed to 0; 3s and 4s were changed to 5). This means that the item score for someone who would be expected to respond with a 1 on an item (based on their ability and the item parameters) could either be changed to a 2 if they were designated to demonstrate MRS or a 0 for ERS. A similar situation could happen for responses of 4. The reason data were simulated in this way is due to the potential for this scenario to happen in real life. It is possible that respondents with MRS and ERS may move responses more towards the middle or extreme regardless of the type of item. That is, we did not want to only move end-point response to the middle for MRS

and the most middle responses to the endpoint for ERS, but also include items/person combinations where responses would be expected to land anywhere on the scale.

Fitting the IRT Models

Once response strings for all simulees were generated based on the GPCM and GGUM, an IRT model (depending on the condition) was fit to each set of response data. This resulted in each dataset and its corresponding aberrant counterpart were fit by the model in their condition. The GPCM was fit to the data using the MIRT function (itemtype = gpcm) in R (Chalmers, 2012), whereas the GGUM was fit to the data using the GGUM function in the GGUM package in R (Tendeiro & Castro-Alvarez, 2020). Any time GPCM was fit to GGUM responses, "Likert-scaling techniques" described in Tay et al. (2011) and Tay and Drasgow (2012) were applied. First, 30% of items on the negative end of the continuum were reverse scored. Next, further reverse coding was completed for any items remaining with negative item-total biserial correlations. This procedure was employed to reduce the identification of an additional factor in ideal-point data which has been pointed out as a potential problem in past research (Tay & Drasgow, 2012; van Schuur & Kiers, 1994; Williams, 2015).

Evaluating Model Fit

To evaluate model fit, multiple aspects were examined including dimensionality, goodness-of-fit, information criteria, and quality of parameter estimates. Dimensionality was assessed using parallel analysis in R ('paran' package; Dinno, 2018). The method utilized implemented Horn's technique for quantitatively and graphically determining the number of factors retained in a Principal Components Analysis (PCA) while adjusting for the sample error-induced inflation. This method was chosen because other methods such as scree plots and the Kaiser rule, have been shown to overestimate dimensionality in data (Zwick & Velicer, 1984).

Information criterion based statistics, Akaike Information Criterion (AIC; Akaike, 1974) and Bayes Information Criterion (BIC; Schwarz, 1978) were used to compare relative model fit. These indices take into account the number of parameters being estimated and were computed as follows:

$$AIC = 2k - 2\ln(\hat{L}) \tag{32}$$

$$BIC = k\ln(n) - 2\ln(\hat{L}) \tag{33}$$

where k is the number of parameters estimated by the model, $\hat{L}$ is the maximized value of the likelihood function for the model, and n is the sample size. Lower AIC and BIC values indicate better fit. Nye et al. (2020) found that AIC and BIC criteria were able to distinguish the correct model fit for GGUM versus the graded response model (GRM) in a simulation study. However, it is emphasized that these are relative model fit indices used to compare the relative fit and do not illustrate absolute model fit.

When determining which model (unfolding or dominance) fit the data better, results from the Tay et al. (2011) study suggest researchers may use the doubles and triples adjusted χ^2 ratios. Nye et al. (2020) were also able to use this method to uncover the appropriate model fit to GRM versus GGUM data. Adjusted χ^2/df ratios for pairs and triplets of items in this study were reported for model comparisons across conditions (Drasgow et al., 1995). Chi-square statistics were computed using the observed frequencies of item *j* response option *z* (O_{jz}) and the expected frequencies (E_{jz}) based on the estimated item parameters and distribution of abilities.

$$\chi_j^2 = \sum_{z=0}^{C} \frac{(O_{jz} - E_{jz})^2}{E_{jz}}, \tag{34}$$

with $E_{jz} = N \int P_{jz}(\theta)\phi(\theta)d\theta$, where $\phi(\theta)$ is the standard normal density function. This equation is written for single items but is often generalized to apply to pairs of items (doublets)

and triples of items (triplets) as these have been shown to be more reliable estimates of model (mis)fit than single items (Drasgow et al., 1995). Because χ^2 statistics are heavily dependent on sample size, they suggest adjusting the χ^2/df statistic to a sample size of 3,000 to make it more generalizable across samples of different sizes (also see Lahuis & Clark, 2009). The resulting formula is:

$$\chi^2/df \; = \; \frac{3{,}000(\chi_j^2 - df)}{N} + df \tag{35}$$

The degrees of freedom in Equation 4 depend on the number of singlets, doublets or triplets used. It has been suggested that values of χ^2/df larger than 3 be considered heuristically indicative of model misfit (Tendeiro, 2017).

The quality of parameter estimates was also compared to reveal any differences in parameter recovery. Procedures in Tendeiro (2017) were followed by comparing bias, mean absolute deviation (MAD), and the correlation (COR) between true and estimated parameters. The following equations were used:

$$MAD = \sum_{t=1}^{T} |\hat{\gamma}_t - \gamma_t^{TRUE}|/T \tag{36}$$

$$BIAS = \sum_{t=1}^{T} (\hat{\gamma}_t - \gamma_t^{TRUE})/T \tag{37}$$

$$COR = cor(\hat{\gamma}_t, \gamma_t^{TRUE}), t = 1, \dots, T, \tag{38}$$

where $\widehat{\gamma_t}$ is the estimated parameter and γ_t^{TRUE} is the true simulated value for the parameter representing either $\alpha_j, \delta_j, \tau_{jk}, or\ \theta_n$ for the GGUM parameters or $\alpha_j,\ b_{jk}, or\ \theta_n$ for the GPCM, and *T* is the corresponding total number of that parameter. For example, *T* is equal to the number of items for α_j and δ_j. *T* equals the number of items times *C* (the number of observed response categories minus 1) for τ_{jk}, and *T* is equal to the sample size for θ_n. The MAD, bias, and correlations were computed and averaged over all replications for each condition. Standard deviations for these averages were also reported to examine the variability in parameter bias.

When fitting GPCM to GGUM data or the GGUM to the GPCM data, the comparison of some estimated parameters is difficult to interpret. For example, the location parameter in dominance IRT models represents the theta associated with a .50 probability of choosing a response. However, in unfolding models, the location parameter represents the theta associated with the highest probability of endorsement. Furthermore, discrimination parameters for non-monotonically increasing items may result in negative values. Thus, for model cross-fit conditions, only the parameter quality for person scoring is presented.

Results

Dimensionality

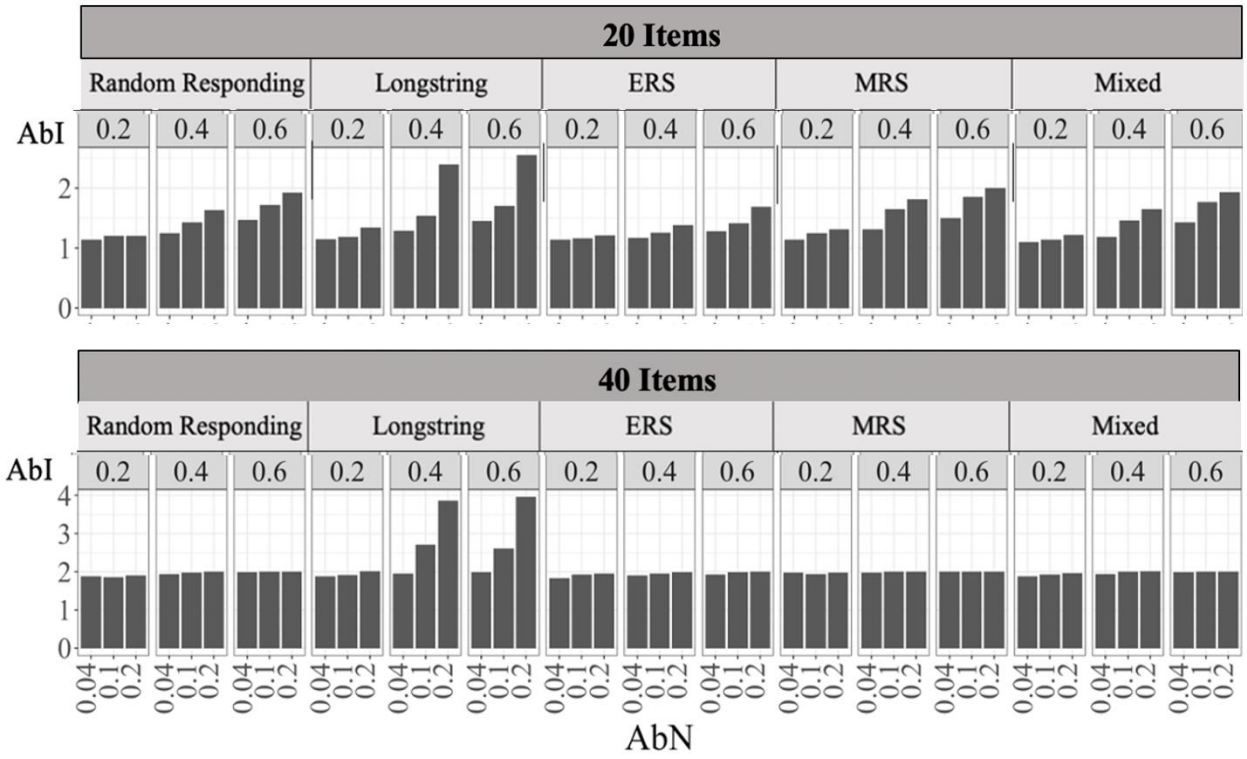
A parallel analysis using Horn's technique was performed when investigating dimensionality for all datasets under each condition. The number of factors retained for each analysis was averaged across replications and trends were examined across conditions for *AbI*, *AbN*, the number of items, the types of aberrant responding, and the two types of data (dominance and ideal point). Using parallel analysis, the number of factors retained are expected to be one for the dominance response data, whereas two factors are expected with ideal-point response data (Tay et al., 2011; Tay & Drasgow, 2012; Williams, 2015). The mean number of factors retained for clean datasets generated using a dominance response process (GPCM) was 1.11 for datasets with 20 items and 1.84 for datasets with 40 items. Across all GPCM datasets with aberrant data, the mean number of factors retained was 1.44 for datasets with 20 items and 2.06 for datasets with 40 items. In contrast to the dominance GPCM datasets, adding items and aberrant responses to the unfolding GGUM data had a very small impact on the number of factors retained. For clean datasets generated according to GGUM, the average number of factors retained for both 20-item and 40-item datasets was 2.00. Adding aberrant data to the GGUM

datasets had a negligible impact on dimensionality assessment, where the average number of factors retained were 2.01 for 20-item aberrant datasets and 2.04 for 40-item aberrant datasets.

Further investigation of the trends in dimensionality assessment across datasets of varying aberrant response conditions are illustrated in Figures 6 (GPCM) and 7 (GGUM). The 20-item graph in Figure 2 shows the upward trends in factor retention for the GPCM datasets with increasing proportions of aberrant responses. This was most evident in the case of increasing longstring responses. When 20% of the sample was simulated to have longstring responses (*AbN* = .20) for at least 40% of the items (*AbI*= .4 or .6), at least 1 additional factor was inappropriately retained 100% of the time. A spurious dimension was almost always found for the 40-item datasets, where proportions of aberrant responding had a smaller effect on factor retention. The exception to this finding was when the *AbType* was specified as longstring responses. In this case, increasing proportions of aberrant responding to *AbN* = .2 and *AbI*=.4 or .6, resulted in an average of nearly 4 factors being retained.

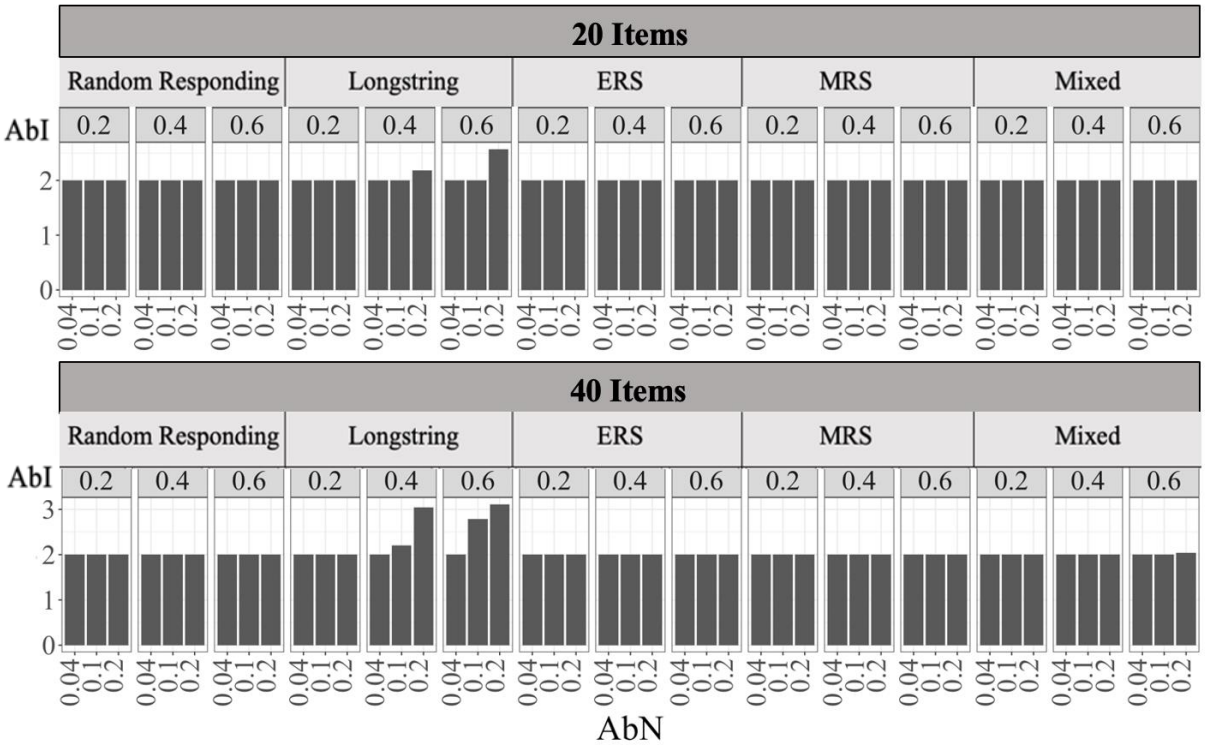
Longstring responses had a similar effect on the GGUM data to a slightly lesser extent (Figure 7). As reflected by the means, regardless of the number of items (20 or 40), 2 factors were nearly always retained. However, when longstring responses were the specified type of aberrant response, and the proportion of aberrant respondents was 20% (*AbN* =.20) with 60% of item responses aberrant within each specified vector (*AbI* = .60), about half of the replications resulted in an additional 3rd factor being retained in the 20-item datasets. The same condition in the 40-item datasets (and even when *AbI* =.4) resulted in an additional 3rd factor being retained 100% of the time.

Figure 6. Trends in the Average Number of Factors Retained by Condition in GPCM Data



Note. ERS = Extreme Response Style. MRS = Midpoint Respoinse Style. Random = Random responders. *AbI* = Proportion of items within response vector designated as aberrant. *AbN* = Proportion of simulees designated to have aberrant response vectors.

Figure 7. Trends in the Average Number of Factors Retained by Condition in GGUM Data



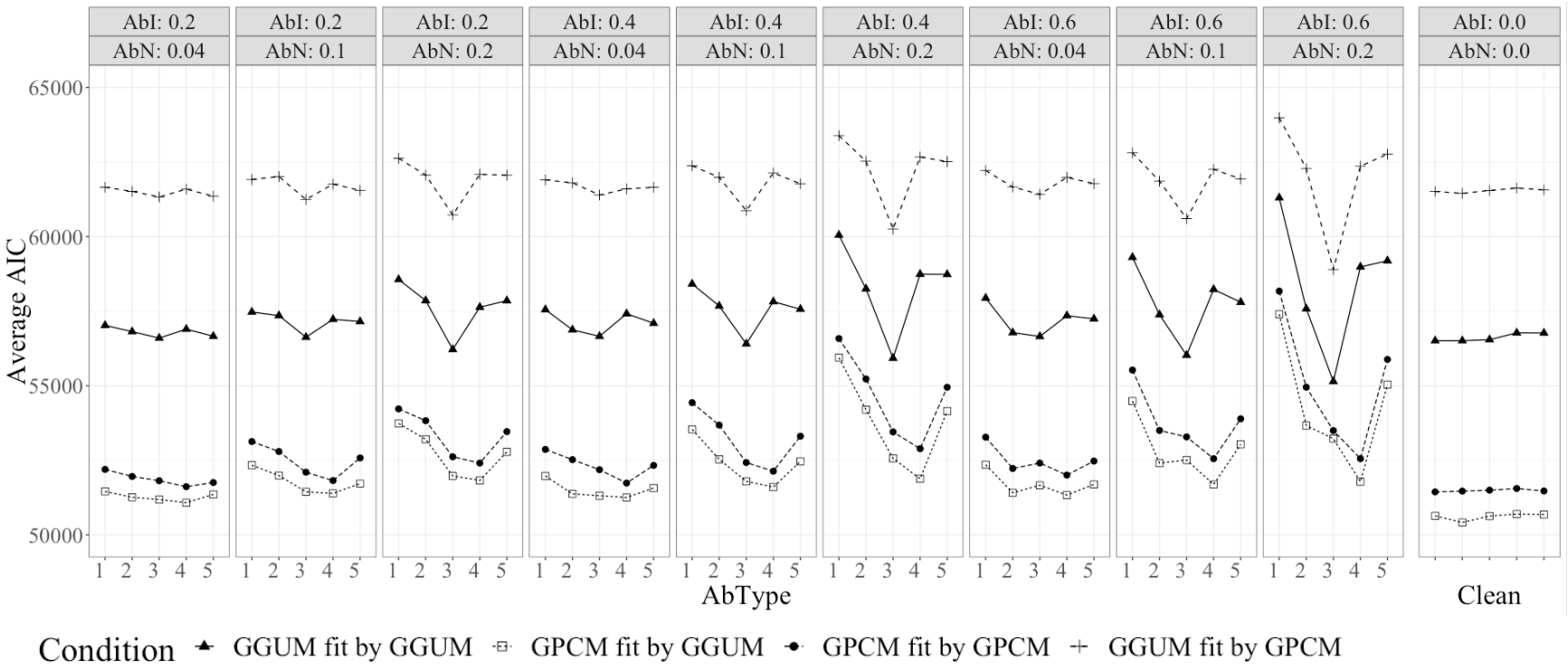
Note. ERS = Extreme Response Style. MRS = Midpoint Respoinse Style. Random = Random responders. *AbI* = Proportion of items within response vector designated as aberrant. *AbN* = Proportion of simulees designated to have aberrant response vectors.

Information Criteria

To compare the relative fit of each model to the data, information criterion fit indices AIC and BIC were utilized because both indices not only consider the non-nested structure of the two models in the study, but they also factor in penalties for additional parameters in more complicated models like GPCM and GGUM. A lower AIC and BIC value indicates better fit. Because these criteria penalize for model complexity, in general the AIC and BIC values for the GPCM fit to data are expected to be lower than the values for the GGUM since GGUM is a more complex model. Although BIC values were slightly higher than the AIC values due to the stricter penalty for model complexity, the results for both AIC and BIC were very similar. Thus, only the AIC results are presented for the 20-item and the 40-item datasets (Figure 8 and 9, respectively).

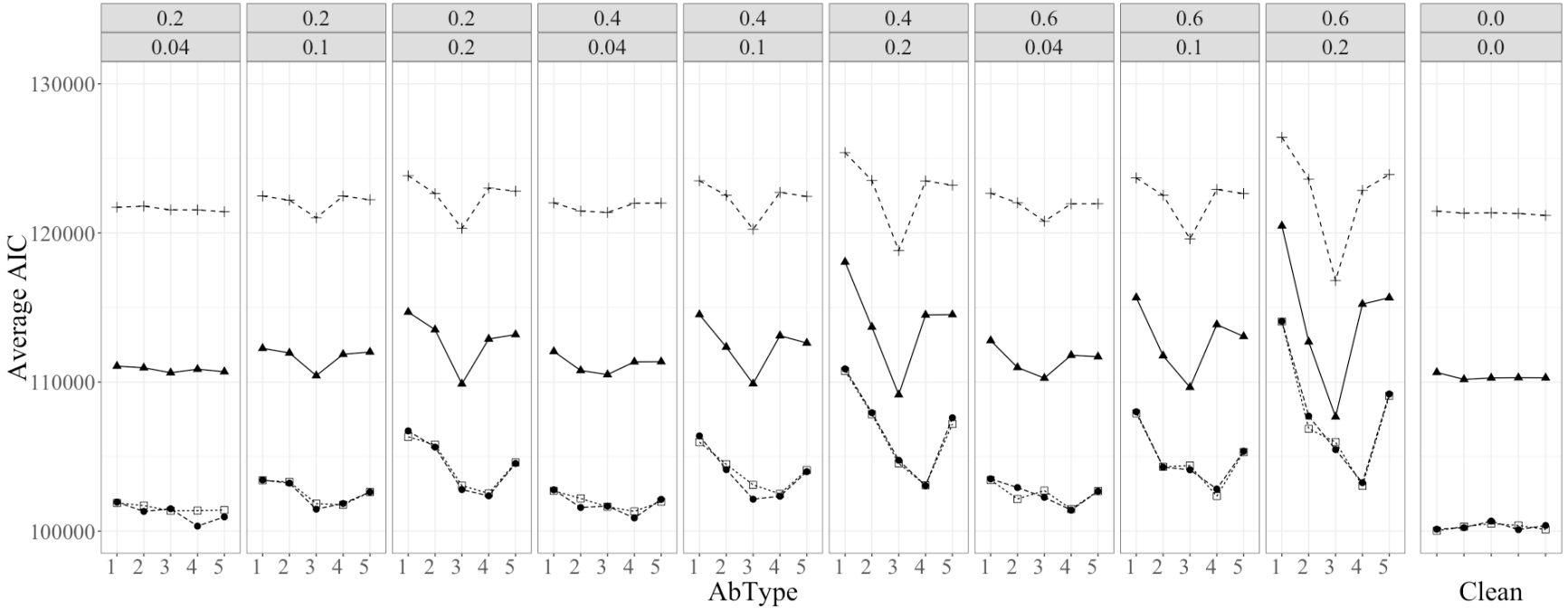
The GPCM fit to the GGUM data had the relatively worse fit as indicated by the AIC and BIC, whereas GGUM fit the GGUM data better with lower AIC values for all conditions. For 20-item datasets, when GGUM was fit to the GPCM data (represented by the hollow squares), slightly lower AIC values resulted compared to when GPCM was fit to the GPCM data (represented by the solid circles). The difference between the fit of the two models (GGUM and GPCM) for the GPCM data disappeared for the 40-item conditions with essentially the same AIC and BIC values for both models. For both types of data, random responding tended to have the most negative impact on data model fit compared to the other types of aberrant responding according to the AIC and BIC. MRS appeared to be more problematic for GGUM data, where ERS was the most problematic for GPCM data. Conversely, increased ERS responding often resulted in improved model fit for the GGUM data compared to other aberrant conditions, and even the clean datasets. When comparing to the clean datasets generated by both GPCM and GGUM, longstring responses had a negative effect on model fit, though this result was much more pronounced for GPCM data than the GGUM data.

Figure 8. Average AIC for Each Model-Data Fit by Condition for 20 Items



Note. AbType 1= Random Responding. AbType 2 = Longstring. AbType 3= Extreme Response Style. AbType 4 = Midpoint Respoinse Style. AbType 5 = Mixed Aberrant Responding. *AbI* = Proportion of items within an aberrant response vector that have aberrant responses. *AbN* = Proportion of simulees that were designated as aberrant responders. Clean datasets were generated before each condition of aberrant responding, but only shown for one arbitrary setting of *AbI*=.4 and *AbN*=.2 in the figure above.

Figure 9. Average AIC for Each Model-Data Fit by Condition for 40 Items



Note. AbType 1= Random Responding. AbType 2 = Longstring. AbType 3= Extreme Response Style. AbType 4 = Midpoint Respoinse Style. AbType 5 = Mixed Aberrant Responding. *AbI* = Proportion of items within an aberrant response vector that have aberrant responses. *AbN* = Proportion of simulees that were designated as aberrant responders. Clean datasets were generated before each condition of aberrant responding, but only shown for one arbitrary setting of *AbI*=.4 and *AbN*=.2 in the figure above.

χ^2/df Ratios

The adjusted χ^2 ratio introduced by Drasgow et al. (1995) was utilized for the study and results for the percent of item triplets flagged for appropriate model fitting conditions and cross-fitting model conditions are presented in Figure 6. Because results were very similar across scale lengths and for item doublets and triplets, only the item triplet results are discussed and presented in Figure 10 for only the 20-item conditions. Results for the clean datasets are first discussed, then the datasets with aberrant responses included.

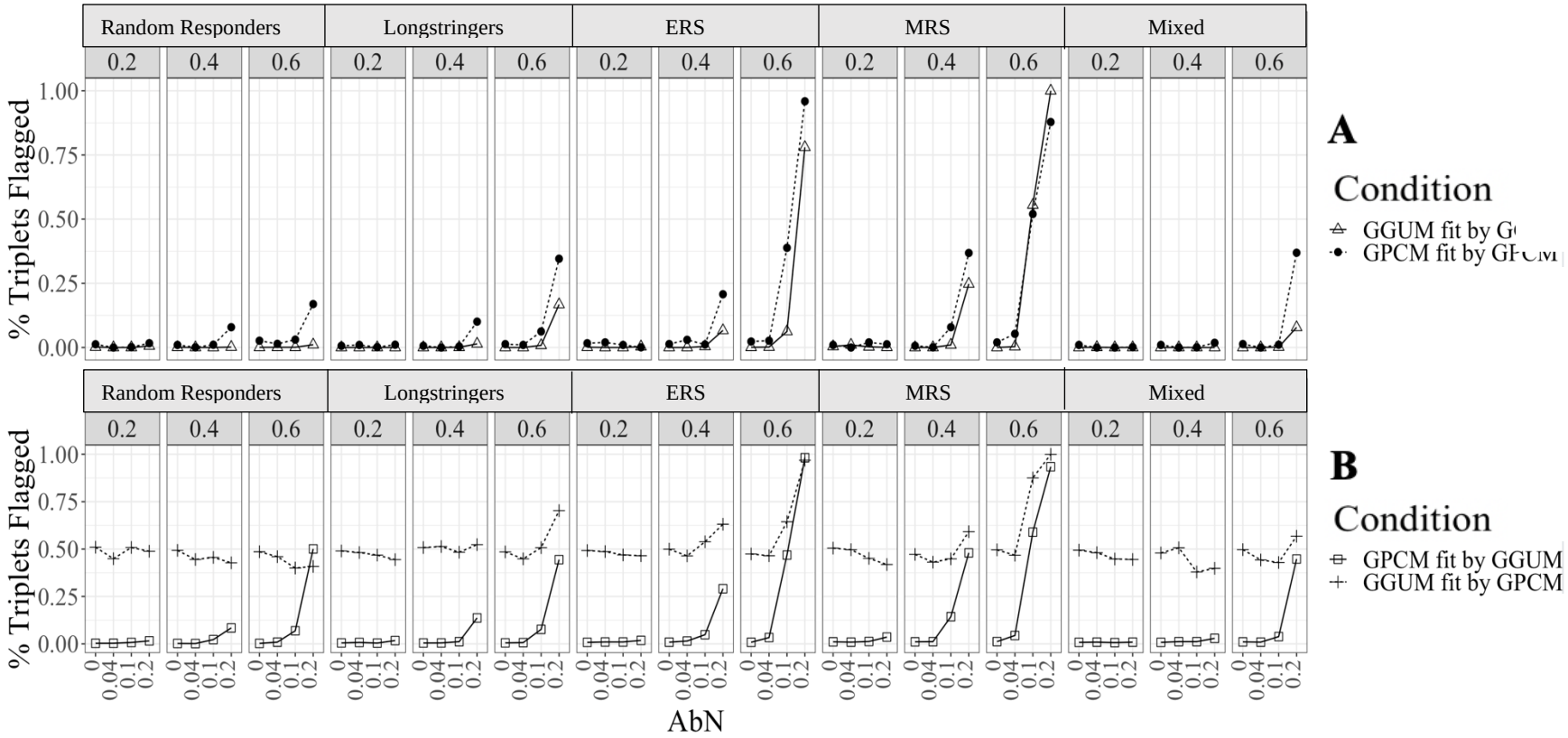
The χ^2/df ratios for item doublets and triplets consistently pointed to the appropriate model for the clean GGUM data (i.e., a larger proportion of item triplets were flagged for cross-fitting model conditions than when the appropriate model was used for the data). When GPCM was cross-fit to GGUM data, even in the clean data, 54.4% of the item doublets were flagged and 49.2% of the item triplets were flagged. In comparison, when the GGUM was appropriately fit to the GGUM datasets with no aberrant responses, less than .08% of the item doublets and .06% of triplets were flagged. In the clean GPCM datasets, the χ^2/df ratios for item doublets and triplets did not as clearly point to the appropriate model. Less than 1.4% of the item doublets and triplets were flagged for clean GPCM datasets fit by the GPCM compared to the 4.2% and 0.7% of the item doublets and triplets being flagged when GGUM was fit to the GPCM data.

Using χ^2/df for doublets and triplets, indications of poor fit were minimal with random responding, longstrings, ERS and MRS when 10% or less of participants exhibit aberrant responding and the proportion of response strings aberrant is 40% or less. It was only when 20% or more participants exhibit aberrant responding (or when 10% exhibit 60% *AbI*) that a significant number of item doublets and triplets were flagged. Of the aberrant response types in the study, MRS and ERS tended to impact the χ^2/df ratios the most in both GPCM and GGUM

datasets, with random responding and longstrings resulting in fewer doublets and triplets being flagged. Even when the appropriate model was used to fit the GGUM data, up to 100% of item doublets and triplets were flagged for the most extreme aberrant conditions of MRS in the study (*AbN* = .2; *AbI* = .6). This suggests the adjusted chi-square doublets and triplets procedure may be effective in detecting the appropriate model for GGUM data except in the case where all doublets and triplets are flagged (*AbN* = .2; *AbI* = .6). Even in the most contaminated GGUM datasets (other than MRS), the χ^2/df ratios for doublets and triplets of items pointed to the correct model fit. In the GPCM datasets, the χ^2/df ratios for item doublets generally pointed to the GPCM for the correct model fit, though this was not as clear as with the GGUM datasets. Specifically, when ERS, MRS, and longstrings were present, and *AbN* = .20 with *AbI* = .60, a high proportion of item doublets and triplets were flagged regardless of the model used to fit the GPCM data.

Overall, when GGUM was applied to the GGUM data, MRS had the largest negative impact on model fit, while ERS tended to have the largest negative effect within the GPCM datasets. When the inappropriate model was used to fit the data, both the *AbI* and *AbN* conditions seemed to impact model fit. Generally, higher *AbI* conditions resulted in worse model fit as indicated by flagged item doublets and triplets. The exception to this trend involved GGUM data being fit by GPCM where no clear trend existed. In this case, the more random responses in a response vector generated by the GGUM, the better GPCM fit the data.

Figure 10. Average Percent of adjusted χ^2Triplets Flagged by Condition



Note. Graph A (Top) reflects appropriate model fit for the datasets. Graph B (Bottom) reflects inappropriate cross-fit of models for the datasets. ERS = Extreme Response Style. MRS = Midpoint Response Style. *AbI* = Proportion of items within an aberrant response vector that have aberrant responses. *AbN* = Proportion of simulees that were designated as aberrant responders. Results presented for the 20-item condition. For each replication of each condition, clean datasets were generated and included in the graphs for comparison purposes (*AbN* = 0). For each model generating mechanism, model fit condition, *AbI*, and *AbType*, there were 3 *AbN* conditions of 100 replications each. Thus, clean datasets for this graph were aggregated and include 300 replications each.

Quality of Parameter Recovery

Using the appropriate model. To assess the quality of parameter recovery across the various conditions, measures of BIAS, MAD, and COR were averaged across replications. For a graphical illustration of the person parameter (theta) recovery across conditions, see Figure 11. Tables 1, 2, and 3 show further details (split by *AbN* condition) on the quality of parameter recovery by *AbI* conditions .2, .4 and .6, respectively. Sections A and B of the tables show results for conditions where the appropriate model was used to fit the data (i.e., Section A = GGUM data fit by GGUM and Section B = GPCM data fit by GPCM). In the lowest aberrant conditions (e.g., *AbN* = .04, *AbI* = .20; Table 7), parameter recovery was good (comparable to the clean data) when the appropriate model is used. The estimated person parameters (θ_n) were strongly related to the true simulated parameters, with correlations being at or above .92, even with datasets that included *AbN*=.20 and *AbI*=.60 (Table 9). In this more extreme aberrant condition, person parameters were recovered best in cases where GGUM data was fit by the GGUM (r = 0.95). Aberrant responses also had very little impact on BIAS and MAD values for person parameter estimates in both GGUM and GPCM datasets when the appropriate model was used to fit the data. For example, when the average person parameter bias for the clean datasets was subtracted from the average person parameter bias for datasets with aberrant responses, the average differences in BIAS were less than .01 across all conditions. The average differences (between clean and aberrant datasets) in person parameter MAD ranged from 0 to .11 depending on *AbI* and *AbN* and were less than .04 across all conditions.

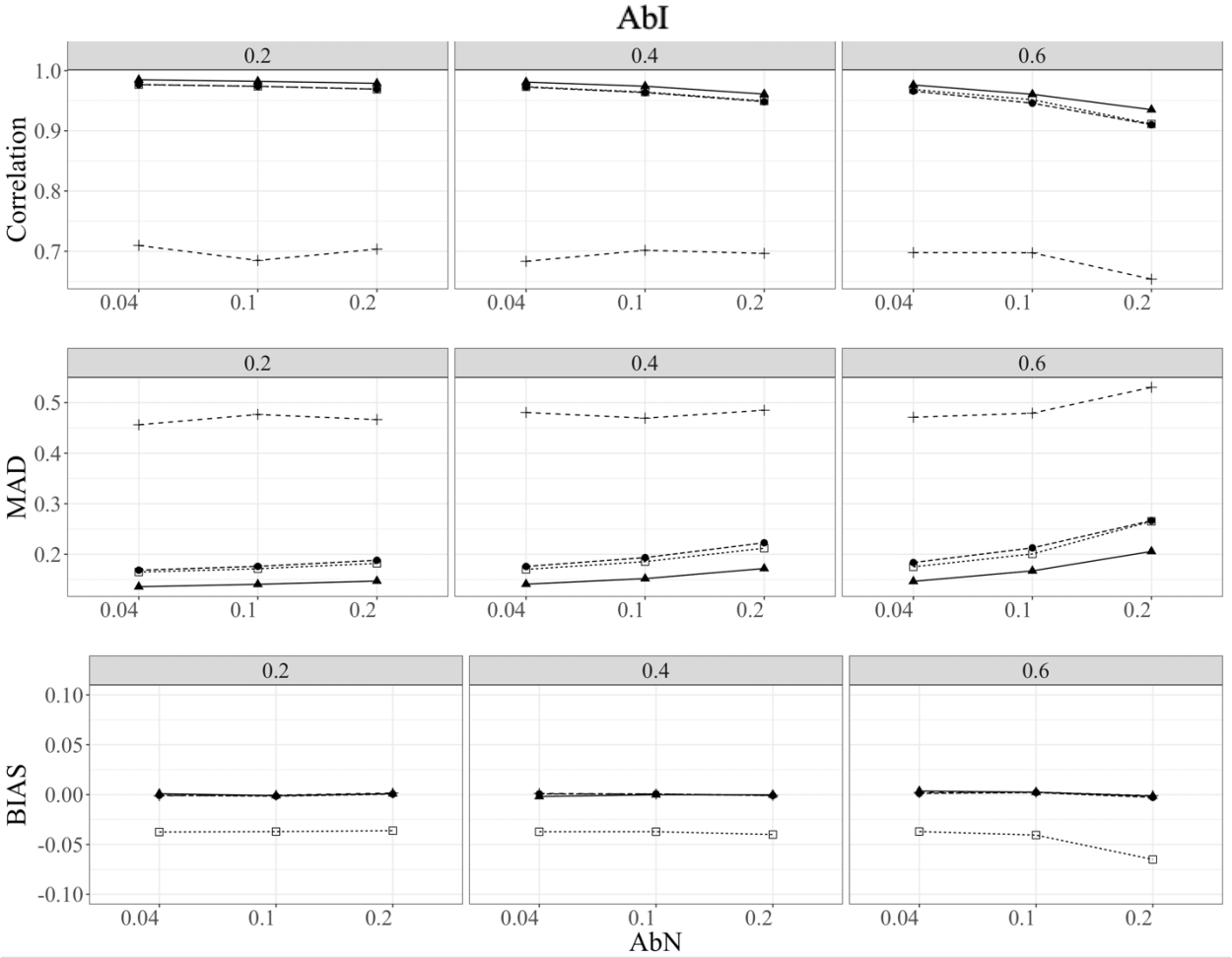
Item location parameters were also strongly correlated to their true values across conditions (Tables 7, 8, and 9). For the GGUM, item location parameters (δ_j) were recovered well, with correlations at or above .99 even in the most contaminated datasets. Similarly, the b-

parameters in the GPCM had correlations at or above .93. Figure 12 shows the discrimination parameter recovery in both GPCM and GGUM conditions, where parameter estimates were very accurate in the datasets with no aberrant data when the appropriate model was used (r = .98 for both). Adding aberrant data decreased parameter recovery accuracy for the discrimination parameters in both cases, where the parameters were generally underestimated as reflected by the negative BIAS results. Comparing the discrimination parameter recovery in the clean datasets to the most aberrant datasets (*AbN* = .20; *AbI* = .60), when the appropriate model was used to fit the GGUM data, correlation between true and estimated parameters dropped from .98 to .85 and dropped from .98 to .77 in the GPCM datasets. The threshold parameters in the GGUM datasets seemed to be the most impacted by aberrant data. With the clean datasets, threshold parameter estimates correlated with the true parameters at .94, and with the most aberrant data condition this dropped to .67.

Cross-fitting conditions. For the conditions where the GGUM was fit to the GPCM data, and the GPCM was fit to the GGUM data, worse parameter recovery was anticipated due to the imposed misspecification on the conditions. However, when the GGUM was fit to the GPCM data (Table 7, section C; Figure 11 – hollow square), estimated person parameters correlated with their true values at .98 for the clean data. Even when *AbN* = .20 and *AbI* = .60, GGUM was able to recover person parameters for the GPCM data with a correlation between true and estimated parameters of .92. These correlations matched those when the correct model was used for the GPCM data. In contrast, when the GPCM was fit to the clean GGUM data (Table 1, section D; Figure 6 – plus sign), the correlation between estimated and true person parameters decreased to .77 (compared to the .98 using the appropriate GGUM model) and the MAD increased to .40 (compared to the .12 using the appropriate GGUM model). The quality of the

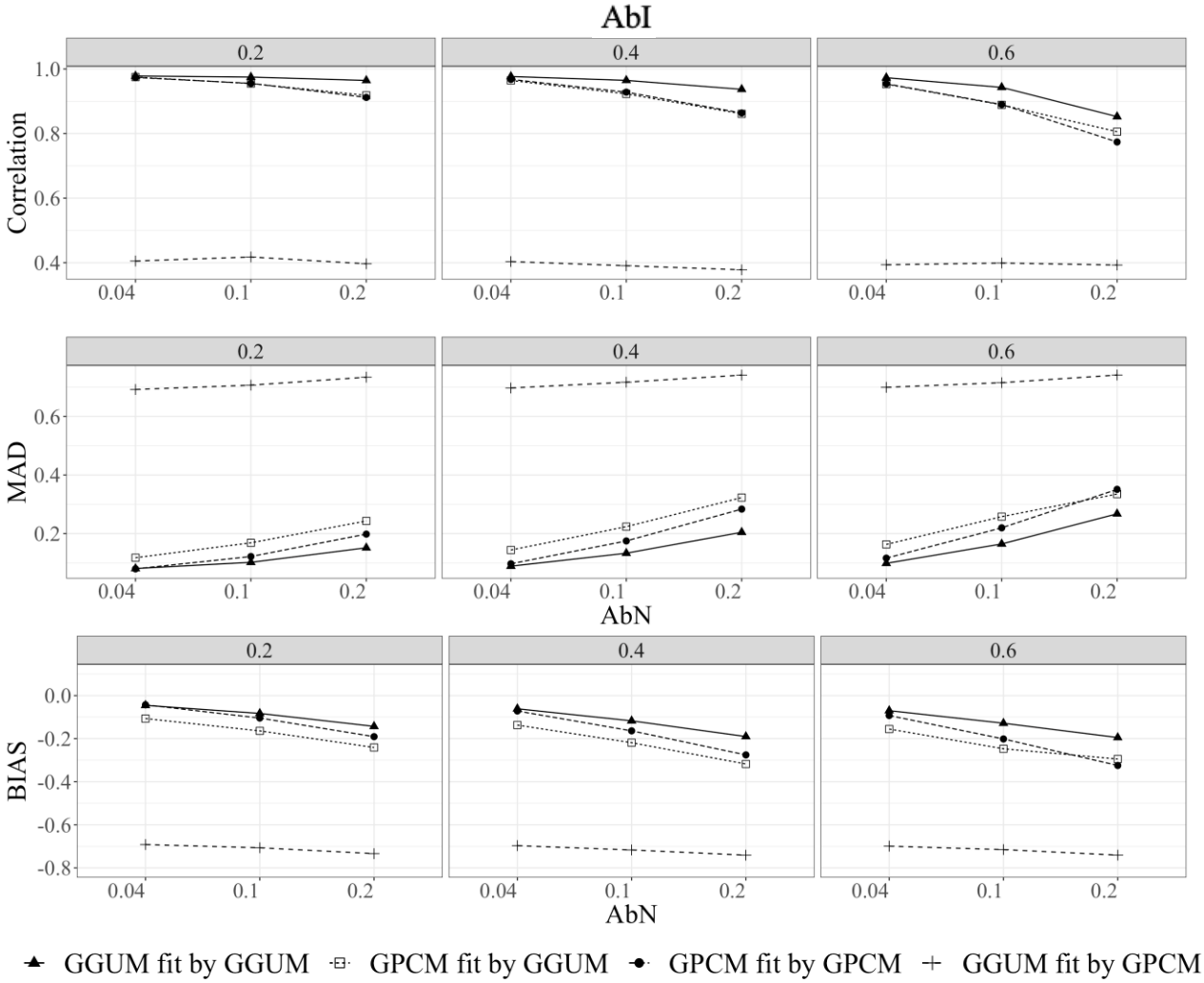
person parameter recovery for the aberrant datasets under this model misfit condition worsened substantially with a correlation of .70 and a MAD of .49. Figure 12 highlights the severity of the discrimination parameter underestimation that occurred when GPCM was fit to the GGUM data. The correlation, MAD, and BIAS of the discrimination parameter was substantially worse with the addition of aberrant responses when GPCM fit GGUM, and the correlation between true and estimated parameters was always below .45. However, when GGUM was fit to the GPCM data, the correlation between true and estimated discrimination parameters were nearly the same as when GPCM fit the GPCM data, and even slightly higher in the most aberrant condition.

Figure 11. Average Theta (Person) Parameter Recovery Measures by Condition



Note. Graphs reflect results for 20-item conditions; *AbI* = Proportion of items within an aberrant response vector that have aberrant responses. *AbN* = Proportion of simulees that were designated as aberrant responders.

Figure 12. Average Discrimination Parameter Recovery Measures by Condition



Note. Graphs reflect results for 20-item conditions; *AbI* = Proportion of items within an aberrant response vector that have aberrant responses. *AbN* = Proportion of simulees that were designated as aberrant responders.

Table 7. Quality of Parameter Recovery for *AbN* = .04

	$Data_{Clean}$			$Data_{Aberrant}$ (*AbI* = .2)			$Data_{Aberrant}$ (*AbI* = .4)			$Data_{Aberrant}$ (*AbI* = .6)		
Parameter	BIAS	MAD	COR	BIAS	MAD	COR	BIAS	MAD	COR	BIAS	MAD	COR
A. GGUM Fit to GGUM Data												
α_j												
M	-0.03	0.07	0.98	-0.09	0.11	0.96	-0.14	0.15	0.95	-0.16	0.19	0.92
(SD)	(0.03)	(0.02)	(0.01)	(0.04)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)	(0.06)	(0.05)	(0.07)
δ_j												
M	<0.01	0.06	1.00	<0.01	0.06	1.00	<0.01	0.06	1.00	<0.01	0.06	1.00
(SD)	(0.04)	(0.04)	(0.02)	(0.06)	(0.08)	(0.01)	(0.05)	(0.04	(0.01)	(0.04)	(0.02)	(0.01)
τ_{jk}												
M	-0.03	0.09	0.94	-0.03	0.09	0.94	-0.03	0.10	0.93	-0.03	0.10	0.93
(SD)	(0.03)	(0.02)	(0.05)	(0.03)	(0.05)	(0.07)	(0.03)	(0.02)	(0.05)	(0.03)	(0.02)	(0.05)
θ_n												
M	<0.01	0.12	0.99	<0.01	0.12	0.99	<0.01	0.13	0.99	<0.01	0.13	0.98
(SD)	(0.04)	(0.02)	(<0.01)	(0.04)	(0.02)	(0.01)	(0.04)	(0.02)	(0.01)	(0.04)	(0.02)	(0.01)
B. GPCM Fit to GPCM Data												
a_j												
M	0.01	0.06	0.98	-0.04	0.08	0.98	-0.07	0.10	0.97	-0.09	0.12	0.96
(SD)	(0.03)	(0.01)	(0.01)	(0.04)	(0.02)	(0.02)	(0.04)	(0.03)	(0.03)	(0.05)	(0.04)	(0.04)
b_j												
M	<0.01	0.09	0.99	<0.01	0.11	0.99	<0.01	0.14	0.98	<0.01	0.17	0.97
(SD)	(0.03)	(0.01)	(<0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.04)	(0.04)	(0.03)
θ_n												
M	<0.01	0.14	0.98	<0.01	0.15	0.98	<0.01	0.15	0.98	<0.01	0.16	0.97
(SD)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.03)	(0.01)
C. GGUM Fit to GPCM Data												
θ_n												
M	-0.03	0.14	0.98	-0.03	0.14	0.98	-0.03	0.15	0.98	-0.03	0.15	0.97
(SD)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)
D. GPCM Fit to GGUM Data												
θ_n												
M	<0.01	0.40	0.77	<0.01	0.40	0.77	<0.01	0.42	0.76	<0.01	0.41	0.77
(SD)	(0.03)	(0.26)	(0.33)	(0.03)	(0.26)	(0.33)	(0.03)	(0.27)	(0.34)	(0.03)	(0.25)	(0.32)

Table 8. Quality of Parameter Recovery for *AbN* = .10

	$Data_{Clean}$			$Data_{Aberrant}$ (*AbI* = .2)			$Data_{Aberrant}$ (*AbI* = .4)			$Data_{Aberrant}$ (*AbI* = .6)		
Parameter	BIAS	MAD	COR	BIAS	MAD	COR	BIAS	MAD	COR	BIAS	MAD	COR
A. GGUM Fit to GGUM Data												
α_j												
M	-0.03	0.07	0.98	-0.09	0.10	0.98	-0.12	0.13	0.97	-0.13	0.16	0.94
(SD)	(0.03)	(0.02)	(0.01)	(0.05)	(0.03)	(0.08)	(0.05	(0.04)	(0.08)	(0.08)	(0.05)	(0.15)
δ_j												
M	<0.01	0.06	1.00	<0.01	0.06	1.00	<0.01	0.06	1.00	<0.01	0.17	1.00
(SD)	(0.05)	(0.03)	(0.01)	(0.05)	(0.05)	(0.02)	(0.05)	(0.03)	(0.01)	(0.05)	(0.03)	(0.01)
τ_{jk}												
M	-0.03	0.09	0.94	-0.01	0.09	0.93	-0.01	0.11	0.91	-0.01	0.13	0.87
(SD)	(0.03)	(0.02)	(0.05)	(0.04)	(0.03)	(0.07)	(0.04)	(0.02)	(0.07)	(0.04)	(0.04)	(0.10)
θ_n												
M	<0.01	0.12	0.99	<0.01	0.13	0.99	<0.01	0.14	0.98	<0.01	0.15	0.97
(SD)	(0.04)	(0.02)	(<0.01)	(0.04)	(0.02)	(0.01)	(0.04)	(0.02)	(0.01)	(0.04)	(0.03)	(0.02)
B. GPCM Fit to GPCM Data												
a_j												
M	0.01	0.06	0.98	-0.10	0.12	0.96	-0.16	0.17	0.93	-0.20	0.22	0.89
(SD)	(0.03)	(0.01)	(0.01)	(0.05)	(0.04)	(0.03)	(0.07)	(0.06)	(0.06)	(0.08)	(0.07)	(0.08)
b_j												
M	<0.01	0.09	0.99	<0.01	0.15	0.98	<0.01	0.21	0.96	<0.01	0.27	0.93
(SD)	(0.03)	(0.01)	(<0.01)	(0.03)	(0.03)	(0.02)	(0.04)	(0.06)	(0.04)	(0.06)	(0.08)	(0.05)
θ_n												
M	<0.01	0.14	0.98	<0.01	0.15	0.98	<0.01	0.16	0.97	<0.01	0.18	0.95
(SD)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)
C. GGUM Fit to GPCM Data												
θ_n												
M	-0.03	0.20	0.98	-0.03	0.21	0.98	-0.03	0.24	0.97	-0.03	0.31	0.96
(SD)	(0.04)	(0.05)	(0.01)	(0.03)	(0.05)	(0.01)	(0.04)	(0.06)	(0.01)	(0.04)	(0.14)	(0.02)
D. GPCM Fit to GGUM Data												
θ_n												
M	<0.01	0.40	0.77	<0.01	0.42	0.75	<0.01	0.42	0.76	<0.01	0.44	0.75
(SD)	(0.03)	(0.26)	(0.32)	(0.03)	(0.28)	(0.36)	(0.03)	(0.25)	(0.32)	(0.03)	(0.25)	(0.31)

Table 9. Quality of Parameter Recovery for *AbN* = .20

	$Data_{Clean}$			$Data_{Aberrant}$ (*AbI* = .2)			$Data_{Aberrant}$ (*AbI* = .4)			$Data_{Aberrant}$ (*AbI* = .6)		
Parameter	BIAS	MAD	COR	BIAS	MAD	COR	BIAS	MAD	COR	BIAS	MAD	COR
A. GGUM Fit to GGUM Data												
α_j												
M	-0.03	0.07	0.98	-0.14	0.15	0.97	-0.19	0.20	0.94	-0.20	0.27	0.85
(SD)	(0.03)	(0.02)	(0.01)	(0.05)	(0.04)	(0.03)	(0.08)	(0.07)	(0.04)	(0.14)	(0.07)	(0.10)
δ_j												
M	<0.01	0.06	1.00	<0.01	0.06	1.00	<0.01	0.07	1.00	<0.01	0.11	0.99
(SD)	(0.05)	(0.04)	(0.02)	(0.05)	(0.04)	(0.01)	(0.05)	(0.03)	(0.02)	(0.06)	(0.08)	(0.04)
τ_{jk}												
M	-0.04	0.09	0.94	-0.01	0.11	0.91	-0.01	0.15	0.81	-0.01	0.21	0.67
(SD)	(0.03)	(0.02)	(0.06)	(0.06)	(0.03)	(0.07)	(0.06)	(0.05)	(0.16)	(0.06)	(0.10)	(0.32)
θ_n												
M	<0.01	0.12	0.99	<0.01	0.13	0.98	<0.01	0.16	0.97	<0.01	0.19	0.94
(SD)	(0.04)	(0.02)	(<0.01)	(0.04)	(0.02)	(0.01)	(0.04)	(0.04)	(0.03)	(0.04)	(0.05)	(0.04)
B. GPCM Fit to GPCM Data												
a_j												
M	0.01	0.06	0.98	-0.19	0.20	0.92	-0.28	0.28	0.86	-0.33	0.35	0.77
(SD)	(0.03)	(0.01)	(<0.01)	(0.08)	(0.07)	(0.07)	(0.11)	(0.09)	(0.10)	(0.13)	(0.11)	(0.13)
b_j												
M	<0.01	0.09	0.99	<0.01	0.15	0.98	<0.01	0.21	0.96	<0.01	0.27	0.93
(SD)	(0.03)	(0.01)	(<0.01)	(0.03)	(0.03)	(0.02)	(0.04)	(0.06)	(0.04)	(0.06)	(0.08)	(0.05)
θ_n												
M	<0.01	0.14	0.98	<0.01	0.17	0.98	<0.01	0.20	0.95	<0.01	0.25	0.92
(SD)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.03)	(0.05)	(0.05)
C. GGUM Fit to GPCM Data												
θ_n												
M	-0.03	0.14	0.98	-0.03	0.16	0.98	-0.03	0.19	0.96	-0.05	0.24	0.92
(SD)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.01)	(0.04)	(0.03)	(0.02)	(0.06)	(0.08)	(0.05)
D. GPCM Fit to GGUM Data												
θ_n												
M	<0.01	0.40	0.77	<0.01	0.41	0.77	<0.01	0.44	0.75	<0.01	0.49	0.70
(SD)	(0.03)	(0.25)	(0.32)	(0.03)	(0.24)	(0.31)	(0.03)	(0.23)	(0.30)	(0.03)	(0.25)	(0.32)

Discussion

In this study, data simulated using unfolding IRT and dominance IRT models, specifically the GGUM and the GPCM, were compared for model fit under varying conditions of aberrant data contamination. One of the first tasks in empirical research, prior to testing model fit, is to test model assumptions. Both the GPCM and GGUM models assume that the probability of a response is a function of a single underlying latent trait or a unidimensional composite of skills. Literature regarding how to assess dimensionality with cumulative, dominance IRT models like the GPCM is immense, but still lacking with regards to unfolding models like the GGUM. Results for the uncontaminated GGUM datasets coincided with other studies that have found an additional spurious factor appearing for unfolding data (Tay et al., 2011; Tay & Drasgow, 2012; Williams, 2015). In this study, a second factor was also identified in the clean GPCM datasets approximately 11% of the time for the 20-item condition and 85% of the time for 40 items. Other research has found that the likelihood of falsely identifying a second factor may increase with test length in dominance data fitting a 2-parameter logistic model (Gessaroli & De Champlain, 1996).

This study adds to the literature by comparing the impacts of different types of aberrant responses on the dimensionality of the data. Even with 20% of the sample responding with random responses, MRS, ERS, or a combination of various types of aberrant responding, trends in factor retention did differ from the trends for clean GGUM datasets when using parallel analysis. The one type of aberrant responding that impacted factor structure for the GGUM data was longstrings which resulted in an extra factor, especially with the 40-item condition. Aberrant responses in the dominance (GPCM) datasets had a stronger impact on dimensionality findings, with longstrings having the greatest influence to increase dimensionality. However, MRS and

random responding also displayed an increase in the number of factors identified in the 20-item condition, with ERS having a smaller impact. Additionally, including more variables (40 items instead of 20) resulted in additional factors being retained for all aberrant conditions for the GPCM datasets, but not for the GGUM datasets (with the exception of the condition with longstringers). Overall, longstring aberrant responders have the largest impact on increasing dimensionality results for both dominance and ideal-point response data.

The results for both AIC and BIC in this study are consistent with the notion posited by other studies that the GGUM is able to fit GPCM data as well as the GPCM model, but the GPCM model is not able to fit the GGUM data as well as the GGUM model (Chernyshenko et al., 2007; Stark et al., 2006). Separating the results by the degree and type of aberrant responding revealed that the impact on model-fit based on the information criteria depends on the type of aberrant responses. As observed in prior studies with dominance data, increases in random responders resulted in poorer model fit for both GPCM and GGUM data (Liu et al., 2019). Although longstrings did not have as large of an impact on model fit as random responding (according to AIC and BIC), the results indicated poorer model fit as their presence in the data increased. These results do not correspond to research that has found increased reliability for empirical data with longstrings included (DeSimone & Harms, 2017), however differences in the item characteristics, percentage of response strings, and models being tested may impact these results. Future research is needed to better understand when longstrings have a positive versus negative effect on model fit.

ERS and MRS appear to have very different impacts on GGUM data compared to GPCM data. In the current study, the more extreme conditions of ERS in a GGUM dataset resulted in an improved model fit, whereas adding MRS to the GGUM datasets resulted in worse fit. Although

both ERS and MRS were detrimental to model fit for GPCM data, the trend for these two forms of aberrant responding were the opposite of that on GGUM data with ERS having a more negative impact. Further research is needed to investigate this phenomenon more thoroughly.

Regarding the adjusted χ^2/df ratios for doublets and triplets of items, the simulation results suggest that the method is useful for assessing the relative fit for the GPCM versus the GGUM in most cases when GGUM data is being used. That is, the flag rates consistently pointed to the correct model for the clean datasets and conditions where the proportion of aberrant response strings was .1 or less. This supports the findings from Tay et al. (2011) where only clean datasets were used. Overall, model misfit is identified at a much higher rate for GPCM applied to GGUM data as compared to GGUM applied to GPCM data. This study adds to the literature by confirming the χ^2/df ratio method is useful in pointing to the correct model applied to the GGUM data, even when aberrant responses were included. The procedure identifies poor data while still pointing to the best model for the data. However, the procedure may hit a ceiling effect when there is a fairly large amount of MRS (e.g., when samples consist of 20% aberrant respondees who respond to at least 60% of the items using MRS). In comparison, with GPCM data, both the GGUM and GPCM models provided similar results when using the χ^2/df ratio method indicating that the GGUM model may fit GPCM data and detect the presence of aberrant responses within the data equally as well as GPCM.

Finally, results based on the quality of parameter recovery indicated that when the GGUM was appropriately applied to GGUM data, person parameter recovery was acceptable under all conditions with correlations between estimated and true parameters ranging from .94 in the most extreme aberrant conditions for the study to .99 in the clean datasets. Person parameter recovery using GPCM on GPCM data was slightly more affected by aberrant data with

correlations ranging from .92 in the most extreme aberrant conditions to .98 in clean datasets. Similar to previous research on aberrant responding in GPCM datasets (Jin et al., 2018), the GPCM underestimated the discrimination parameter for items in GPCM datasets where aberrant responding was present. This finding was also apparent when fitting GGUM to the GGUM datasets with aberrant responses, though to a slightly lesser degree. Examination of parameter recovery under cross-fitting conditions gave further evidence that the GGUM can fit GPCM data relatively well under clean and less extreme aberrant conditions. On the contrary, the GPCM was not able to recover person parameters as effectively for GGUM data. Although this finding has been shown using fit plots and correlations in other studies (Chernyshenko et al., 2007; Stark et al., 2006), this study revealed that this phenomenon was evident even when aberrant responses were relatively low (*AbN* = .04 and *AbI* = .20).

Limitations

As with many simulation studies, the choice of conditions is a natural and important limitation to the study. First, only two models are compared (GGUM, GPCM). Future research should consider other polytomous unfolding and dominance IRT models. Additionally, the types of aberrant responding and ranges of contamination levels (*AbI* and *AbN*) could be extended. For example, spuriously high and spuriously low responding are a common focus in several studies investigating aberrant responding (Li & Olejnik, 1997; Y. Liu et al., 2009; Tendeiro & Meijer, 2014; Xia & Zheng, 2018). The item parameter recovery was adequate when using the appropriate model to fit the data, but the condition with the highest proportion of aberrant respondees in a sample was 20%. Although conditions were chosen to mimic realistic and common conditions researchers face in empirical studies, some have estimated the proportion of aberrant responding in a sample to be as high as 60% (Berry et al., 1992). Item parameter

recovery in conditions where higher proportions of aberrant responding is present would add to the research. Further, the way aberrant responding was simulated in combination with the characteristics of the generated data could have had an impact on results. For example, there were more generated extreme scores (0,5) than generated midpoint scores (3, 4) in the GGUM datasets. That is, the ratio of extreme to middle scores ranged from 1.40 to 2.08, which could have impacted the relative magnitudes of the MRS and ERS effects on the GGUM data. Lastly, 100 replications of each condition were conducted in this study. Ideally one might include a higher number per condition, however the computer time needed per condition for the GGUM parameter estimation made more than 100 replications prohibitive of realistic completion times. Comparisons across studies with GGUM analyses will be important for interpreting the stability of results.

Conclusions

As mentioned previously, researchers will not know if model-data misfit is due to the presence of aberrant responding, or if it is due to model misspecification. An advantage of the current simulation study was the control over these factors. Aberrant responding and model misspecification (based on the underlying response process) was manipulated to provide insight on the impacts of these factors that may arise with real data. As results demonstrated how aberrant data can severely impact model fit, it is suggested to carefully examine the quality of the data before making conclusions about model-data fit or misfit. As aberrant data has been shown to affect model fit, and model fit may affect the detection of aberrant responding, close investigation of these factors is warranted in future research in the field. With most aberrant data research conducted with dominance response models and considering the finding that an unfolding IRT model framework may fit both ideal point and dominance response data

effectively, further research is recommended that studies the effectiveness of current aberrant response data detection under an unfolding model framework.

CHAPTER 5

STUDY 2

Performance of Nonparametric Person-Fit Statistics with Unfolding versus Dominance Response Models

ABSTRACT

Person-fit analyses are commonly used to detect aberrant responding in self-report data. Nonparametric person fit statistics such as G^P (Molenaar, 1991), G_N^P (Emons, 2008), H^T (Sijtsma, 1986; Sijtsma & Meijer, 1992), and $U3^P$ (Van der Flier, 1980) do not require fitting a parametric test theory model and have been shown to perform well in comparison with other person-fit statistics in the context of dominance response data (Emons, 2008; Karabatsos, 2003; Tendeiro & Meijer, 2014; Turner, 2018). However, ideal point response models are increasingly being applied to non-cognitive constructs(Freund & Lohbeck, 2021; Jin et al., 2021; Kartal & Dirlik, 2021; Kutlar et al., 2020). As these models do not exhibit the same relationship between item-level responses and latent traits as dominance models, person-fit statistics that were developed for dominance models may not be as effective for ideal point data. Further, detection of different types of aberrant responding has primarily focused on dominance response data, thus the types of impacts different aberrant behaviors have on the detection rates of person-fit statistics applied to ideal point data is unclear. This study demonstrates the performance of nonparametric person-fit statistics in detecting aberrant responding under an unfolding model context in comparison to a dominance context. Results for dominance data indicate that increases in detection rates depend, among other factors, on the type of aberrant responding and the person-fit statistic used. Detection of random responding, longstrings, and extreme response style increases for three person-fit statistics (G^P, G_N^P, $U3^P$) when the proportion of aberrant responses within a

participant's response vector increases while the overall proportion of participants with aberrant data decreases. G^P was most effective in identifying random responding, whereas G_N^P was most effective for extreme response styles. None of the person-fit statistics effectively identified midpoint response style in our dominance data conditions. In comparison, the detection of aberrant responses in ideal point response data was ineffective using the four nonparametric person-fit statistics, with slightly higher type I error and power less than 0.25. Additional research is needed to identify or develop nonparametric or parametric person-fit statistics effective for different types of aberrant behavior exhibited in ideal point response data.

Performance of Nonparametric Person-Fit Statistics with Unfolding versus Dominance Response Models

With the increased use of attitudinal and personality measures in organizational behavior and human resource management, the detection of aberrant responding has become increasingly important for many applied researchers and analysts. Aberrant responding has been shown to affect psychometric properties of a scale and ultimately may lead to erroneous measurements and classifications for participants (DeSimone et al., 2018; McGrath et al., 2010). Person-fit analyses have become a popular option for detecting aberrant responding. In the field of psychometrics, person-fit refers to how well participants' responses fit what would be expected based on the stipulated measurement model (Sijtsma & Meijer, 2001). This is not to be confused with person-job or person-organization fit, which refers to how well a person or applicant will fit with a particular job or organization. Rupp (2013) notes that applications of person-fit statistics have gained popularity with measures in many fields, including assessments of personality, attitudes, health outcomes, psychological traits, and education-related characteristics or skills.

Most person-fit analyses have been investigated with data assumed to fit a dominance response model (Armstrong et al., 2007; Conijn et al., 2014; Dimitrov & Smith, 2006; Emons, 2008; Glas & Meijer, 2003; Karabatsos, 2003; Sijtsma & Meijer, 2001; St-Onge et al., 2011; Tendeiro & Meijer, 2014, 2014; Turner, 2018), where it is assumed that the probability of endorsement monotonically increases as the underlying latent trait increases. However, a growing body of literature suggests that non-cognitive data is often better fit by models that assume an ideal point response process, also known as unfolding models (Chernyshenko et al., 2007; Drasgow et al., 2010; Stark et al., 2006; Weekers & Meijer, 2008). The ideal point response process considers the distance between person and item locations on the underlying

continuum for the construct being measured. Researchers have used unfolding models to successfully describe non-cognitive data from assessments of creativity using the Gough's Creative Personality Scale (Zampetakis, 2010), conscientiousness (Carter et al., 2014), personality inventory of self-judgement on the order-facet (a feature of conscientiousness; Chernyshenko et al., 2007; Weekers & Meijer, 2008), 16 personality factor subscales (Stark, 2006), control preferences in medical contexts (Control preferences Scale; Degner et al., 1997), attitude and affect constructs (LaPalme et al., 2018), censorship data (C.-W. Liu & Wang, 2019), and job satisfaction as measured by the Job Descriptive Index (Carter & Dalal, 2010). Nonetheless, dominance models are more widely used than unfolding models. As such, further research to understand the application of unfolding models with non-cognitive data, and the use of data management procedures such as person-fit analyses, is warranted.

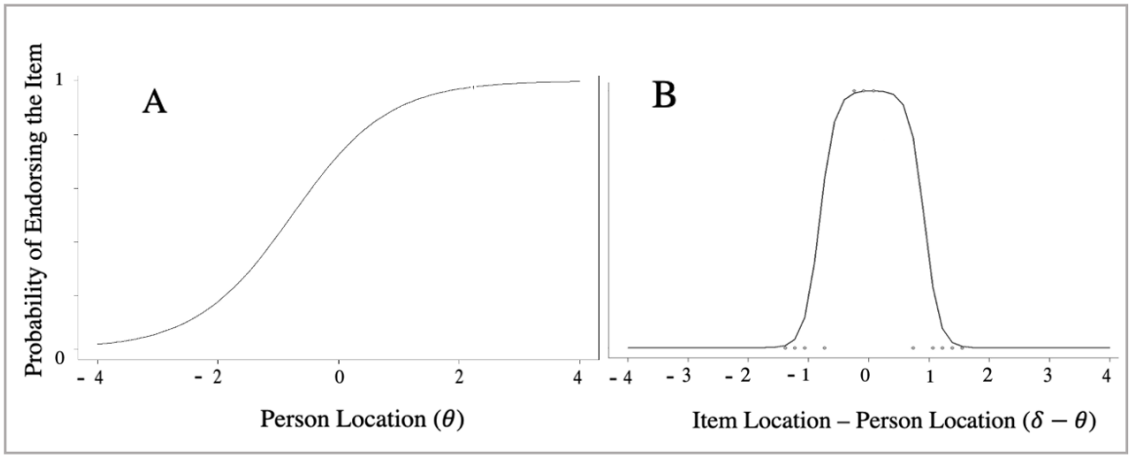
Despite a considerable number of studies that have investigated the impact of assuming an incorrect response process on model fit, little attention has been placed on the performance of person-fit statistics in the context of different underlying response processes. Applying person-fit statistics identified as being most effective with dominance response model data may not be appropriate for ideal point response data due to the differing distribution assumptions. Further, there is a dearth of research on the identification of different types of aberrant responding with ideal point response models using person-fit statistics. The aim of the current study is to examine the performance of several polytomous nonparametric person-fit statistics in identifying four types of aberrant response strings when applied to data that fit an ideal point model vs a dominance response model.

Literature Review

Ideal Point Versus Dominance Response Processes

The way participants respond to items may differ based on the type of relationship between the participant's and item's location on a continuum representing the construct of interest. In the dominance response process, as the participant's ability or trait level increases, the probability of endorsing an item increases regardless of the item's location on the continuum. For example, when measuring perceived social skills with items like "My social skills are at least as good as those of an average person," it may be reasonable to assume that the higher the perceived social skills a respondent has, the more likely he or she will endorse this item (example from Chernyshenko et al., 2007, p. 104). In contrast, the ideal point process is based on a notion conceptualized by Thurstone (1928) and termed by Coombs (1964), that assumes that the probability of endorsing an item increases as the item more closely reflects the person's standing on the construct being measured. With this approach, the probability of endorsing an item increases as the difference between the person's and item's location on the continuum representing the construct decreases. Thus, a notable difference between dominance and ideal point models can be illustrated using their respective item response functions (IRFs). Figure 13 provides a visual representation of these differences for dichotomous items.

Figure 13. IRFs Based on Dominant (A) and Ideal Point (B) Models



Note. Figure A demonstrates a monotonically increasing item response function (IRF), where the probability of endorsing the item increases with increasing person locations. Figure B demonstrates a non-monotonic IRF, where the probability of endorsing the item is the greatest when the person and item locations match.

When items are polytomous, intermediary step functions are used to model the transition, or "stepping," to successively higher score categories. Four common approaches for defining step functions include adjacent category, continuation ratio, cumulative, and nominal (Penfield, 2014). Within the adjacent category approach for dominance models is the generalized partial credit model (GPCM; Muraki, 1992), where step functions are specified by the two-parameter logistic model (Birnbaum 1957; as cited in Hambleton & Swaminathan, 1985) and an item-level discrimination parameter is estimated. Many unfolding models use the same or similar approaches to defining step functions in the case of fitting polytomous data. A number of probabilistic unfolding models have been developed over the past few decades, including the Squared Logistic Model (SLM; Andrich, 1988), PARELLA model (Andrich, 1988), Hyperbolic Cosine Model for dichotomous data (HCM; Andrich & Luo, 1993), General Hyperbolic Cosine Model for polytomous data (GHCM; Andrich, 1996), Graded Unfolding Model (GUM; Roberts

& Laughlin, 1996), and the Generalized Graded Unfolding Model (Roberts et al., 2000). The GGUM is one of the more widely used unfolding models and includes characteristics similar to the GPCM that make it an optimal choice for comparison. GGUM can be used with both polytomous and dichotomous data. Discrimination parameters and the thresholds for each response option are allowed to vary across items. This allows the IRFs to exhibit different distributional shapes across items (Stark et al., 2006). Further, the GGUM package in R (Tendeiro & Castro-Alvarez, 2019) is open source and freely accessible to researchers. Although many unfolding models have been used for ideal point data, the GGUM has been recognized as a very popular choice for applied studies in the non-cognitive field (Joo et al., 2019). Because the current study used the GGUM proposed by Roberts et al. (2000), where each subjective response follows the Generalized Partial Credit Model (GPCM; Muraki, 1992), it was decided to use the GPCM to model the dominance data to maximize comparability of results.

Aberrant Data Types

Aberrant responding can stem from several behaviors that include providing invalid responses due to insufficient effort and providing responses influenced by factors other than content. Random responding and invariant responding are two very common forms of insufficient effort (DeSimone et al., 2018). Both types of responding are characterized by answering items without regard to item content. Invariant responding is answering with the same option for a number of successive items, also known as longstring responding (Huang et al., 2012; Johnson, 2005; Karabatsos, 2003). In comparison, random responders (often labeled as careless, haphazard, or inconsistent responders) vary their responses, possibly in attempt to simulate attentive responding (DeSimone et al., 2018; Huang et al., 2012; McGrath et al., 2010). Additionally, some participants allow response styles to influence their response choices which

can result in misfitting item scores. Extreme response style (ERS) refers to people tending to choose the upper or lower extreme categories, regardless of the item content (Greenleaf, 1992). People who tend to choose the middle response option are deemed as exhibiting a middle response style (MRS; Baumgartner & Steenkamp, 2001), or mid-lining. These types of aberrant responding have been a concern in a variety of contexts such as organizational psychology (Huang & DeSimone, 2021; Schroeders et al., 2022), student evaluations (AlQuraan, 2019), personality inventories (Huang et al., 2012; Johnson, 2005; Niessen et al., 2016) and attitudinal surveys (Bowling et al., 2016; Iaconelli & Wolters, 2020) as they contribute to the misrepresentation of participants' true construct levels. It is important for researchers to be able to detect these types of responses that increase measurement error in a dataset and to be aware of the potential adverse effects they may have on analyses and decisions made from those analyses.

Nonparametric Person-Fit Statistics

Person-fit statistics are one of many methods for examining response behaviors on cognitive and non-cognitive assessments. Two general types of person-fit statistics include parametric and nonparametric statistics. Parametric statistics are computed by measuring the disparity between the observed data and the estimated response predictions resulting from a parametric IRT model. In contrast, nonparametric person-fit statistics do not rely on parametric IRT-based models, but rather are computed from the observed response data (Karabatsos, 2003). The focus of this study is on the performance of four nonparametric person-fit statistics (H^T, $U3^P$, G_N^P, and G^P) that have been shown to perform well in comparison with other person-fit statistics, however primarily with data that have a dominance model structure (Emons, 2008; Karabatsos, 2003; Tendeiro & Meijer, 2014; Turner, 2018). Each are described briefly below.

G^p. When items conform to a Guttman scale, they are arranged hierarchically so that the endorsement of one item suggests the endorsement of items located lower on the underlying latent continuum of the construct being measured. The number of Guttman errors for polytomous items, to be defined below, can be summarized using the G^p statistic (Molenaar, 1991). First, the item step difficulties (π_{jx_j}) for step *x* of item *j*, are computed as the proportion of respondents who passed step x_j or higher on the item. Using an example similar to Emons (2008), suppose a four-category item (item 1) with response options 0 = strongly disagree, 1 = disagree, 2 = agree, and 3= strongly agree had step difficulties of $\pi_{11} = .75$, $\pi_{12} = .50$, and $\pi_{13} = .18$. This means that 75% of the respondents passed the first step (chose a category higher than the first option of strongly disagree), 50% passed the second step (chose an option higher than or equal to the third category), and 18% passed the third step (chose the fourth category). Now suppose item 2, with the same response options, had step difficulties equal to $\pi_{21} = .65$, $\pi_{22} = .36$, and $\pi_{23} = .08$. If a respondent selected 'disagree' for item 1 ($x_1 = 1$) and 'strongly agree' for item 2 ($x_2 = 3$), the ordered vector (based on item step difficulties from least difficult to most difficult) of item step scores for this respondent would be (***Y*** = ordered vector; y_k = element *k* of vector ***Y***):

Table 10. Example Item Steps Vector

Ordered item steps	$\pi_{11} = .75$	$\pi_{21} = .65$	$\pi_{12} = .50$	$\pi_{22} = .36$	$\pi_{13} = .18$	$\pi_{23} = .08$
Y	1	1	0	1	0	1

In summary, this respondent passed the second step of item 2 but failed to pass the "easier" (or more commonly selected) second step of item 1. Additionally, this respondent passed the third step of item 2 but failed to pass the "easier" third step of item 1. Because the G^p statistic counts all such pairwise Guttmann ordering errors of all possible item-step pairs, the G^p for this respondent would be equal to three. The formal equation for the G^p statistic is given by:

$$G^p = \sum_{l<k}^{JC} y_k(1 - y_l), \tag{39}$$

where *J* is the number of items which here are assumed to all have the same number of response categories (*V*), and thus the same number of item steps (*C*) for each item. If *C* = 1, the G^p statistic is specified for dichotomous items. In equation 1, y_l represents all elements of vector $\boldsymbol{Y}$ that are prior to y_k. The more Guttmann errors recorded for a participant, the greater the G^p statistic, indicating greater person misfit. In dominance data contexts, the G^p statistic has been shown to identify aberrant responding relatively well compared to other nonparametric person-fit statistics when the behavior reflects careless responding that is random or respondents failing to notice when items are worded in the opposite direction. However, G^p has not been as effective (compared to G_N^P and $U3^P$) in identifying ERS with dominance response data (Emons, 2008). Given the design of the fit statistic, we would expect G^p to effectively indicate misfit in dominance data where monotonically increasing response functions are assumed. However, it may not be expected to function as effectively for ideal point data where greater agreement is not necessarily expected for items located lower on the underlying latent continuum.

G_N^P. Because the maximum possible G^p depends on the sum score (X_+) of the respondent and the ordering of the item steps, it becomes difficult to compare G^p statistics across different X_+ scores. One solution is to norm the G^p statistic using the following equation (Emons, 2008):

$$G_N^P = \frac{G^p}{\max(G^p|X_+)}. \tag{40}$$

Thus, the normed number of Guttmann errors, G_N^P, is simply the G^p statistic after being normalized to have a range of [0, 1]. Unlike G^p, the G_N^P statistic has been demonstrated as being effective in identifying aberrant responding such as ERS with dominance data (Emons, 2008). In its dichotomous form, it has not been as effective in identifying cheating with dominance data (Karabatsos, 2003). Due to their popularity and effectiveness with certain conditions such as

careless and inattentive responders (e.g., Emons, 2008), both G_N^P and G^p are included in the current study. Similar to G^p, G_N^P assumes monotonically increasing IRFs, which could possibly result in poorer aberrant responding detection for ideal point data than data that reflect a dominance response process.

H^T. The H^T statistic (Sijtsma, 1986; Sijtsma & Meijer, 1992) is a modified version of Mokken's (1971) H_j index, which allows items to be scaled to the Guttman (1944) model. By transposing the item by person matrix, Sijtsma (1986) was able to apply the H_j index procedure to persons rather than items and detect respondents that do not conform to the Guttman model. The H^T statistic is computed from a matrix of *J* columns of items and *N* rows of participants, with each element in the matrix representing an item score. Suppose $\boldsymbol{X}_n$ represents the item-score vector composed of $j = 1 \ldots J$ item-scores for participant n. The total score for item *j* ($\boldsymbol{T}_j$) is computed as the sum of all participants' scores for that particular item. The vector $\boldsymbol{T}$ is composed of the total scores for all items (T_1 *to* T_J) and the vector $\boldsymbol{T}_n = \boldsymbol{T} - \boldsymbol{X}_n$. That is, $\boldsymbol{T}_n$ is the vector of all item-score totals excluding participant *n*. Finally, the H^T statistic for participant *n* is computed as follows:

$$H_n^T = \frac{Cov(\boldsymbol{X}_n, \boldsymbol{T}_n)}{Cov_{max}(\boldsymbol{X}_n, \boldsymbol{T}_n)}, \tag{41}$$

where $Cov(\boldsymbol{X}_n, \boldsymbol{T}_n)$ is the covariance between participant *n*'s item-scores and the item-score totals for the remaining participants in the sample (excluding that participant). $Cov_{max}(\boldsymbol{X}_n, \boldsymbol{T}_n)$ is the maximum covariance possible between $\boldsymbol{X}_n$ and $\boldsymbol{T}_n$ given the marginal distributions.

The H_n^T person fit statistic can be used to assess the degree to which a respondent's item responses match the same ordering as the item-score totals. This statistic is included in the study because it has been found to have superior detection efficiency in several studies (Beck et al., 2019; Karabatsos, 2003; Tendeiro & Meijer, 2014). In dominance data contexts, H^T has been

found suitable to detect cheating, creative responding, spuriously high, spuriously low, and careless responders relatively well compared to other person fit statistics (Karabatsos, 2003; St-Onge et al., 2011; Tendeiro & Meijer, 2014). The overall coefficient H^T can be used to summarize the individual H_n^T statistics for all participants in a sample (Ligtvoet et al., 2010):

$$H^T = \frac{\sum_n^N Cov(\boldsymbol{X}_n, \boldsymbol{T}_n)}{\sum_n^N Cov_{max}(\boldsymbol{X}_n, \boldsymbol{T}_n)}. \quad (42)$$

The overall H^T coefficient will be high if a clear ordering exists among items and the item response functions are spread apart. However, if item response functions overlap, the overall H^T coefficient will be relatively low and the H_n^T statistics will be less stable. If H^T for the overall sample is low, it may not be an appropriate indicator to use for the dataset.

$U3^P$. The $U3$ person-fit statistic (Van der Flier, 1980) was generalized for application to polytomous items by Emons (2008), resulting in the creation of the $U3^P$ statistic. This statistic is included in the investigation because several studies have found it to have comparative, if not better performance, than other parametric and nonparametric person-fit statistics (Emons, 2008; Karabatsos, 2003; Tendeiro & Meijer, 2014; Turner, 2018). The statistic can be defined in a few steps. First, the sum of the log-odds of the item step difficulties for the steps that were passed by the participant, $W(\boldsymbol{y})$, is computed as follows (Emons, 2008):

$$W(\boldsymbol{y}) = \sum_{k=1}^{JC} y_k \log\left(\frac{\pi_k}{1-\pi_k}\right), \quad (43)$$

where $\boldsymbol{Y}$ is an observed response vector for J items with C+1 response categories, y_k is the item-step score for item-step k (taking a value of 1 if step k is passed or 0 if step k is not passed), and π_k is the item-step difficulty for item-step k. Next, $W(\boldsymbol{y})$ is normed which results in the $U3^P$ person-fit statistic as follows:

$$U3^P = \frac{\max(W|X_+) - W(\boldsymbol{y})}{\max(W|X_+) - \min(W|X_+)}, \quad (44)$$

where X_+ is the sum score computed as $X_+ = \sum_{k=1}^{IV} y_k$. The $\max(W|X_+)$ can only be obtained if the following holds:

$$\max(W|X_+) = \sum_{k=1}^{X_+} logit(\pi_k). \quad (45)$$

The $\min(W|X_+)$ cannot be expressed in closed form due to the structural dependencies between the item-step scores. That is, based on Guttman scaling principals, it is assumed that passing a step for an individual item means that all easier steps of that same item are also passed. Emons (2008) proposed to use a recursion algorithm to compute $\min(W|X_+)$. For more details on the recursion algorithm, see the Appendix from Emons (2008). Several studies have demonstrated how $U3^P$ may be an effective person-fit statistic to detect careless responding that is random, spuriously high or low, longstring, and mixed aberrant responding (Emons, 2008; St-Onge et al., 2011; Turner, 2018). Rudner (1983) showed how the $U3^P$ statistic was effective in identifying spuriously high and low scores on longer tests (e.g., 85 items) but not on "shorter tests" (e.g., 45 items).

A researcher's motivation to evaluate the fit of a model to their data is often rooted in their intention to use that model for the estimation of true scores for a latent trait versus when they use observed composite scores. As researchers are likely to apply nonparametric person-fit statistics to data that use observed composite scores for person-level outcomes, it is important to research the functioning of these person-fit statistics with data that more closely fit an ideal point response model than a dominance model. To the best of our knowledge, no nonparametric person-fit statistics have been created to detect the types of aberrant responding used in this study under an ideal point context and there is minimal research on the effectiveness of the current procedures.

Purpose

Aberrant responding has been demonstrated to have negative impacts on data outcomes (Clark et al., 2003; Credé, 2010; DeSimone et al., 2018; Woods, 2006) which can result in decreased accuracy of decisions made from those outcomes. It is important for researchers to be able to detect varying types of aberrant responses and to be aware of the potential adverse effects they may have on datasets and analyses. Because of the limited research available on the use of person-fit statistics with ideal point response data, researchers have recommended that studies investigate their use with unfolding frameworks (Drasgow et al., 2010; Lee et al., 2014; Liu & Zhang, 2020; Polak et al., 2012; Tendeiro, 2017). The purpose of the current study is to examine the performance of several polytomous, nonparametric person-fit statistics in identifying four types of aberrant response tendencies (random responding, long-stringing, ERS, MRS) when applied to data that fit one of two underlying response processes, the common dominance response model vs the less common ideal point response model. Although a large body of literature exists covering the performance of the aforementioned person-fit statistics, many have not been studied with the types of aberrant data in this study (see Chapter 2 Literature Review of this dissertation for a review of methodology for person fit analysis research). Additionally, including the dominance data in the current study offers a vis-à-vis between both types of data. Nearly all previous studies assume an underlying dominance response process. However, in the last fifteen years, an ideal point response process has been recognized as more appropriate than a dominance response process for several types of non-cognitive data and the increased use of unfolding models reflects that (Carter et al., 2014; Chernyshenko et al., 2007; Stark et al., 2006; Weekers & Meijer, 2008; Zampetakis, 2010). Including both types of data provides a means for comparison of the performance of these person-fit statistics in the two different contexts.

Because the ordering of persons based on latent trait scores may be severely affected by the underlying item response process (Stark et al., 2006), it is reasonable to question the applicability of the findings from previous person-fit studies using dominance IRT models to an unfolding model context. Misuse of these common nonparametric person-fit statistics with data that more accurately fit an ideal point response model may result in either false negatives or false positives for aberrant response strings. This study will attempt to inform decisions made for analytical procedures in settings where unfolding models are appropriate. The research questions for this investigation include:

1) How do the selected nonparametric person-fit statistics (H^T, $U3^P$, G_N^P, and G^P) perform in identifying aberrant responding under an unfolding model versus a dominance model framework?
    a. Do the data fit the minimum requirements for applying each statistic?
    b. How do the type I error and detection rates for each person-fit statistic for each type of aberrant behavior (random responding, longstrings, ERS, MRS) compare for unfolding and dominance data?
    c. How do the trends and magnitudes for detection and type I error rates compare for the study conditions (e.g., test lengths, contamination levels) under unfolding and dominance frameworks?

Methods

To investigate the research questions, data were generated using a fixed sample size of 1,000 for 6-point item responses. Five conditions were varied: type of aberrant responding, proportion of aberrant responders, proportion of aberrant responses within a response vector, test length, and data model. Five types of aberrant responding (*AbType*) were simulated: extreme

response style (ERS), midpoint response style (MRS), random responding, longstrings, and a 'mixed aberrant response' condition which combined all four aberrant response types. The mixed *AbType* condition was used to simulate realistic situations where a sample may be composed of several types of aberrant responders simultaneously. Three proportions of aberrant responders with misfitting item scores were considered (*AbN* = .04, .10, .20). To simulate the response vectors for each aberrant respondent, three proportions of aberrant responses to the items within each aberrant response string were used (*AbI*= .20, .40, .60). Additionally, two test lengths of 20 and 40 items were included. Therefore, combined, the simulation study is based on a total of 2 (data generating mechanisms: GGUM and GPCM) × 3 (proportion of aberrant responders in the sample, *AbN*) × 3 (proportion of aberrant responses in response vectors, *AbI*) × 2 (test lengths) × 5 (types of aberrant responding and response styles) = 180 fully crossed conditions. The number of replications per condition was 100. Each replication for each condition began with the generation of perfect model-fitting data to obtain a baseline for "clean", non-aberrant data and for the purposes of computing type I error for the four person fit statistics (H^T, $U3^P$, G_N^P, and G^P). All code for generating and estimating model parameters, model fit statistics, and person-fit statistics was written in R (R Core Team, 2016) and is available on OSF.

The procedure for the study can be summarized in five steps:

1) Generate "clean" non-aberrant data using GGUM for the unfolding IRT model and GPCM for the dominance IRT model.
2) Generate aberrant data and replace "clean" response vectors with the designated random, longstring, ERS, MRS, or mixed aberrant response strings.
3) Test item ordering for the H^T procedure.
4) Compute the four person-fit statistics for each simulee.

5) Calculate the type I error rates (falsely identifying a simulee as aberrant when their responses were “clean”) and detection rates (correctly identifying a simulee as aberrant).

Methods for each of the steps are detailed below.

Data Generation (Step 1)

Person parameters were randomly drawn from the standard normal distribution for GGUM and GPCM datasets. For the GGUM datasets, the GenData.GGUM function from the GGUM package in R was used to generate all item and person parameters as well as the item scores (Tendeiro & Castro-Alvarez, 2021). The item discrimination parameters (α_j) were randomly sampled from a uniform distribution [0.5, 2.0]. The item location parameters (δ_j) were randomly sampled from the standard normal distribution truncated between -2.0 and 2.0. The truncation was implemented due to reports of extreme values of δ_i sometimes leading to issues of low accuracy and variability of MML estimates under the GGUM (Roberts & Thompson, 2011 as cited in Tendeiro, 2017). The locations of the threshold parameters (τ_{jk}), relative to the location of the *j*th item, were recursively generated using procedures described in Roberts et al. (2002).

Similar to the GGUM datasets, the GPCM data were simulated using item discrimination parameters sampled from a uniform distribution [0.5, 2.0] and item difficulty parameters sampled from the standard normal distribution $N(0,1)$. Item category thresholds, d_{jk}, for step k of item j were simulated by taking the sequential cumulative sum of five numbers drawn from a random uniform distribution between .3 and 1. Using this interval ensured that the distance between categories would be at least .30. If thresholds are too close, some categories may be chosen infrequently (Chalmers, 2012). Next, each number in the set of sequential cumulative

sums was transformed by centering around the mean. In order for the model to be identified, the initial item category threshold, d_{j0}, was set to 0 (Muraki, 1992). The sim_gpcm function in R (PP package; Reif & Steinfield, 2021) was used to simulate the GPCM response data.

Generation of Aberrant (Misfitting) Responses (Step 2)

To simulate random responding, 20%, 40%, or 60% (depending on *AbI* condition) of the responses in an aberrant response vector were replaced with randomly sampled integers drawn from a uniform distribution [0, 5]. To mimic longstring responses, first an initial starting position in the response vector was randomly generated. Next, a single integer was drawn at random from a uniform distribution [0, 5] and replaced the specified proportion of consecutive items ($AbI \times number\ of\ items$) starting at the randomly drawn initial position in the vector (DeSimone et al., 2018). In attempt to simulate MRS responses for 6-point items, endpoint item scores were replaced with the closest midpoint response. Item scores adjacent to the endpoints were also replaced with the closest midpoint response (i.e., on the 6-point scale ranging from 0 to 5, items scores of 0 and 1 were replaced with a 2, and scores of 4 and 5 were replaced with a 3) to model the behavior of any responses that were not midpoint values being changed to the nearest midpoint (Liu et al., 2017). To simulate responses that reflect ERS, the four middle item scores were changed to the corresponding endpoint responses (i.e., 1s and 2s were changed to 0 and 3s and 4s were changed to 5).

Testing Item Ordering (Step 3)

Before drawing conclusions from nonparametric person-fit statistics that depend on invariant item ordering, it is advised to test the ordering of items (Van der Ark, 2007). For example, if the overall H^T coefficient (summary of H_i^T for all participants) is less than 0.3, researchers suggest that invariant item ordering may be too unstable to be useful (Ligtvoet et al.,

2010). For every condition in the study, the overall H^T coefficient was computed and averaged across replications to assess how well the simulated items were ordered.

Computing Nonparametric Person-Fit Statistics (Step 4)

The values for $U3^P$, G_N^P, and G^P were computed for each simulee using the PerFit package in R (Tendeiro, Meijer, and Niessen, 2016). As previously described, the H^T statistic is essentially a modified version of Mokken's (1971) H_i statistic.

Cutoff Criteria for Aberrant Identification. A decision rule was needed to flag a simulee as aberrant or not. A common method used to determine the cutoff criteria for each person-fit statistic is to use the 5% quantile as the cutoff value (e.g., Emons, 2008; Magis et al., 2012; Tendeiro, 2017). In this study, item parameters estimated from each aberrant dataset were used to simulate 20 datasets, and the distribution of each person-fit statistic value was examined for each dataset. Within each of the 20 replications, a cutoff for each statistic was then determined by finding the value associated with the critical value for a 5% probability of a type I error. For example, for the G^p statistic, the more Guttmann errors a respondent has, the greater the G^p statistic, indicating greater person misfit. Thus, in the equation $P(G_p \geq \text{value}_{\text{critical}}) = .05$, the person-fit statistic critical value ($\text{value}_{\text{critical}}$) was used as the cutoff criteria. Next, the mean cut-off score for the 20 model-fitting datasets was then used as the cut-off for that replication and condition. Due to time restrictions only 20 replications were used in determining cutoffs. Ideally, more than 20 replications would be used in practice, however, results were considered sufficiently consistent for the current study (e.g., the mean and standard error for the $U3^P$ statistic under the GPCM model was 0.146 and 0.005, respectively). The highest standard error relative to its mean was for the H^T statistic under the GGUM, where the mean was -0.188 and

the standard error was 0.01. The mean and standard error for the cut-offs used for each person-fit statistic are reported in Table 11. This process was also conducted for a subset of clean datasets to compare type I error rates of uncontaminated data to varying levels of contamination.

Table 11. Means of Person-Fit Statistic Cut-off

	$U3^P$		G^p		G_N^P		H^T	
Condition	*M*	*(SE)*	*M*	*(SE)*	*M*	*(SE)*	*M*	*(SE)*
GGUM	0.525	(0.011)	2121.684	(51.488)	0.541	(0.011)	-0.188	(0.010)
GPCM	0.146	(0.005)	635.647	(9.945)	0.152	(0.003)	0.379	(0.009)

Evaluating Performance of Person-Fit Statistics (Step 5)

To assess the performance of each person-fit statistic, both type I error (false positive) and detection (true positive) rates were computed. Type I error was computed in two ways for comparison. First, type I error was computed for all simulees incorrectly flagged as being aberrant within a clean data condition. This was used as a base condition for comparison. Second, type I error was computed as the proportion of non-aberrant simulees that was incorrectly flagged as being aberrant by the person-fit statistic in the datasets that included aberrant response strings. Detection rates (true positives) were computed as the proportion of aberrant simulees that was correctly flagged as aberrant by the person-fit statistic. Additionally, accuracy rates are summarized and reported. Accuracy was computed as the sum of correctly classified simulees (true positives and true negatives), divided by the total sample size. All rates were averaged over all conditions and replications. While accuracy may be a useful measure in providing a different perspective regarding the rate at which participants are correctly classified, this measure may be skewed when the proportions of aberrant and non-aberrant participants are not equal. For example, in conditions where only 4% of the sample was simulated to be aberrant, a method that does not flag anyone would still have a 96% accuracy rate. However, these values are provided as a means to compare performance of the same condition across methods.

Results

Before examining the results, item ordering was tested using the overall H^T coefficient [Figure 14A (20 items) and Figure 14B (40 items)]. If the overall H^T coefficient is less than 0.3, researchers suggest that invariant item ordering may be too unstable to be useful (Ligtvoet et al., 2010). For each condition applied to the GPCM datasets, the average overall H^T coefficient was above this criterion, with the lowest value of 0.41 for 20-item datasets having 20% aberrant responders with 60% of aberrant response vectors containing random responding. The GPCM datasets with only 4% aberrant simulees with 20% aberrant items replaced using MRS, had the highest average overall H^T coefficient (0.56). In comparison, the item ordering criterion was lower than recommended for several of the GGUM datasets, with average values ranging from 0.24 to 0.31. Random responding had the most negative impact on overall H^T for GGUM datasets. Similar to the GPCM datasets, the most extreme conditions of random responding (*AbN* = .20, *AbI* = .60) resulted in the lowest average overall H^T coefficient (0.24). Trends in the GPCM datasets, and to a slightly lesser degree in the GGUM datasets, indicated that item ordering decreased as the proportion of aberrant responders increased and as aberrant items within an aberrant response vector increased. The exception to this trend occurred with longstring vectors when the proportion of aberrant responses in a vector made up over half of the items (*AbI* = .60), in which case the coefficient H^T slightly increased. Increasing the number of items from 20 to 40 did not substantially impact the coefficient H^T in either type of dataset. In the following sections, the degree to which type I error and power is impacted for the person-fit statistics with the unfolding versus dominance data is investigated.

Figure 14A. Average Overall Coefficient H^T by Condition (20 items)

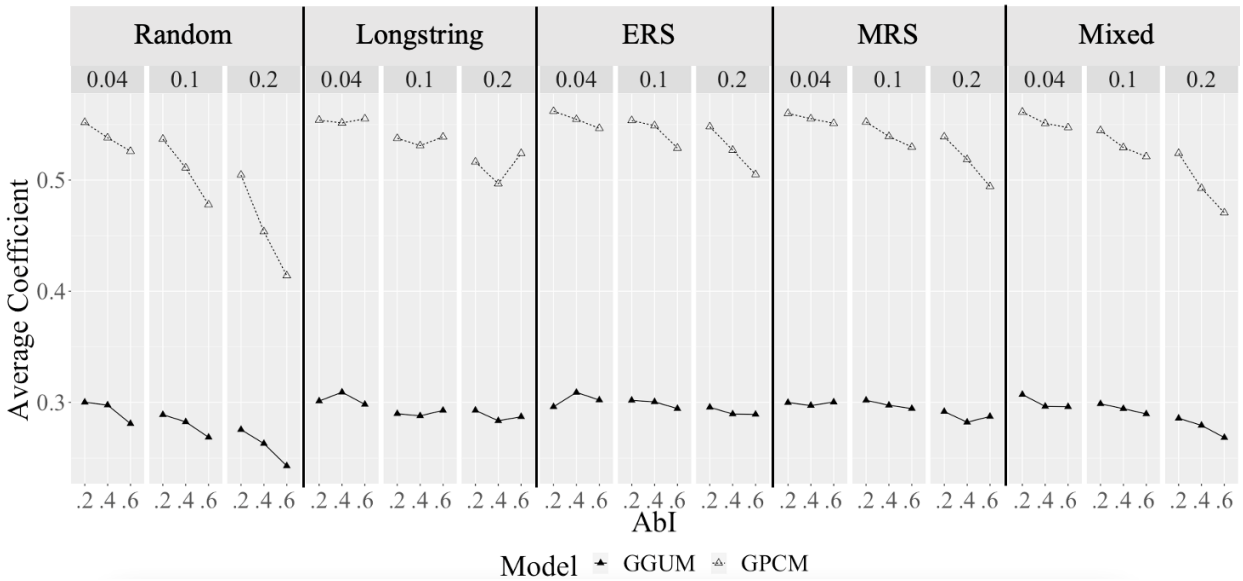
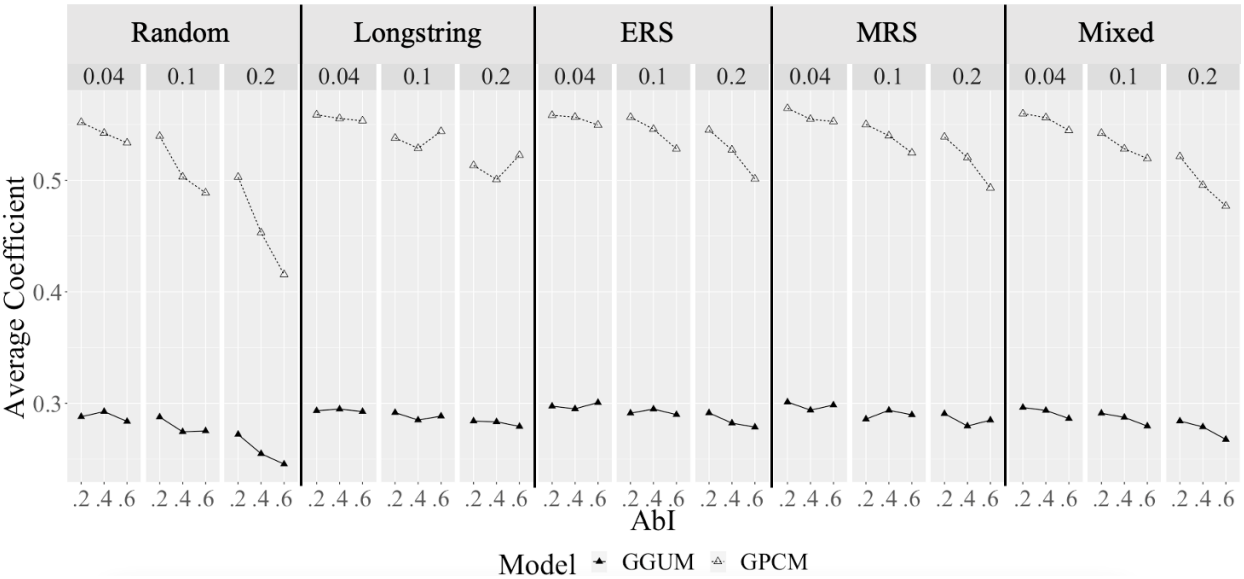


Figure 14B. Average Overall Coefficient H^T by Condition (40 items)



Note. ERS = Extreme Response Style. MRS = Midpoint Response Style. *AbI* = Proportion of items within response vector designated as aberrant. *AbN* = Proportion of simulees designated to have aberrant response vectors.

Type I Error

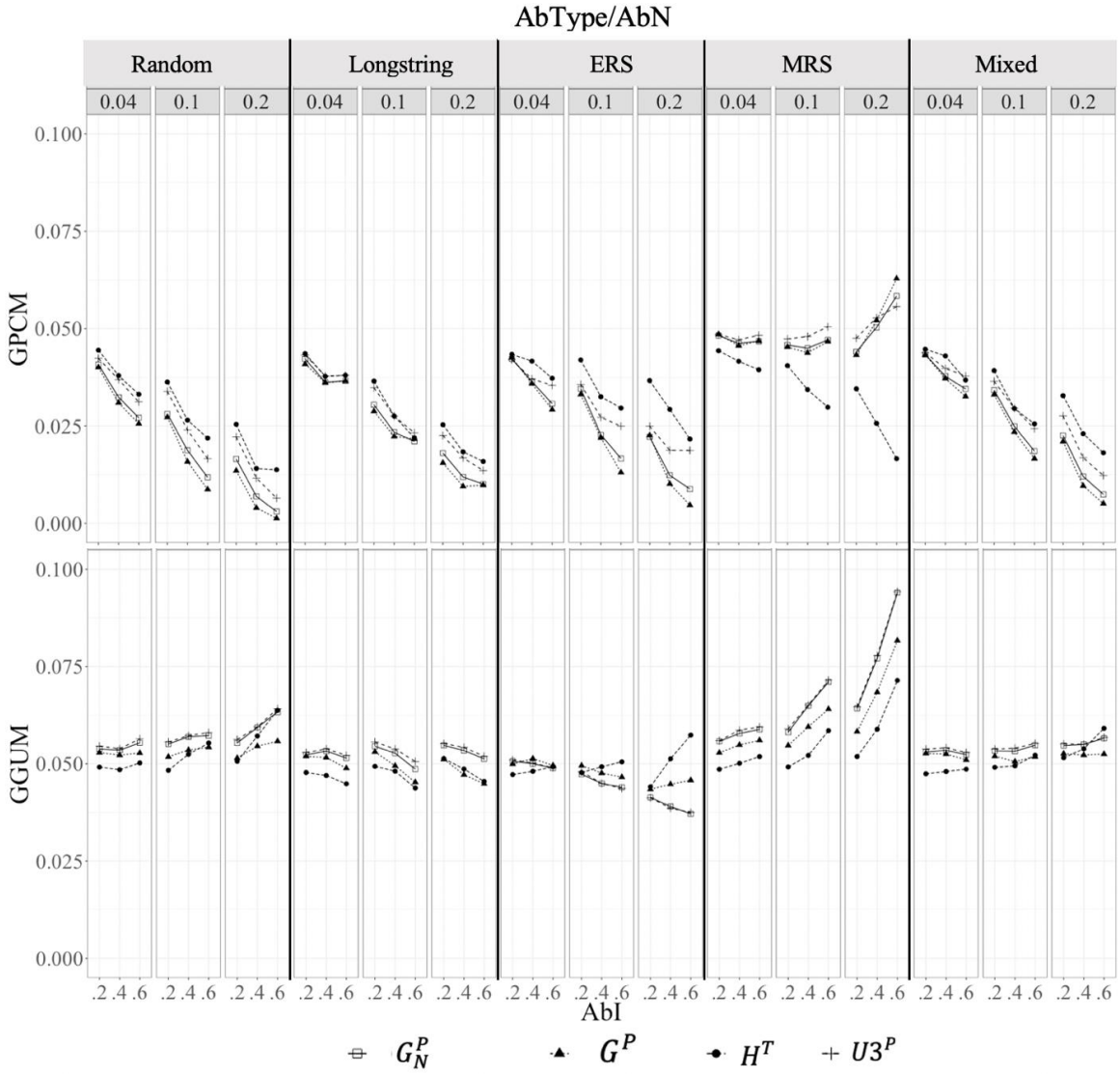
First, results for clean data without aberrant response behavior (i.e., *AbN* = *AbI* = 0) were analyzed to ensure detection methods were working as expected in both unfolding (GGUM) and dominance (GPCM) datasets. Using the clean datasets, each person-fit statistic incorrectly

identified approximately 4 to 6% of the simulees as aberrant. This matches the expectation based on the nominal Type I error rate and corresponds with the cutoff criteria determined by the 5% probability of a Type I error. Next, datasets that included aberrant responses were examined to determine the impact of aberrant responding on the false detection of non-aberrant simulees. The person-fit statistics were computed using the entire dataset (aberrant vectors included) with the cutoff criteria obtained from the means of the replicated model fitting data. To compute the Type I error for each dataset, the proportion of non-aberrant vectors flagged by the person-fit criteria was recorded. Figure 3 illustrates Type I error rates for each type of aberrant response, *AbN*, and *AbI* condition using the four person-fit statistics for the GPCM and GGUM datasets. The 20- and 40-item datasets had very similar Type I error rates and trends. Thus, only the 20-item datasets are shown in Figure 15. For the GPCM datasets (top graph in Figure 15), increasing the *AbI* condition tended to result in more conservative Type I error rates (i.e., type I error rates were less than 0.02 in the highest *AbI* condition of .6 except when $U3^P$, G_N^P, and G^P were used on the MRS datasets). In the *AbN* =.20 conditions, where $U3^P$, G_N^P, and G^P were used on the MRS datasets, Type I error increased with larger proportions of *AbI*.

Similar to results when increasing *AbI*, increased *AbN* (proportion of simulees with aberrant responses) in the GPCM datasets generally resulted in more conservative Type I error rates, except when $U3^P$, G_N^P, and G^P were used on the MRS datasets. In the GGUM datasets, increasing *AbN* had a less pronounced impact on Type I error. GGUM datasets with MRS were most impacted by increasing *AbN*, resulting in increased Type I error rates. Using GGUM data, type I error rates for many conditions were near the nominal .05, except when MRS was present and type I error rates increased to .094. Overall, type I error rates for the person-fit statistics applied to GPCM datasets were more conservative than the rates for GGUM data. Type I error

rates for the GGUM data tended to be slightly above .05 for non-aberrant simulees in many aberrant conditions, with the largest inflation occurring when MRS is present.

Figure 15. Average Type I Error for GPCM and GGUM Datasets (20 items)



Note. ERS = Extreme Response Style. MRS = Midpoint Response Style. *AbI* = Proportion of items within response vector designated as aberrant. *AbN* = Proportion of simulees designated to have aberrant response vectors.

Power Under the Dominance Context

Figure 16 shows a comparison of the power (true positive detection rates) between the four nonparametric person-fit statistics in the dominance GPCM datasets with respect to the

number of items and the type of aberrant responding (ERS, MRS, longstrings, random responders and mixed aberrant responders). For each of these five aberrant response types, results are further broken down by *AbI* and *AbN* conditions. Over all four person-fit statistics, random responders were the easiest to detect within the GPCM datasets, with detection rates as high as 96% using G^P. ERS was also well detected by $U3^P$, G_N^P, and G^P, especially in conditions where *AbI* = 0.60 and *AbN* = 0.04 (Power = 0.96, 0.97, and 0.89, respectively). Generally, when 40% or 60% of the aberrant participants' responses were random or exhibited ERS, $U3^P$, G_N^P, and G^P were more likely to identify the participant as aberrant, whereas when only 20% of the responses were aberrant, detection rates were lower using these methods. Conversely, H^T was ineffective in detecting ERS in GPCM data under all conditions.

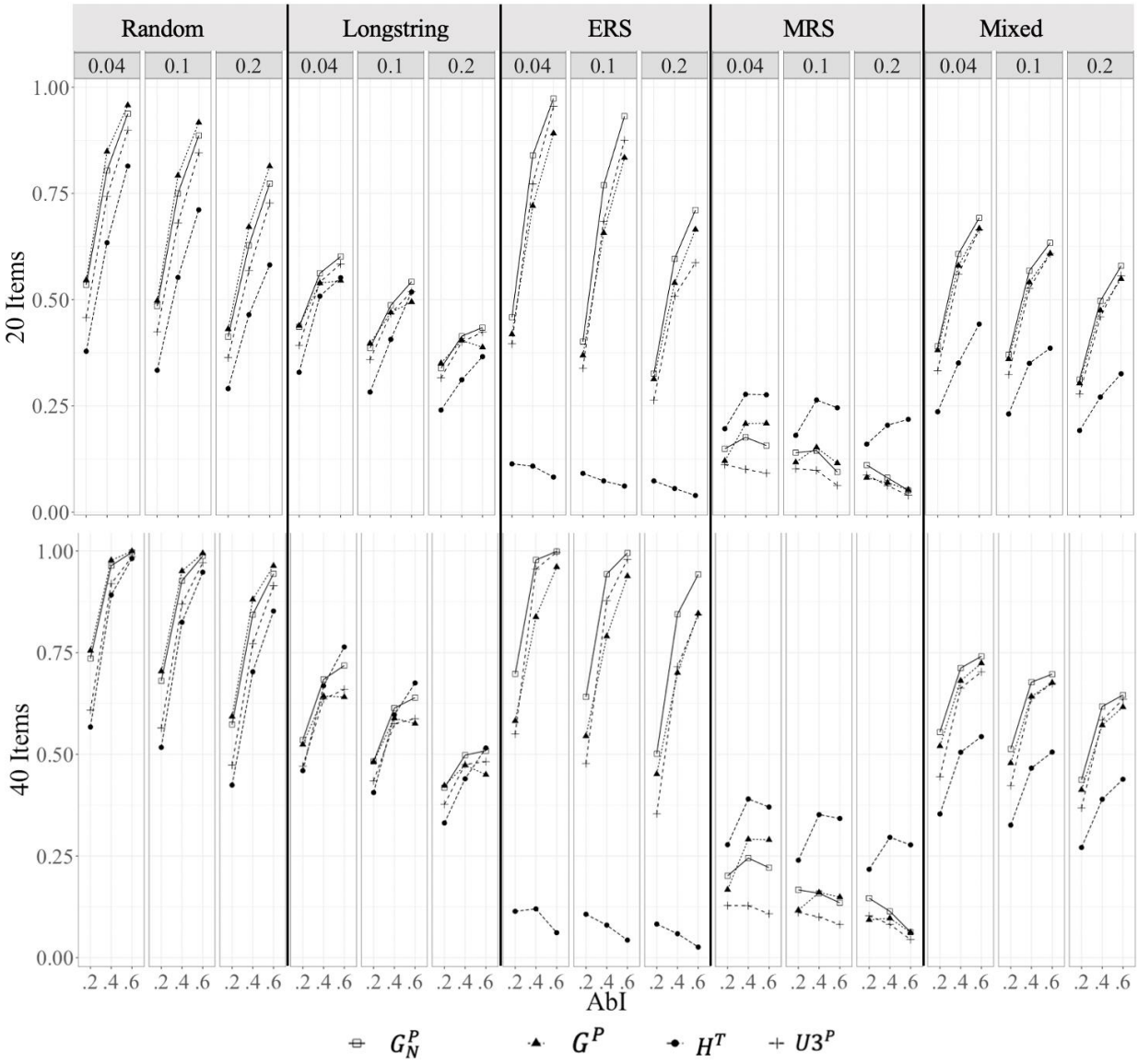
Patterns reveal that for three of the four person-fit statistics ($U3^P$, G_N^P, and G^P), aberrant responding due to MRS is the most difficult to detect in GPCM data (Figure 16). The fourth person fit statistic (H^T) had the most difficulty detecting ERS, with MRS being the second-most difficult aberrant response type to detect. Although H^T was the best in detecting MRS, it still had relatively lower power (the highest condition was .39 for 40 items). Longstrings were the second most difficult to detect in GPCM data overall, with G_N^P having the highest rate in the 20-item datasets with an average power = .60 (AbI = .6 and AbN = .04 condition) and H^T having the highest rate in the 40-item datasets with an average power = .76 (AbI = .6 and AbN = .04 condition). Datasets with a mixed composition of aberrant responders resulted in similar detection rates to the datasets with longstrings.

As the proportion of aberrant examinees in a sample (*AbN*) increased, the detection rates generally decreased. Within each of the two test length conditions (20 and 40 items), the pattern of detection rates shown in Figure 4 illustrate how detection rates tended to slightly increase with

increasing test lengths. In the shorter tests (20 items), G_N^P seemed to outperform the other statistics when responders exhibit ERS, longstringing, or when mixed aberrant responders are present. G^P was slightly better at identifying random responders. H^T was favored when responders exhibit MRS, however none of the four person-fit statistics were effective at identifying MRS in GPCM data. For tests with 40 items, G_N^P had the highest power for conditions with ERS and mixed aberrant responders, H^T had the highest power for conditions with MRS and longstring responders, and G^P had the highest power for detecting random responders.

For random responding, longstrings, and mixed aberrant responses, as the proportion of aberrant responses within aberrant vectors (*AbI*) increased, the detection rates generally increased as well. The exception to this trend occurred when the power of G^P increased from *AbI* = .2 to .4 and then decreased from *AbI* = .4 to .6. For ERS, three of the four statistics ($U3^P$, G_N^P, and G^P) showed a clear trend of increasing power with increasing *AbI*. However, when H^T was applied to the GPCM data with ERS, not only were detection rates lower, but they also decreased with increasing *AbI*. Increasing the levels of *AbI* with MRS had differential impacts on the detection rates for each person-fit statistic.

Figure 16. Power of Person Fit Statistics in GPCM Datasets by Type of Aberrant Response and Proportions of Aberrant Responding



Note. ERS = Extreme Response Style. MRS = Midpoint Response Style. *AbI* = Proportion of items within response vector designated as aberrant. *AbN* = Proportion of simulees designated to have aberrant response vectors.

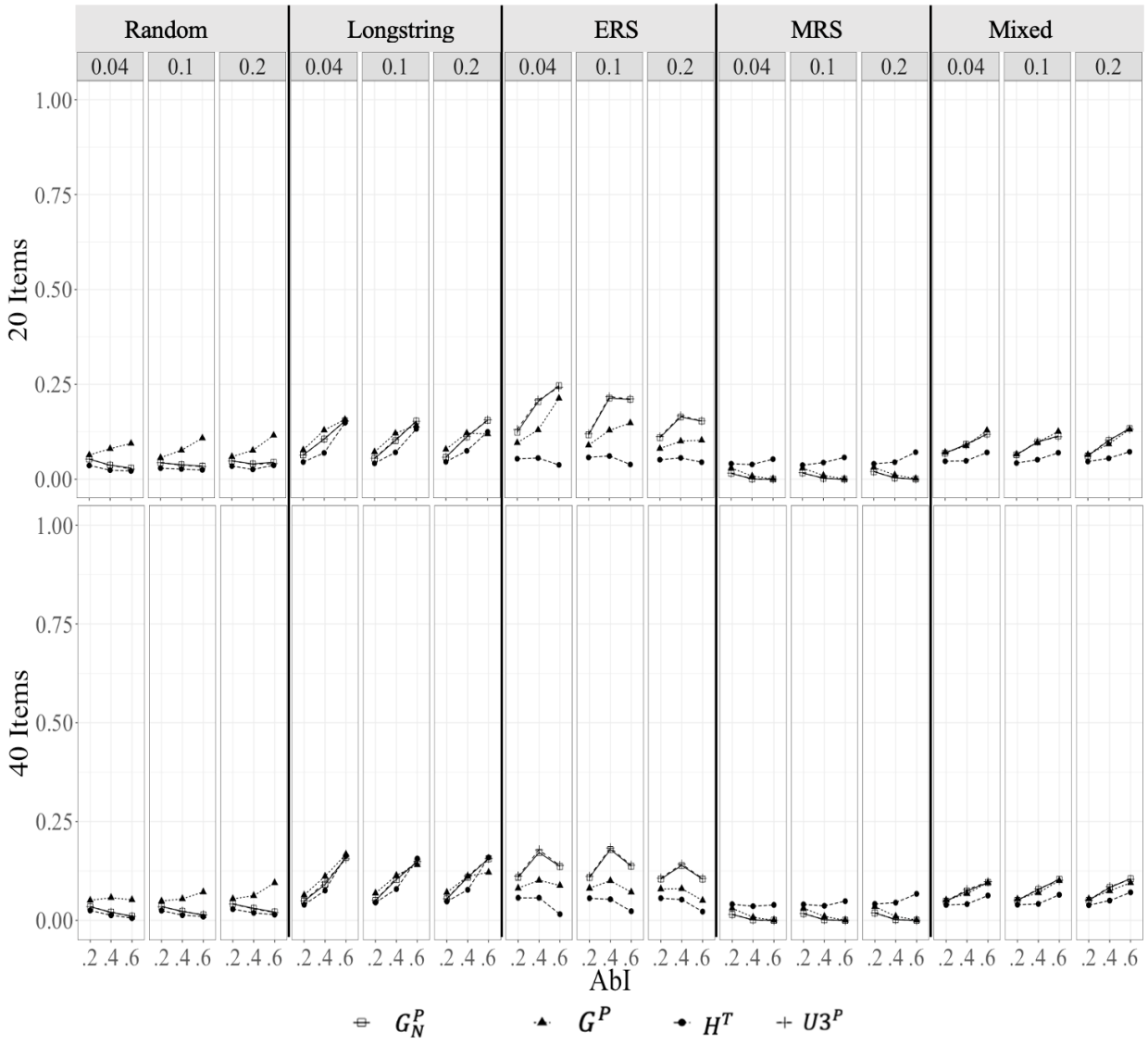
Power Under Unfolding Context

Figure 17 shows a comparison of the power (true positive detection rates) between the four nonparametric person-fit statistics in the ideal point GGUM datasets. The most notable difference for the person-fit statistics under the unfolding context (GGUM data) compared to the dominance context (GPCM data) was the universal drop in power levels. Although all power rates are below necessary levels, trends are described as they may contribute to future research directions. For the 20-item datasets, ERS had slightly higher detection rates than the other types of aberrant responding. However, the highest power on average for ERS in the 20-item tests was .25 with the G_N^P statistic under the condition where few aberrant responders were present in the sample (*AbI* = 0.04) and higher proportions of aberrant items were present within the aberrant response vectors (*AbN* = 0.60). $U3^P$ had similar power under this condition (0.24), followed by G^P (0.21). H^T had the lowest power in detecting ERS, reaching only 0.06 in the condition where *AbI* = 0.10 and *AbN* = 0.04. For the 40-item datasets, longstrings and ERS had slightly higher detection rates than random responding and MRS, however these true positive rates were not substantially higher than type I error rates for the data. As the proportion of longstring values within vectors (AbI) increased, detection increased for all four person-fit statistics. However, as AbI increased from .4 to .6 for ERS, detection began to decrease. Though less apparent than in the GPCM datasets, MRS seemed to be the most difficult type of aberrant response to detect in GGUM data using these four person-fit statistics, with random responding having similar levels. Similar to the GPCM datasets, H^T was the best at detecting MRS, though detection rates were similar to its detection of ERS which was close to type I error rates (< 0.08).

The trend of decreasing detection rates for increasing the proportion of aberrant responders in a sample (*AbN*), as seen in the GPCM datasets, was only observed in the ERS

condition and was not as pronounced in the GGUM datasets. The comparison of detection ability between the four nonparametric person-fit statistics with respect to test length (20 items or 40 items) among the GGUM datasets revealed an opposite trend from the GPCM datasets, where the longer tests resulted in slightly lower detection rates for random responding, ERS, and mixed aberrant conditions. For longstrings and MRS, test length seemed to have a negligible effect. In both the shorter and longer tests, $U3^P$ and G_N^P seemed to outperform the other statistics when responders exhibit ERS, while G^P was slightly more effective when random responders were present. $U3^P$, G_N^P, and G^P all performed similarly when aberrant responding types were mixed, and H^T was favored when responders exhibit MRS. Overall, the detection rates using the four nonparametric person-fit statistics in this study were not effective at identifying ERS, MRS, longstrings, nor random responders in data that fit an unfolding model using GGUM.

Figure 17. Power of Person Fit Statistics in GGUM Datasets by Type of Aberrant Response and Proportions of Aberrant Responding



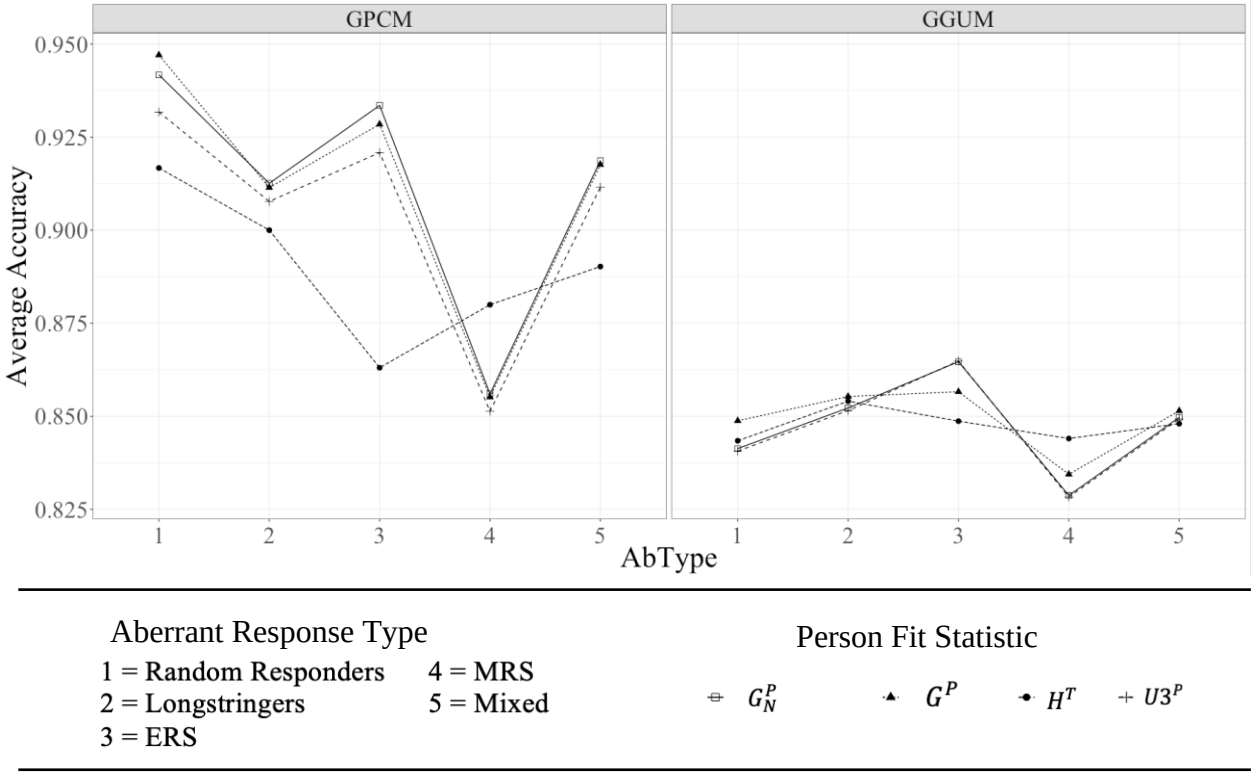
Note. ERS = Extreme Response Style. MRS = Midpoint Response Style. *AbI* = Proportion of items within response vector designated as aberrant. *AbN* = Proportion of simulees designated to have aberrant response vectors.

Accuracy

To provide a different perspective of classification evaluation, accuracy rates were computed for each condition and replication. As mentioned previously, accuracy may be skewed when the proportions of aberrant and non-aberrant response vectors are not the same, giving more weight to the group that makes up the larger proportion of the sample. In this study, the larger group was always the non-aberrant simulees. In the "cleanest" datasets, the non-aberrant group made up 96% of the sample. Even in the most extreme aberrant conditions, the non-aberrant group made up 80% of the sample. Thus, high accuracy rates were obtained by correctly identifying non-aberrant response vectors, even if the aberrant vectors were infrequently classified correctly. Due to this issue of comparing across different levels of *AbI* and *AbN*, accuracy was aggregated using the mean across all *AbI* and *AbN* levels to focus comparisons on the accuracy of person-fit statistics as applied to different aberrant behavior. Figure 18 illustrates this comparison across both types of data (GPCM and GGUM). The test length condition (20 or 40 items) had a very small influence on the accuracy rates, where cases with 40 items had either the same or slightly higher averages than the conditions with 20 items. Thus, results are shown for conditions with 20 items only. Overall averaged accuracy rates were lower for the GGUM datasets which ranged from .83 to .86 in comparison to the GPCM datasets which ranged from .85 to .95. For the GPCM datasets, G^P and G_N^P, had the highest accuracy in detecting random responding, longstrings, ERS, and mixed aberrant conditions, while H^T had the highest accuracy for detecting MRS. H^T had lower accuracy than the other three statistics for all types of aberrant responding except MRS. For the GGUM datasets, G^P had slightly higher accuracy for random responding and G_N^P had slightly higher accuracy for ERS. Similar to the GPCM datasets, H^T had the highest accuracy for detecting MRS. All person fit statistics had very similar accuracy rates

for both longstring and mixed aberrant responding conditions. It is important to reiterate that the accuracy rates for GGUM are much lower than what would be desired when the proportions of non-aberrant responders are .80 to .96. Further, none of the person-fit accuracy rates for detecting MRS in GPCM data are sufficient.

Figure 18. Average Accuracy Rates for 20-item Datasets by AbType and Person-Fit Statistic



Note. ERS = Extreme Response Style. MRS = Midpoint Response Style.

Discussion

This study focused on the application of four nonparametric person-fit statistics that have been shown to work well under certain conditions with dominance data (Emons, 2008; Karabatsos, 2003; Tendeiro & Meijer, 2014), and tested whether similar trends exist when applying them in an unfolding context. Under the dominance conditions, several trends found in

the study were consistent with prior research. For example, as the proportion of aberrant responses within participants' response strings increases, the detection rates generally increase, indicating the degree of aberrant responding within response vectors makes a difference in misfit detection (Karabatsos, 2003; Rudner, 1983). Additionally, the study supports the finding that increasing the test length (from 20 items to 40 items) increases detection rates under the dominance context (Karabatsos, 2003; Meijer et al., 1994, 1996; Rudner, 1983; Tendeiro & Meijer, 2014). Results also coincide with Emons (2008) where the normed nonparametric person-fit statistics G_N^P and $U3^P$ tend to outperform the non-normed G^P statistic in detecting responses that reflect a tendency to choose extreme response categories. However, our results are contrary to prior research (Karabatsos, 2003) that indicates H^T may be more effective than the dichotomous versions of G^P, G_N^P, and $U3^P$ in detecting random responding, as our study found H^T to be the least effective of these four under our data conditions. Sinharay (2017) also displayed conditions of random responding where $U3$ is as effective as H^T for random responding identification.

This study also adds to the research on extreme and midpoint response styles with certain person-fit statistics. The results demonstrate how MRS may be very difficult to detect in GPCM data when using G^P, G_N^P, and $U3^P$, and although H^T has the highest power for detecting this type of aberrant responding, it is still ineffective. Further, under the study conditions, H^T has a very difficult time detecting ERS. This is different than findings from research that has found H^T to perform relatively well in detecting spuriously high and spuriously low responses, however ERS is a form of extreme responding that is different from spuriously high or low data as ERS occurs in a bidirectional manner for participants (i.e., responses are more likely to be both strong

agreement and strong disagreement for a single person) rather than a unidirectional response tendency (spuriously high, spuriously low) as simulated in Emons (2009).

Results reveal a clear disparity in the performance of the four nonparametric person-fit statistics for dominance and ideal point response data. While each of the four statistics has relative success for three of the four aberrant conditions using dominance response data (as modeled by the GPCM), a serious concern is highlighted when using these nonparametric person-fit statistics with ideal point response data (as modeled by the GGUM). First, we are unaware of published research on the detection of random responding or longstrings in an ideal point response context. Our findings indicate that none of the four nonparametric person-fit statistics that we investigated effectively detect either of these aberrant conditions in unfolding model data. Tendeiro (2017) had relative success detecting MRS under an unfolding model context using the parametric person-fit statistics $l_{z(p)}$ and $l^*_{z(p)}$, however these statistics were unable to detect ERS in the conditions of his study. His study did not include the use of nonparametric person-fit statistics. In our study, the results for detecting ERS with the normed nonparametric person-fit statistic G_N^P and $U3^P$ had similar rates to the $l_{z(p)}$ and $l^*_{z(p)}$ under similar conditions in Tendeiro (2017). However, none of our nonparametric person-fit statistics were able to detect MRS as well as $l_{z(p)}$ and $l^*_{z(p)}$.

Low power for detecting aberrant responding under an unfolding model context is possibly because the item ordering criteria underlying many person-fit statistics were not met for several of the GGUM datasets (Figure 1). Though this was anticipated by the researchers, the extent of this effect on detection rates and type I error was uncertain. In several conditions for the current study, power for the unfolding (GGUM) data is similar to expected type I error rates.

Accuracy rates provide a perspective that focuses on correctly classified respondents, which could be useful to compare person-fit statistics as applied to different aberrant behavior. Results illustrate how the person-fit statistics have relatively high accuracy rates for three types of aberrant data detection in GPCM data, but differ by aberrant response type(s) present in the sample. The lowest accuracy rates were found when the person-fit statistics were applied to data with MRS, a finding consistent across both GPCM and GGUM datasets. Under the dominance context (GPCM datasets), the G^P statistic applied to data with random responding has the highest accuracy overall, while the G_N^P statistic was relatively accurate in detecting ERS. The four person-fit statistics were less effective in identifying longstrings in GPCM data and were ineffective in identifying MRS. The accuracy rates for all four person-fit statistics applied to the five aberrant data condition in the unfolding context (GGUM datasets) were insufficient.

Limitations

Several limitations of the study are worth discussing, as well as additional areas of research that can be investigated further. First, this study did not include the manipulation of item parameters as a simulation condition. Of particular interest may be the item discrimination parameter since previous research has shown this parameter to have potentially large effects on the detection power for nonparametric person-fit statistics (St-Onge et al., 2011; Tendeiro & Meijer, 2014). The current study simulated response data using discrimination parameters randomly sampled from a uniform distribution [0.5, 2.0]. Extending this range to 2.5 for example, may have increased power under certain conditions.

Second, related to manipulating item parameters, the degree of intersecting item response functions (IRFs) could be investigated. The practicality of assuming non-intersecting IRFs is that it aligns with the use of the sum score for ordering persons (Emons, 2008; Sijtsma & Molenaar,

2002). In replications where the item discrimination parameters were relatively low, the intersection of IRFs is expected. This characteristic of the study could have influenced Type I error rates and power. Previous research suggests that if the overall $H^T \geq .30$ and the percentage of negative H^T values < 10, it may be assumed that IRFs do not intersect. However if one or more of these conditions is violated, it may be assumed that for a substantial number of persons, item ordering is different (Sijtsma & Meijer, 1992). This could be one contributing factor to the low power in the GGUM datasets where many overall H^T were less than .30. However, although increasing item discrimination parameters within a dominance model may be expected to increase power in detecting certain aberrant responses due to the reduction in intersecting IRFs, this would not necessarily be expected for an unfolding model. Future research may benefit from manipulating the intersection/non-intersection of IRFs by including different IRT models to generate data that satisfy the non-intersecting IRFs assumption.

Third, it is likely that the way the aberrant responding was simulated had method effects on the results. For example, the way ERS and MRS were simulated differs from previous studies such as Tendeiro (2017) and Emons (2008). Tendeiro (2017) simulated the response styles such that middle responses were replaced with the most extreme responses for ERS (or vice versa for MRS). Thus, a score of 3 in this study (on a scale of 0 to 5) would have been replaced by a 5, however scores of 4 would not have been replaced with a 5. Including only 2-point differences and not 1-point differences could have increased the power of aberrant data detection. Other ways of simulating extreme response data such as the manipulation of threshold structures (Emons, 2008; Johnson, 2004; Rossi et al., 2001) could also impact the results. Future research could investigate the differential method effects of different procedures for modeling the extreme and midpoint response styles.

Conclusions

This study examined nonparametric person-fit statistics that have been found to perform just as well, if not better than parametric person-fit statistics under various conditions when applied to data using dominance response models (Emons, 2008; Karabatsos, 2003; Sinharay, 2017; Tendeiro & Meijer, 2014). Specifically, their application to data that more appropriately fits an unfolding model was investigated. Results provided insight on how these nonparametric person-fit statistics perform under various conditions and types of aberrant responding when applied to data that reflect an unfolding response process in comparison to a dominance response process. Results indicate that the nonparametric person fit statistics G^P, $G^P_{N,}$ and $U3^P$ have strong power in detecting extreme response style and random responding with data that fit a GPCM model, and relatively low power in detecting longstrings. Although H^T was the most effective at identifying midpoint response style with GPCM data, the detection levels were very low. Other types of person-fit statistics or procedures (e.g., latent class confirmatory factor analysis models [Moors, 2008], multidimensional nominal response models [Johnson & Bolt, 2010], IRTrees [De Boeck & Partchev, 2012]) may yield more success in identifying midpoint response style in dominance data.

None of the four nonparametric person-fit statistics were effective in detecting aberrant responding with GGUM data. Tendeiro (2017) studied two parametric person fit statistics ($l_{z(p)}$ and $l^*_{z(p)}$) under an unfolding model context (GGUM) where the detection rates for extreme response styles were low however detection of midpoint response style patterns using the $l^*_{z(p)}$ person-fit statistic were promising in many conditions. As the four nonparametric person fit statistics in this study had low power in the unfolding conditions, future research should extend the current study to include other types of nonparametric and parametric person fit statistics for

varying types of aberrant responding with ideal point models. Specifically, we are unaware of person-fit statistics that have been identified as effectively identifying random responding, longstrings, or extreme responding in unfolding models. The development of nonparametric or parametric person-fit statistics designed specifically for unfolding data may be needed. This work might build on research such as Mair, Borg, and Rusch (2016) who apply goodness-of-fit assessments to unfolding data, or research that implements nonparametric unfolding models (e.g., MUDFOLD) to create new person-fit statistics. However, currently, researchers should be advised against applying the popular nonparametric person-fit statistics in this study to unfolding data. Many questions still remain regarding how person-fit statistics perform assuming an underlying ideal point response process. Indeed, we echo prior researchers' recommendations (e.g., Drasgow et al., 2010; Ferrando, 2007; Polak et al., 2012) that further research is warranted to expand our knowledge about the use of these statistics within the growing field of unfolding models.

CHAPTER 6

STUDY 3

Impacts of Misspecification of Underlying Response Processes on the Performance of Nonparametric and Parametric Person-Fit Statistics

ABSTRACT

Person-fit analyses are widely used to detect aberrant responding that can impact model fit and analytical results. However, little attention has been placed on the performance of person-fit statistics in the context of different underlying response processes (e.g., dominance versus ideal point). This study examines the Type I error and power rates of two popular parametric person-fit statistics ($l_{z(p)}$ and $l^*_{z(p)}$) under various aberrant conditions for data generated according to the GGUM and GPCM, reflecting ideal point and dominance response processes respectively. Results are benchmarked against two nonparametric person-fit statistics (G_N^P and G^P). Results indicate that no person-fit statistic was robust against model misspecification when GPCM was fit to GGUM data, as Type I error was severely inflated. Conversely, results were comparable for GPCM data, regardless of fitting the GPCM or GGUM to the data. When the correct model was specified, parametric person-fit statistics were more effective than nonparametric statistics in detecting random responding, longstringing, and midpoint response style in GGUM data, however not extreme response style. Person-fit results varied across type of aberrant responding for GPCM data. As ideal point methods are more appropriate for many noncognitive assessments and are becoming more widely used, results from the study provide practitioners details on the performance of these person-fit statistics under various conditions for data with different underlying response processes.

Impacts of Misspecification of Underlying Response Processes on the Performance of Nonparametric and Parametric Person-Fit Statistics

Research concerning the detection of aberrant patterns in item scores can help researchers learn about individual responding behavior. For example, person-fit statistics have been used to identify examinees who tend to choose extreme or middle response options, lack motivation, or exhibit cheating behavior (e.g., Cizek & Wollack, 2016; Conijn et al., 2014; Emons, 2008; Karabatsos, 2003; Tendeiro, 2017). Aberrant responding can affect the psychometric properties of a scale, which ultimately may lead to negatively impacting the quality of the measure and decisions made using the measure (DeSimone et al., 2018; McGrath et al., 2010). Though numerous studies have investigated the performance of person-fit statistics, the current study focuses on their use under different assumed underlying response processes (dominance and ideal point) and the impacts of model misspecification.

Aberrant Responding

Aberrant responses result from numerous possible behaviors and characteristics of the respondent. For example, misfitting item scores could result from respondents exhibiting Extreme Response Style (ERS), which refers to people tending to choose the upper or lower extreme categories, regardless of the item content (Bachman & O'Malley, 1984; Baumgartner & Steenkamp, 2001; Chen et al., 1995; Greenleaf, 1992; Hui & Triandis, 1985; Marin et al., 1992; Weijters et al., 2010). People who tend to choose the middle response option regardless of item content have been classified as using a Middle Response Style (MRS; Baumgartner & Steenkamp, 2001; Chen et al., 1995; Stening & Everett, 1984), or mid-lining. The presence of ERS and MRS have been found to bias parameter estimates in data that exhibit a dominance (Jin & Wang, 2014) or an ideal point response process (C.-W. Liu & Wang, 2019). Further,

respondents may also lack the cognitive effort to provide meaningful responses. This can manifest itself when respondents give random responses or invariant responses (longstrings) to a set of items regardless of the item content. The literature has revealed many potentially adverse effects of aberrant data and thus its detection using various methods (e.g., person-fit statistics) has gained substantial attention in several fields (Rupp, 2013).

Person-Fit Statistics

"Person-fit," also referred to as "appropriateness measurement," refers to the degree to which a person's item response pattern departs from what is expected based on an item response theory (IRT) model or the response patterns of other persons in the group (Drasgow et al., 1985; Meijer et al., 1994). In general, research applications have focused on two classes of person-fit statistics: parametric and nonparametric. Parametric person-fit statistics generally measure the disparity between the observed data and the estimated response predictions resulting from an IRT model's parameter estimates. Conversely, nonparametric person-fit statistics do not rely on IRT-based estimates, but rather are computed using the observed response data to the items in a dataset (Karabatsos, 2003). The focus of this study is on the performance of two parametric person-fit statistics ($l_{z(p)}$ and $l^*_{z(p)}$) that are widely used among researchers (Armstrong et al., 2007; Conijn et al., 2014; M. Hong et al., 2021; Hong et al., 2020; Magis et al., 2012; Meijer & Tendeiro, 2012; Nering & Meijer, 1998; Avşar, 2021; Seo & Weiss, 2013; Sinharay, 2016b, 2017; St-Onge et al., 2011; Tendeiro, 2017; Torre & Deng, 2008; Xia & Zheng, 2018). Additionally, two non-parametric person-fit statistics ($G_N^P,$ and G^P) are included for comparison purposes that have been shown to perform well in comparison with other person-fit statistics in a variety of aberrant conditions (Emons, 2008; Karabatsos, 2003).

Parametric Person-fit Statistics ($l_{z(p)}$ and $l^*_{z(p)}$). The $l_{z(p)}$ and $l^*_{z(p)}$ statistics stem from possibly the most well-known parametric person-fit statistic, the likelihood statistic, *l* (Levine & Rubin, 1979). The l statistic measures the log-likelihood fit of a response to an item with the prediction based on an IRT model (Karabatsos, 2003). The binomial log-likelihood statistic is computed as follows:

$$l = \sum_{j=1}^{J}[\, X_{nj}(\ln P_j(\theta)) + \left(1 - X_{nj}\right)\left(ln\; Q_j(\theta)\right)], \tag{46}$$

where $P_j(\theta)$ represents the probability of a correct response on item *j* given the person's estimated ability or trait level (θ), and $Q_j(\theta)$ represents the probability of an incorrect response on item *j* $[Q_j(\theta) = 1 - P_j(\theta)]$. The limitation of the likelihood-statistic is that it is not standardized and the distribution under a fitting IRT model is unknown. Drasgow et al. (1985) proposed a standardized normal version of the likelihood statistic, l_z, using the mean and variance in the following way:

$$l_z = \frac{l - E(l)}{\sqrt{var(l)}}, \tag{47}$$

$$E(l) = \sum_{j=1}^{J}\{P_j(\theta)\ln(P_j(\theta)) + Q_j(\theta)\ln(Q_j(\theta)), \text{ and} \tag{48}$$

$$var(l) = \sum_{j=1}^{J} P_j(\theta)Q_j(\theta)\left(log\,\frac{P_j(\theta)}{Q_j(\theta)}\right)^2. \tag{49}$$

For the polytomous case, where probabilities of a correct response reflect passing the k^{th} threshold from one response category to the next, and *C* is the total number of response categories minus 1, the mean and variance for Equation 2 are defined as follows (Sinharay, 2016):

$$E(l) = \sum_{j=1}^{J}\sum_{k=1}^{C}[\,[\,P_{jk}(\theta)\big(\ln P_{jk}(\theta)\big)]], \text{ and} \tag{50}$$

$$var(l) = \sum_{j=1}^{J} \sum_{k_1=1}^{C_j} \sum_{k_2=1}^{C_j} P_{jk_1}(\theta) P_{jk_2}(\theta) \ln\left(P_{jk_1}(\theta)\right) \ln\left(\frac{P_{jk_1}(\theta)}{P_{jk_2}(\theta)}\right). \quad (51)$$

This polytomous version is known as $l_{z(p)}$, and has become a very popular choice for person-fit analyses. Still, it is only when true theta values are used, that this statistic can be assumed to have an asymptotically standard normal distribution (Molenaar & Hoijtink, 1990). In practice, it is unrealistic to assume true theta values are available. Consequently, a modified version of l_z, l_z^* , was proposed by Snijders (2001) that addresses this concern by accounting for the sampling variability of the estimated theta parameters. A thorough and helpful explanation of the computational formulas involved in calculating l_z^* can be found in Magis et al. (2012). Sinharay (2016) further extended this corrected version for polytomous cases, $l_{z(p)}^*$.

The majority of published research concerning the performance of l_z and l_z^* for dichotomous and polytomous data has been conducted using dominance IRT models (Armstrong et al., 2007; Conijn et al., 2014; de la Torre & Deng, 2008; Emons, 2008, 2009; Karabatsos, 2003; Turner, 2018). Emons (2008) used $l_{z(p)}$ as a benchmark against several nonparametric person-fit statistics and found that for careless responding (as simulated by drawing random numbers from a uniform distribution), $l_{z(p)}$had the highest detection rates. Extreme responding was more difficult to detect than careless responding by all person-fit statistics in the study, particularly when the tendency to choose extreme options is exhibited on less than half of the items. Normed nonparametric person-fit statistics ($G_N^P, U3^P$) used in their study had higher detection rates than parametric statistics for most of the conditions under extreme responding. Other studies (Niessen et al., 2016; Turner, 2018) have also found the l_z statistic to have comparatively higher detection rates for random and careless responding under a dominance framework.

In the single study investigating the performance of $l_{z(p)}$ and $l^*_{z(p)}$ when data have an ideal point response (or unfolding) framework, Tendeiro (2017) found the parametric person-fit statistics to have relatively low detection rates for extreme response style (mean across conditions = .17; first and third quartiles = .06 and .22 respectively). Overall, the parametric person-fit statistics were conservative (low type I error rates), but midpoint response style had relatively higher detection rates (mean = .45; first and third quartiles = .17 and .72 respectively). Although Tendeiro (2017) found very similar results for the $l_{z(p)}$ and $l^*_{z(p)}$ statistics in an unfolding context, the current study will include both statistics to test whether they perform similarly under additional aberrant conditions, and in comparison to nonparametric statistics, including conditions where assumed models are misspecified.

Nonparametric Person-fit Statistics. In contrast to parametric person-fit statistics, nonparametric person-fit statistics do not rely on the predicted responses based on an IRT model's parameter estimates (Karabatsos, 2003). This study includes two nonparametric person-fit statistics ($G_N^P,$ and G^P) that have been shown to outperform other person-fit statistics in various aberrant conditions under a dominance response process setting (Emons, 2008; Karabatsos, 2003). The G^P statistic (Molenaar, 1991) is based on the Guttman scaling procedure where items are arranged in order according to their locations on the underlying continuum for the latent trait being measured, and endorsement of a particular item suggests endorsement of all items located lower on the latent continuum. The G^P statistic summarizes the number of deviations from a perfect Guttman pattern of responses for a participant. Thus, more Guttman errors result in a larger G^P statistic, which indicates a larger degree of person misfit. The following equation is used to compute the G^p statistic:

$$G^p = \sum_{l<k}^{JC} y_k(1 - y_l), \tag{52}$$

where *J* is the number of items and *C* is the number of response categories minus 1 (number of item steps) for each item. The y_l in Equation 7 represents all elements of vector $\boldsymbol{Y}$ that are prior to y_k. To compare the G^p statistics across respondents with different sum scores (X_+), the following equation can be used to normalize the G^p statistic to have a range of [0, 1] (Emons, 2008):

$$G_N^P = \frac{G^p}{\max\left(G^p | X_+\right)}. \tag{53}$$

Both Guttman person-fit statistics have been shown to detect random responding well, and G_N^P has been shown to outperform G^P in detecting ERS in dominance contexts (Emons, 2008). Study 2 demonstrated the limitations of these nonparametric statistics under an ideal point framework due to their dependencies on invariant item ordering assumptions and Guttman scaling procedures. However, no studies have shown a direct comparison of their performance to the performance of parametric person-fit statistics in an unfolding context. It is especially useful to investigate their side-by-side performance for aberrant responding such as ERS, where detection rates using parametric person-fit statistics have been shown to be modest at best (Tendeiro, 2017).

Dominance Response Process

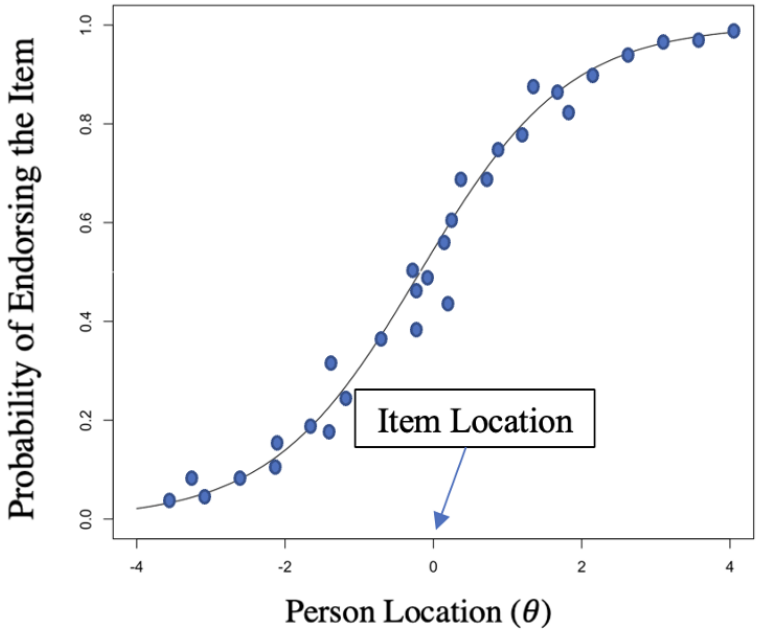
Researchers have long recognized and studied the differences in underlying response processes used by individuals when responding to items. Coombs (1964) used the term "dominance" to describe situations when individuals with higher trait levels dominate, or correctly answer, assessment items. In data that reflect a dominance response process, the probability of endorsing an item increases monotonically with the increase of the latent trait being measured (regardless of the item's location on the continuum). Dominance IRT models reflect this characteristic, which can be illustrated with monotonically increasing item response

functions (IRFs) for dichotomous items (see Figure 19). When items are polytomous, intermediary step functions are modeled. Step functions are defined by modeling the transition, or "stepping" to successively higher score categories. Four popular approaches for defining step functions include adjacent category, continuation ratio, cumulative, and nominal (Penfield, 2014). Within the adjacent category approach is the partial credit model (PCM; Masters, 1982), which models step functions as the probability of success at the adjacent k^{th} step as specified by the Rasch model. The generalized partial credit model (GPCM; Muraki, 1992) also uses the adjacent category approach to defining step functions, but the 2PL model is used and an item-level discrimination parameter is estimated. In the current study, the GPCM was chosen to simulate the dominance data because of its flexibility for estimating different discrimination parameters for items and its comparability with the unfolding model chosen for the study (GGUM; described below). The IRF using the GPCM is given by

$$P(Z_j = z|\theta) = \frac{\exp\{\sum_{k=0}^{z}[a_j(\theta - b_{jk})]\}}{\sum_{r=0}^{C}\{\exp\sum_{k=0}^{r}[a_j(\theta - b_{jk})]\}}, \quad (54)$$

where Z_j represents the observed response (with ability or trait level, θ) to item *j*, and $z = 0, 1, 2, \ldots, C$ with *z* = 0 corresponding to the strongest level of disagreement and *z* = *C* corresponding with the step that reflects the strongest level of agreement. Thus, C is the number of observed response categories minus 1. The discrimination parameter for item *j* is represented by a_j, and b_{jk} is the difficulty parameter or location parameter of the k^{th} step. In the denominator, *r* = 1, 2, . . . , *C* represents the total *C* exponent terms.

Figure 19. Example of IRF Based on Dominant Model

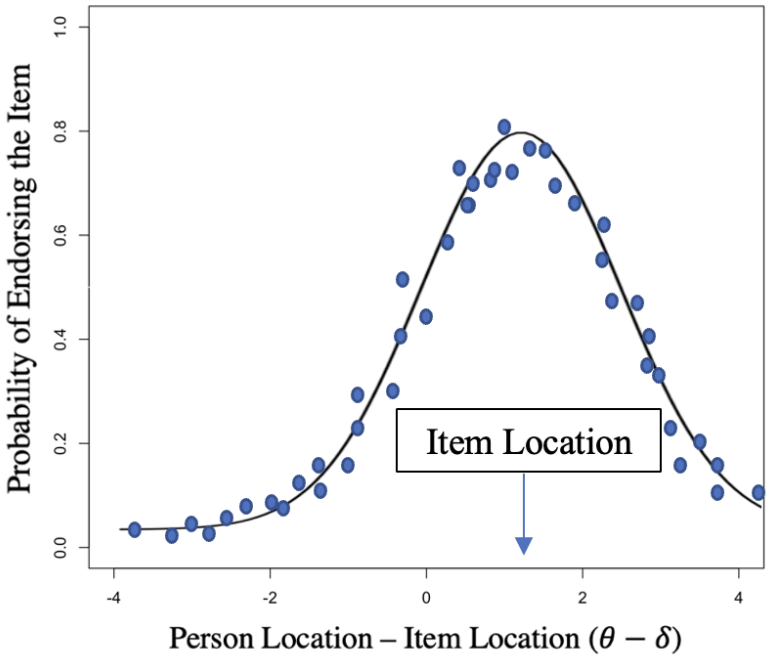


Note. The IRF illustrates a monotonically increasing item response function (IRF).

Ideal Point Response Process

The ideal point response process is based on the notion conceptualized by Thurstone (1928) that assumes that the probability an individual will endorse an item increases to the degree that the item reflects the person's sentiment/trait-level on the construct being measured. Thus, as the difference between the person's and item's location on the underlying latent trait continuum decreases, the probability of endorsing the item increases. This distinction between ideal point and dominance response processes may be best illustrated using a depiction of the IRFs modeled by IRT models with the respective underlying response process assumptions. In contrast to the monotonically increasing IRF modeled by the dominant IRT model in Figure 1, the IRF resulting from an item that reflects an ideal point response process is single-peaked (Figure 20).

Figure 20. Example of IRF Based on Ideal-Point Model



Note. The IRF illustrates a non-monotonic IRF, violating monotonicity assumption for dominance item response theory models.

Differences between items representing ideal point and dominance response approaches can be explained using an example with a neutral item on a mathematics efficacy scale: "I have about average math skills." Someone who perceives themselves to have about average math skills is likely to strongly agree with this item. However, someone who perceives themselves to have below average math skills (their location is below the item location) will likely disagree with the item. Similarly, someone with self-perceived above average math skills (their location is above the item's location) will also likely disagree with the item. The characteristic that two people may disagree with the item for two very different reasons yields an unfolding quality that is problematic for dominance IRT models. However, ideal-point (unfolding) IRT models can be used for data that reflect an ideal point response process because they incorporate subjective response functions that account for disagreeing (or agreeing) with an item due to the individual's

trait level being above or below the item's location on the underlying latent trait continuum. An increasingly popular model used for ideal point response data is the generalized graded unfolding model (GGUM). This investigation used the GGUM proposed by Roberts et al. (2000), where each subjective response follows the GPCM (Equation 7). The formal equation for the GGUM is given by equation 8.

$$P(Z_j = z|\theta) = \frac{\exp\{\alpha_j[z(\theta-\delta_j)-\sum_{k=0}^{z}\tau_{jk}]\}+\exp\{\alpha_j[(M-z)(\theta-\delta_j)-\sum_{k=0}^{z}\tau_{jk}]\}}{\sum_{w=0}^{C}\{\exp\{\alpha_j[w(\theta-\delta_j)-\sum_{k=0}^{w}\tau_{jk}]\}+\exp\{\alpha_j[(M-w)(\theta-\delta_j)-\sum_{k=0}^{w}\tau_{jk}]\}\}}. \quad (55)$$

The additional parameters include *M*, which corresponds to the strongest level of disagreement from *above* the item (*M* is the total number of subjective response categories minus 1), δ_j is the item location parameter, and τ_{jk} is the location of the k^{th} threshold on the latent continuum relative to the location of the j^{th} item. Although a relative increase in complexity is evident, the GGUM package in R (Tendeiro & Castro-Alvarez, 2019) makes it an accessible option for researchers.

Person-Fit Statistic Performance Under Model Misspecification

From a statistical standpoint, the performance of a person-fit statistic depends on several assumptions. One of these, and a particular focus of this study, is the assumption that the correct underlying response process is correctly specified. The performance of person-fit statistics under model misspecification have been studied under dominance frameworks. Meijer and Tendeiro (2012) highlight the importance of investigating model and item fit before assessing person-fit with the l_z and l_z^* statistics for dichotomous data. Without information regarding model and item fit, item response patterns may be flagged by person-fit statistics due model-data misfit or due to actual aberrant responding behavior. In the empirical study, when the better-fitting IRT model (2PL) was fit to the data, the study illustrated how the distribution of the l_z^* was superior and relatively closer to standard normal than the l_z distribution. This is in contrast to the findings by

Magis et al. (2012) where the l_z^* statistic had worse fit for extreme values, which the authors note may have been due to possible misfit of the IRT model used to fit the data (i.e., the 2-parameter logistic model was used instead of the possibly more appropriate 3-parameter logistic model for language assessment data). When Tendeiro and Meijer (2012) intentionally used the wrong IRT model (done by constraining the discrimination parameters), the left tails of the l_z and l_z^* distributions were thicker, though the differences were not as large for the l_z^*. Under the misspecification condition, the l_z distribution appeared to be slightly closer to standard normal probably because of the inflation of the statistic due to the many misfitting scores detected because of model misfit. These results resemble those found in the Magis et al. (2012) paper, supporting the suggested possible explanation for why the l_z^* statistic had a somewhat different distribution than the expected standard normal density (model misfit).

More recently, Hong et al. (2020) used a simulation study along with two empirical examples to reveal potential consequences of model misspecification for common logistic IRT models (i.e., 1PL, 2PL, and 3PL) when using the popular l_z^* and the extended caution index ζ_2^* (Sinharay, 2016a; Tatsuoka, 1984). In the simulation study, data did not include aberrant response vectors. Thus, detected responses by the person-fit statistics reflected type I error rates. Results illustrated that type I error rates were inflated for conditions where models were misspecified compared to when the correct models were fit to the data. When type I error rates were compared conditionally on person parameters (θ), the impact of misspecification was amplified for extreme values of θ, reaching rates of .27 under certain conditions. The study also showed that neither of the person-fit statistics used were robust to model misspecification. The model misfit conditions where the 1PL was fit to either the 2PL or 3PL had the greatest inflated type I error rates. The authors encouraged practitioners to consider specifying relatively complex

IRT models for the purpose of producing person-fit statistics. However further research is warranted to test this hypothesis, especially in the context of unfolding versus dominance models, where the impact of model misspecification on person-fit statistic performance is unstudied.

Purpose

The current research extends these previous studies by including both dominance and unfolding models in a simulation study where conditions can be fully controlled by the researcher to reveal the effects of various factors on the performance of person-fit statistics. In addition to investigating the relative performance of the person-fit statistics under the context of two different underlying response processes, the study will add to the literature regarding the effects of model misspecification as well. Although model fit is routinely part of item analysis procedures, the primary driver for model fit investigation is usually not person-fit analyses. For example, model fit investigations may be conducted for a scale intended to be used for proficiency classification, which may be quite robust to model misspecification (S. E. Hong et al., 2020). Another example includes the retention of items, even though items may show misfit, due to practical concerns such as the need to satisfy test assembly requirements. These examples of model misspecification and the retention of items exhibiting misfit could have severe effects on person-fit analyses and need to be investigated further, especially under an unfolding model context where very little research regarding person-fit analysis has been published. The purpose of this study is to assess the performance of parametric and nonparametric person-fit statistics in identifying varying types of aberrant responders (midpoint and extreme responding, invariant/longstring responding, random responding) in the context of two underlying response processes (dominance versus ideal point) that have either been correctly or incorrectly specified.

It is anticipated that in the cases of poor model fit, the person-fit statistics will not perform as well. One goal is to provide insight on the point at which this hypothesized decline in performance is most apparent. The research questions driving the study include:

1) What are the effects of model (mis)specification of dominance and unfolding models applied to dominance or unfolding data on the performance of parametric and nonparametric person-fit statistics for detecting aberrant data?
    a. How accurately do the selected person-fit statistics identify aberrant responding when a(an):
        i. dominance model is applied to dominance data?
        ii. unfolding model is applied to unfolding data?
        iii. dominance model is applied unfolding data?
        iv. unfolding model is applied to dominance data?
    b. How does the performance of the parametric person-fit statistics $l_{z(p)}$ and $l^*_{z(p)}$ compare to nonparametric person-fit statistics (G^p and G^P_N) under unfolding and dominance model contexts?
        i. Are the trends and magnitudes for detection and type I error rates (e.g., higher detection rates with longer tests) the same under unfolding and dominance frameworks?
    c. What kinds of aberrant behavior are most/least easily detectable via parametric and nonparametric person-fit analyses when using unfolding vs dominant response frameworks, and under what conditions?

Methods

Simulation Factors

A series of simulations was conducted to assess the performance of the person-fit statistics for a fixed sample size of 1,000 simulee responses to 6-point items. The study included 360 completely crossed conditions including 2 data generating mechanisms (GPCM and GGUM) × 2 models used in fitting the data (GPCM and GGUM) × 3 proportions of aberrant responders in the sample (*AbN*: .04, .10, .20) × 3 proportions of aberrant responses within response vectors (*AbI*: 20%, 40%, or 60%) × 5 types of aberrant responding and response style conditions (random responders, longstring, MRS, ERS, mixed) × 2 test lengths (20, 40). A total of 100 replications were generated for each condition. All code for generating and estimating model parameters, model fit statistics, and person-fit statistic values were written in R (R Core Team, 2016). Code is attached in Appendices A through F.

Simulation Procedures

First, data were generated for the "clean" or uncontaminated unfolding data under the GGUM and dominant data under the GPCM. To do this, person parameters were randomly drawn from the standard normal distribution. For the GGUM datasets, the GenData.GGUM function in R was used to generate the person parameters, item parameters, and item scores. The item discrimination parameters (α_j) were randomly sampled from a uniform distribution [0.5, 2.0], and the item location parameters (δ_j) were randomly sampled from the standard normal distribution truncated between -2.0 and 2.0. Using procedures described in Roberts et al. (2002), the locations of the threshold parameters (τ_{jk}), relative to the location of the *j*th item, were recursively generated.

For the GPCM datasets, item discrimination parameters were also randomly sampled from a uniform distribution [0.5, 2.0], and item difficulty parameters were sampled from the standard normal distribution $N(0,1)$. To simulate item category thresholds, d_{jk}, for step k of item j, the sequential cumulative sum was taken for five numbers drawn from a random uniform distribution between .3 and 1. This results in distances of .30 or greater between categories to increase the chance of responses in each category (Chalmers, 2012). Next, each number in the set of sequential cumulative sums was centered around their mean. Additionally, the initial item category threshold, d_{j0}, was set to 0 in order for the model to be identified (Muraki, 1992). The sim_gpcm function in R (PP package; Reif & Steinfeld, 2021) was used to simulate the GPCM response data.

To obtain datasets that included aberrant data, response vectors from the clean datasets were replaced with vectors simulated to be aberrant. The proportion of items (*AbI*) and response vectors (*AbN*) that were replaced depended on the condition. For random responding, the values of randomly sampled numbers from a uniform distribution [0,5] replaced a proportion of items (AbI × the number of items) for each simulee designated to be "aberrant." To simulate longstring responses, a single value randomly sampled from a uniform distribution [0,5] replaced either 20%, 40%, or 60% of consecutive item responses (*AbI*) in an aberrant response vector (DeSimone et al., 2018). To do this, an initial starting position in the response vector was randomly drawn so that the longstring could fulfill the *AbI* condition. For response vectors reflecting MRS, endpoint item scores and item scores adjacent to the endpoints were replaced with the nearest midpoint response (i.e., on the 6-point scale ranging from 0 to 5, items scores of 0 and 1 were replaced with a 2, and scores of 4 and 5 were replaced with a 3). Conversely, for ERS responding, item scores in the middle range were changed to the nearest endpoint responses

(i.e., 1s and 2s were changed to 0; 3s and 4s were changed to 5). An additional "mixed" condition of aberrant responding was included in the study to mimic the realistic testing situation where a mixture of aberrant response types show up in the data. For this condition, an equal proportion of each type of aberrant responding replaced the response vectors for aberrant simulees (random responding, longstrings, ERS and MRS). To clarify, each aberrant simulee was designated with a single type of aberrant responding, and the datasets consisted of a mixture of the four types of aberrant responders for the "mixed" condition.

Model Fit and Parameter Recovery

Once the datasets were created (clean and aberrant), both the GPCM (using the MIRT function in R; Chalmers, 2012) and GGUM (using the GGUM package in R; Tendeiro & Castro-Alvarez, 2019) were used to fit the data under the various conditions. Any time GPCM was fit to GGUM responses, "Likert-scaling techniques" described in Tay et al. (2011) and Tay and Drasgow (2012) were applied by reverse scoring the lower 30% of items on the continuum. Then, any remaining items with negative item-total biserial correlations were also reverse scored. The investigation of model fit, item fit, and parameter recovery was an important precursor to investigating person-fit with the person-fit statistics in the study. As noted in the research, model fit can have an impact on the performance of person-fit statistics (Meijer & Tendeiro, 2012). As Tendeiro (2017) notes, and others have shown (St-Onge et al., 2011), the impact of the operationalization of the aberrant responding entered into the data could have effects on the performance of person-fit statistics. In the cross-fitting conditions (e.g., where the parametric statistics use parameter estimates from the GPCM fit to GGUM data), parameter quality is not expected to be relatively high, even for clean data. In these cases, it is expected for the person-fit statistics to not perform as well. The purpose for this type of condition is to evaluate the

flexibility of the models and investigate the rate at which the person-fit statistic performance declines. One of the advantages to using a simulation for the study was having full control over factors to reveal effects undetectable in realistic empirical studies.

Under each condition, multiple aspects of model fit were examined including dimensionality, information criteria, item-fit, and quality of parameter estimates. Dimensionality was investigated using parallel analysis in R ('paran' package; Dinno, 2018) where Horn's technique was used. This technique adjusts for the sample error-induced inflation while quantitatively and graphically determining the number of factors retained in a Principal Components Analysis (PCA).

Information criterion-based statistics are commonly used in the process of selecting a model. The Akaike Information Criterion (AIC; Akaike, 1974) and Bayes Information Criterion (BIC; Schwarz, 1978) allow for relative model fit and penalize for parameters being added to the model. The BIC penalty is stronger and lower values indicate better fit, as demonstrated in the computations:

$$AIC = 2k - 2\ln(\hat{L}) \tag{56}$$

$$BIC = k\ln(n) - 2\ln(\hat{L}) \tag{57}$$

where k is the number of parameters estimated by the model, $\hat{L}$ is the maximized value of the likelihood function for the model, and n is the sample size. Both the AIC and BIC are used in the current study to demonstrate the relative fit of each model to the data under different conditions.

To further assess model fit, the item-level goodness-of-fit was examined using adjusted χ^2/df ratios (Drasgow et al., 1995) for doublets and triplets of items. This procedure was successfully used by Tay et al. (2011) when comparing relative model fit for dominance and unfolding models. The ordinary chi-square statistic is computed using the observed (O_{jz}) and

expected (E_{jz}) frequencies of item *j* response option *z* based on the estimated item parameters and distribution of abilities.

$$\chi_j^2 = \sum_{z=0}^{C} \frac{\left(O_{jz} - E_{jz}\right)^2}{E_{jz}}, \tag{58}$$

with $E_{jz} = N \int P_{jz}(\theta)\phi(\theta)d\theta$, where $\phi(\theta)$ is the standard normal density function. Equation 11 is written for single items. However, chi-squares for single items have been found to be insensitive to unidimensionality and local independence violations (Drasgow et al., 1995). Thus, Equation 3 is often generalized to apply to pairs of items (doublets) and triples of items (triplets). Additionally, the χ^2 statistic is adjusted to a sample size of 3,000 (Drasgow et al., 1995; Lahuis & Clark, 2009) to make it more generalizable across samples of different sizes. The resulting formula is:

$$\chi^2/df \; = \; \frac{3{,}000(\chi_j^2 - df)}{N} + df \tag{59}$$

where the degrees of freedom (*df*) depend on the number of singlets, doublets or triplets used. A common rule of thumb is that an adjusted χ^2/df larger than 3 be considered indicative of model misfit (Stark et al., 2006; Tay et al., 2011; Tendeiro, 2017).

Additionally, the quality of parameter estimates was compared across conditions to reveal any differences in parameter recovery that may affect the performance of the person-fit statistics. Following the procedures used in Tendeiro (2017), the bias, mean absolute deviation (MAD), and the correlation (COR) between true and estimated parameters were computed using the following equations:

$$MAD = \sum_{t=1}^{T} |\hat{\gamma}_t - \gamma_t^{TRUE}|/T \tag{60}$$

$$BIAS = \sum_{t=1}^{T} (\hat{\gamma}_t - \gamma_t^{TRUE})/T \tag{61}$$

$$COR = cor(\hat{\gamma}_t, \gamma_t^{TRUE}), t = 1, \ldots, T, \tag{62}$$

where $\widehat{\gamma_t}$ is the estimated parameter and γ_t^{TRUE} is the true simulated value for the parameter representing either $\alpha_j, \delta_j, \tau_{jk},$ and θ_n for the GGUM parameters or $\alpha_j,\ b_{jk},$ and θ_n for the GPCM, and *T* is the corresponding total number of that parameter. For example, *T* is equal to the number of items for α_j and δ_j. For τ_{jk}, *T* equals *I* (the number of items) times *C* (the number of observed response categories minus 1), and for n, *T* is equal to the sample size. For each condition, the MAD, bias, and correlations were computed and averaged over all replications. To illustrate the variability in parameter bias, standard deviations for these averages are also reported.

For the cross-fitting conditions (fitting GPCM to GGUM data or the GGUM to the GPCM data), some estimated parameters are not directly comparable. For example, in dominant IRT models the location parameter represents the person parameter associated with a .50 probability of choosing a response. However, in unfolding models, the location parameter represents the person parameter associated with the highest probability of endorsement. The subjective thresholds in GGUM do not exist for GPCM, and the discrimination parameters for non-monotonically increasing items may result in negative values, so averaging over a condition may be misleading. Thus, for conditions where the inappropriate model was applied to the data, parameter quality is discussed for person scoring only.

Computing and Evaluating the Person-fit Statistics

To compute the parametric person-fit statistics used in the study, $l_{z(p)}$ and the $l_{z(p)}^*$, both the GPCM and the GGUM were used to estimate relevant parameters. Code was written in R to compute two versions of the $l_{z(p)}$ and $l_{z(p)}^*$ statistics: one specifying the GPCM and one specifying the GGUM to estimate parameters. When GPCM was fit to the data, results for the $l_{z(p)}$ statistic were checked using the PerFit package in R. When GGUM was fit to the data, R

code closely followed publicly available code written by Tendeiro (2021) on the Open Science Framework (OSF) repository. Once parameters were estimated from each aberrant dataset within each condition and replication, model-fitting data were generated for 1,000 simulees based on these parameter estimates. Person-fit statistics were then computed for the model-fitting data. Next, the value associated with the critical point for a 5% probability of a type I error for each person-fit statistic was found as demonstrated in previous research (Emons, 2008; Tendeiro, 2017). This procedure was replicated 20 times within each replication for all study conditions and the median cutoff value was used as criteria for identifying aberrant simulees. To assess the performance of the $l_{z(p)}$ and $l^*_{z(p)}$ person-fit statistics, both type I error and detection rates were computed for each replication of each condition. Type I error was calculated by computing the detection rate for simulees in a dataset with non-aberrant responses (i.e., calculating the proportion of simulees that were incorrectly flagged as aberrant). Power was also computed for each person-fit statistic as the proportion of simulees that were generated to have aberrant response vectors that were correctly flagged as aberrant.

Results

Dimensionality and Model Fit

Results for dimensionality and model fit are described in detail in Study 1 of this dissertation. Using parallel analysis with Horn's technique, the number of factors retained were averaged across replications to compare dimensionality results across different conditions. The clean data with 20 items generated under a dominance framework (GPCM) were almost always found to be unidimensional (mean factors retained over all replications was 1.11) and more often found to retain 2 factors under the 40-item condition (M = 1.84). Adding aberrant responses to the dominance data resulted in additional factors being retained (M = 1.44 for aberrant datasets

with 20 items and *M* = 2.06 for aberrant datasets with 40 items). Data generated according to an ideal point framework (GGUM), nearly always retained 2 factors for both 20-item and 40-item datasets. Adding aberrant responses to the GGUM datasets had very little impact on dimensionality.

Comparing relative fit of each model to the data using information criteria (AIC and BIC) revealed that the GGUM may be able to fit GPCM data as well as the GPCM. However, the GPCM did not fit GGUM data as well according to these criteria. When adding aberrant responses to the data, random responding seemed to have the greatest negative impact on model fit for both GPCM and GGUM data. ERS had the lowest impact on model fit for GGUM data, in many cases resulting in improved fit compared to the clean data according to the AIC and BIC values. For GPCM data, MRS tended to impact model fit less. Nonetheless, adding aberrant responses to the dominance data always resulted in a decline in model fit.

When the correct model was used to fit the data, or when GGUM was used to fit GPCM data, results based on adjusted χ^2/df for item doublets and triplets showed that no fit issues occurred when the proportion of aberrant responders (*AbN*) was 10% or less and the proportion of aberrant responses within a response vector (*AbI*) was 40% or less. Under these conditions, the percent of flagged triplets was lower than 5% for all types of aberrant responding except MRS, where up to 14% of the triplets were flagged when GGUM fit GPCM data. However, when GPCM was fit to GGUM data, all aberrant and clean conditions resulted in 38% or more of item triplets being flagged. Further, when 20% of the sample was designated as aberrant responders and 60% of the items were replaced with aberrant scores within each response vector for those simulees, item triplets were flagged at rates of 25% or higher for all model fitting conditions except one. When GGUM was correctly fit to GGUM data, even the most severe

conditions of random responding used in the study (*AbI* = 60% and *AbN* = 20%) had negligible effects on the percent of item triplets being flagged. The flag rates were highest for ERS and MRS, reaching 90-100% for all model fitting conditions.

The quality of parameter recovery was also evaluated, as it could potentially affect the performance of the parametric person-fit statistics used in the study. When the correct model was used, or when GGUM was fit to GPCM data, estimated person parameters were always strongly correlated with their true values ($r \geq .91$), even under conditions with *AbN* = 20% and *AbI* = 60%. When GPCM was fit to GGUM data, correlations between estimated and true person parameters ranged from .65 to .71, indicating GPCM is not as capable of recovering person parameters for ideal point data compared to GGUM when used with dominance data. Average discrimination parameter recovery was slightly more affected by factors *AbI* and *AbN*, where correlations between estimated and true values declined and MAD increased with increasing levels of *AbI* and *AbN*. Discrimination parameters were nearly always underestimated. When the appropriate model was used to fit the data, estimated discrimination parameter BIAS ranged from -0.04 to -0.19 for ideal point data, and -0.04 to -0.32 for dominance data. Overall, when *AbN* was 10% or less, discrimination parameters were recovered well with the correlations ranging from .94 to .98 for ideal point data (GGUM), and from .89 to .97 for dominance data (GPCM). Once *AbN* and *AbI* reached 20% and 60% respectively, correlations between estimated and true discrimination parameters declined to .77 for dominance data and .85 for ideal point data. This is important to keep in mind because if the operationalization of these types of aberrant responding under certain conditions affect estimation of model parameters, the adequacy of the performance of the parametric person-fit statistics may be distorted. While it may be difficult to determine whether performance under these certain conditions is due to the

person-fit statistic or the operationalization of the aberrant responding, this mimics what may happen in real research situations. Looking at overall trends and considering the broader picture could reveal situations where these statistics may have difficulty in detecting aberrant responding.

Type I Error

Figure 21 shows the false detection rates (Type I error) of all person-fit statistics under all four model-data fit conditions when 20 items were used. The first model-data fit condition shows Type I error rates when GGUM was appropriately fit to GGUM data. Under this condition, the two parametric person-fit statistics ($l_{z(p)}$ and $l^*_{z(p)}$) tended to have more conservative Type I error rates, compared to the nonparametric person-fit statistics, for all types of aberrant responding except ERS. This was especially true in the case of random responding, where type I error rates were as low as .01 for both $l_{z(p)}$ and $l^*_{z(p)}$ when *AbI* = 60% and *AbN* = 20%. Type I error rates were slightly inflated for the nonparametric person-fit statistics when the aberrant responding was set to MRS under this model-data fit condition. When GPCM was appropriately fit to GPCM data, Type I error rates tended to stay at or below the 5% nominal error rate. In general, a downward trend in Type I error resulted from increasing the aberrant responding conditions *AbI* and *AbN*, except in the case where MRS was the specified aberrant response type. These results were very similar when GGUM was fit to the GPCM data. However, when GPCM was fit to GGUM data, Type I error rates were inflated under all conditions, reaching false detection rates as high as 21%.

In datasets with 40 items (Figure 22), type I error was slightly higher for correctly specified models and when GGUM was applied to GPCM data. However, rates were generally not inflated when the percentage of simulees with aberrant data was 10% or less (*AbN*) with up

to 60% *AbI*. The conditions of 20% simulees with random responding or longstrings also retained sufficient control of type I error. Inflated type I error occurred with GGUM data when 20% of simulees had ERS or MRS. In comparison, when GPCM was misapplied to GGUM data, all person-fit statistics for all aberrant conditions exhibited inflated type I error rates (> .10) with the highest type I error occurring in data with ERS simulees when using $l_{z(p)}$. Across all aberrant conditions when GPCM was misapplied to GGUM data, $l_{z(p)}$ demonstrated the highest type I error rates, whereas $l^*_{z(p)}$ exhibited the second highest type I error rates.

Figure 21. Average Type I Error for Non-Aberrant Simulees in GPCM and GGUM Datasets (20 items)

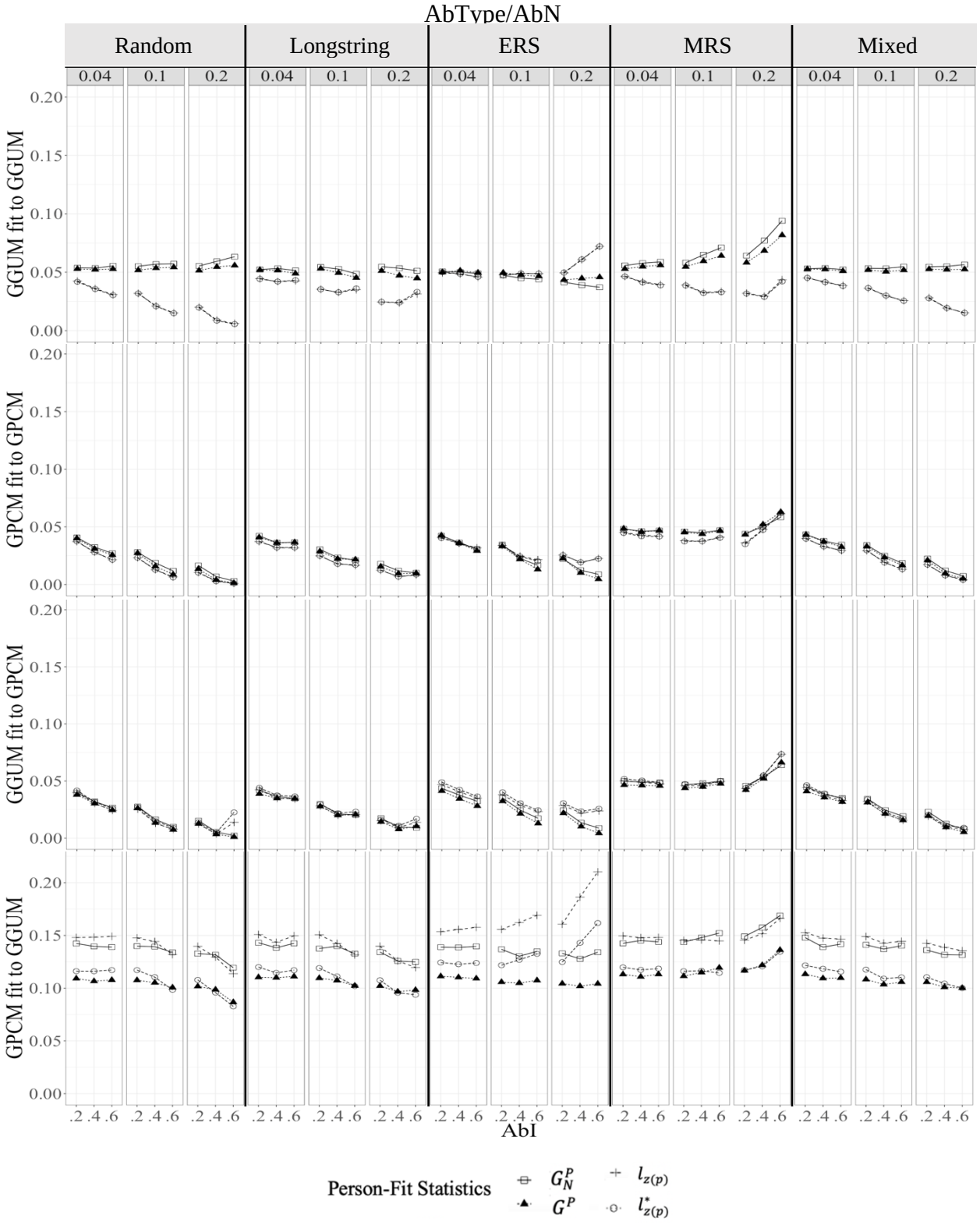
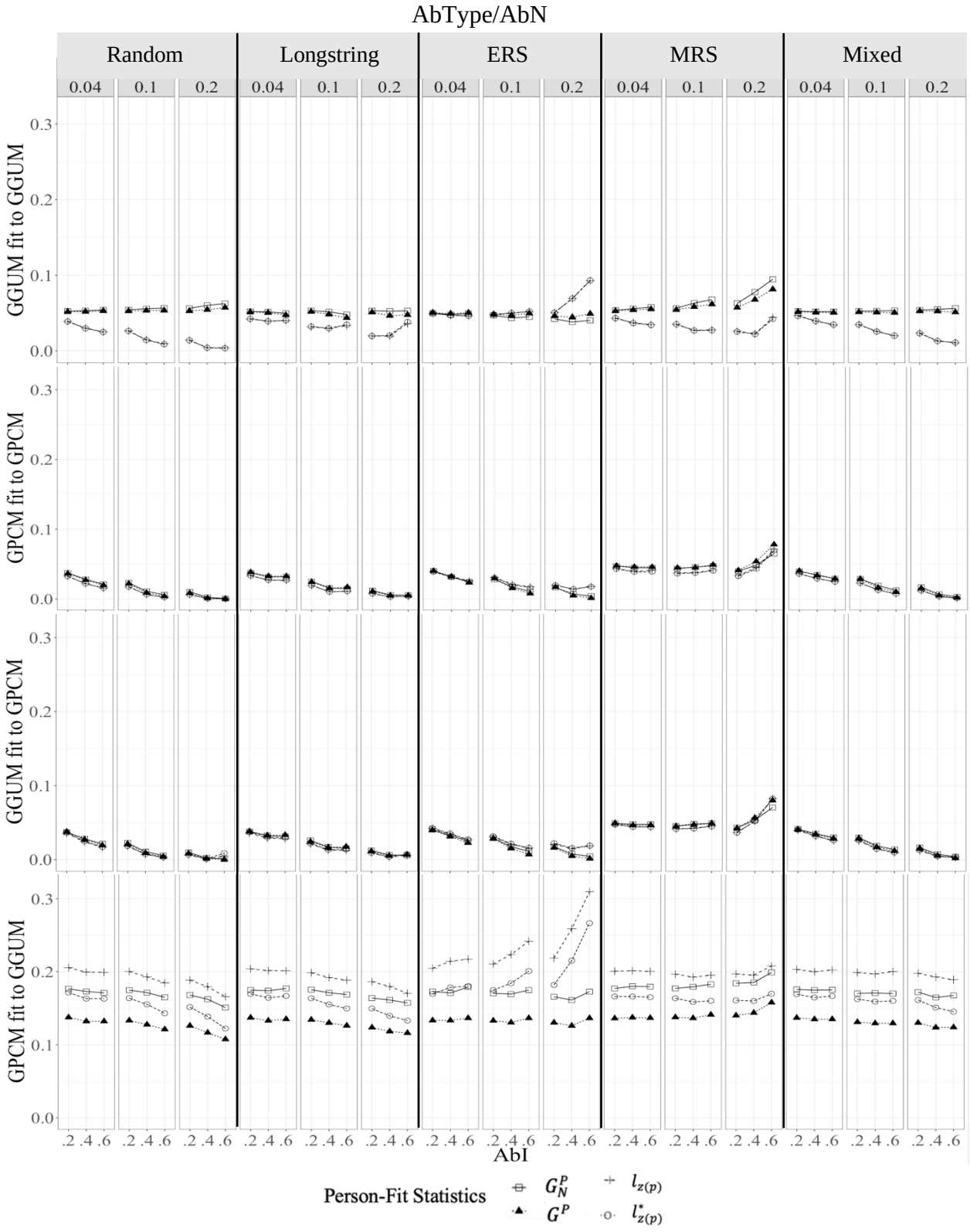


Figure 22. Average Type I Error for Non-Aberrant Simulees in GPCM and GGUM Datasets (40 items)



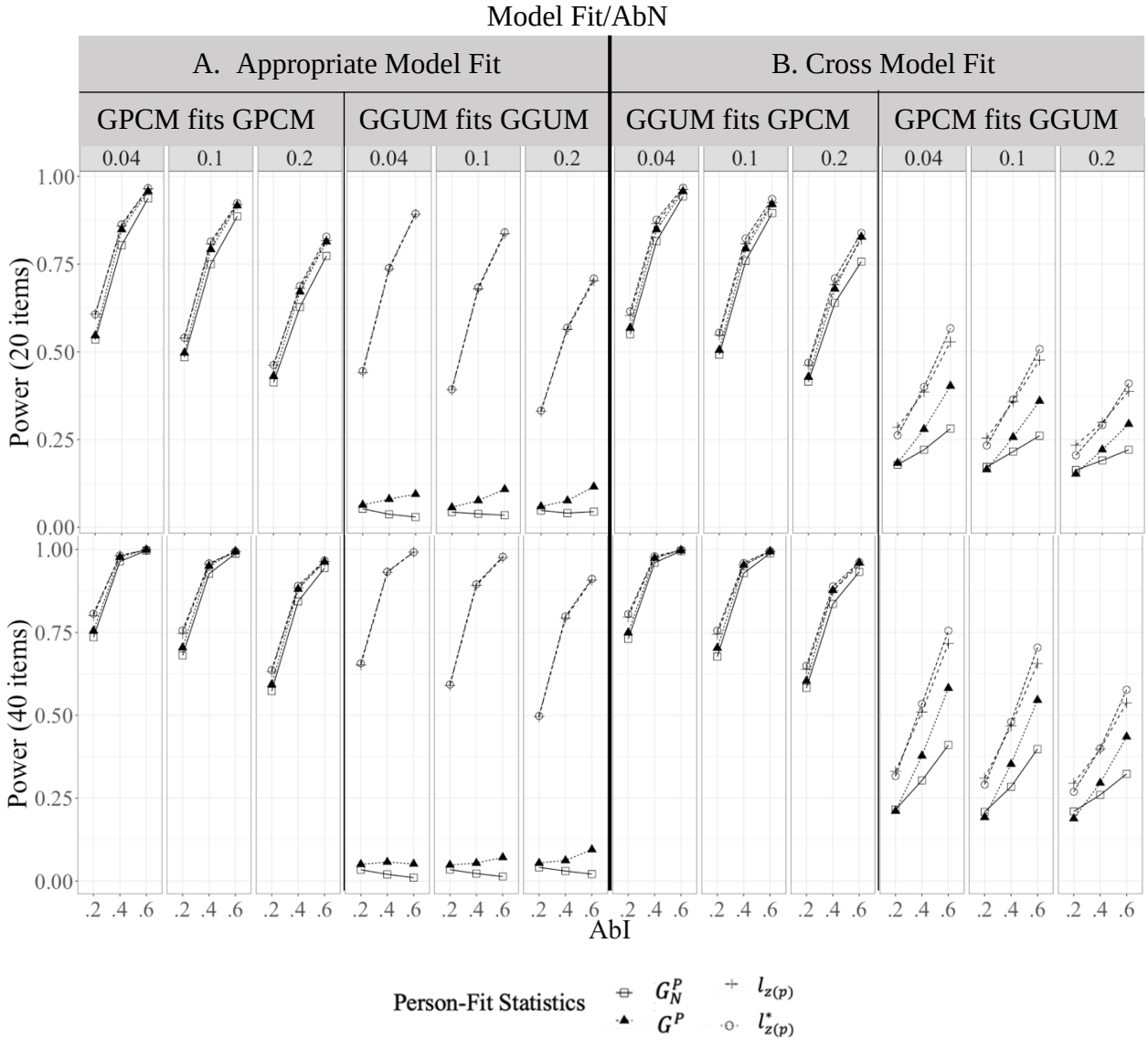
Detection of Random Responding

Figure 23 shows the average power rates for detecting random responding in datasets containing 20 (upper) and 40 items (lower). In general, increasing the number of items and level of *AbI*, and decreasing levels of *AbN*, resulted in higher detection rates for random responding, except in the case for nonparametric person-fit statistics applied to the GGUM fit to GGUM data condition. When the appropriate model was fit to the dominance data (Section A – GPCM fit to GPCM data), $l_{z(p)}$ and $l^*_{z(p)}$ and G^P tended to have slightly higher power rates than the G^P_N statistic, though the gap narrowed when the number of items increased from 20 to 40. Power was highest (97%) for the condition of *AbI* = 60% and *AbN* = 4% using $l_{z(p)}$. When GGUM was fit to the GGUM data, the parametric person-fit statistics clearly outperformed the nonparametric person-fit statistics. However, $l_{z(p)}$ and $l^*_{z(p)}$ had slightly lower power in the ideal point setting where detection rates ranged from .33 to .89 for 20-item datasets, compared to .46 to .97 in the dominance setting.

In the cross-fitting model conditions (when GGUM was fit to GPCM data or when GPCM was fit to GGUM data) shown in section B of Figure 23, trends continued to follow the pattern where increasing the number of items and *AbI* increases the ability for the person-fit statistics to detect aberrant responding. Additionally, decreasing the proportion of aberrant simulees in the sample (*AbN*) resulted in higher power rates. When GGUM was inappropriately fit to GPCM data, the person-fit statistics had very comparable power rates to when the appropriate model was fit to the GPCM data. Conversely, when GPCM was inappropriately fit to the GGUM data, results were drastically different. First, the power to detect random responding using the parametric person-fit statistics, dropped to a range of .21 to .57. Second, the detection rates of the nonparametric person-fit statistics in the ideal point setting increased. However, these

results should be interpreted jointly with the results for Type I error rates. In other words, while the power increased for the nonparametric person-fit statistics in the ideal point setting when the inappropriate model was used to fit the data and compute cut-off criteria, so did the Type I error to a large degree.

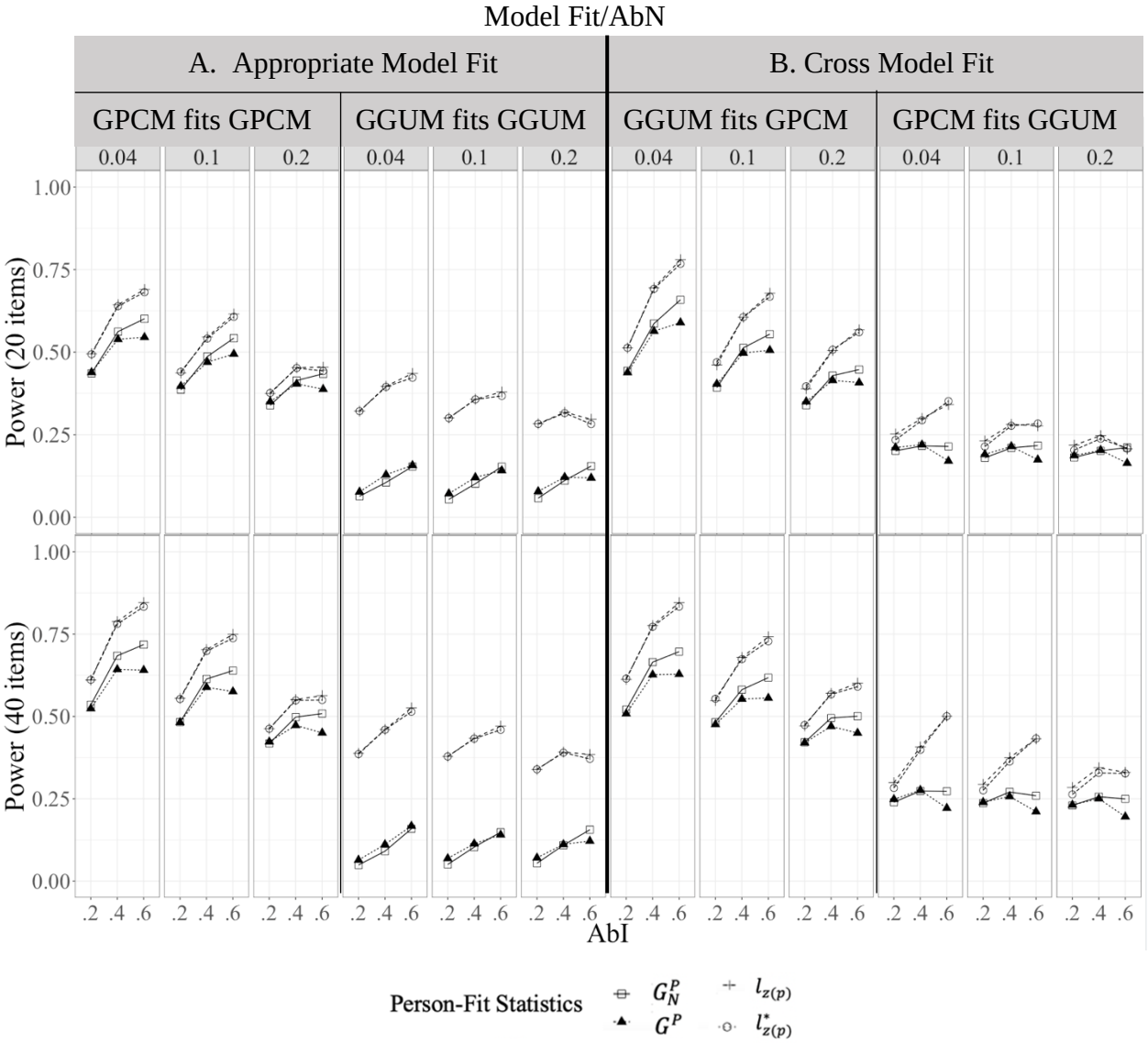
Figure 23. Average Power for Detecting Random Responding



Detection of Longstrings

Longstrings were more difficult to detect than random responding (Figure 24). The $l_{z(p)}$ and $l^*_{z(p)}$ statistics seemed to have the highest power detecting this type of aberrant responding compared to the other person-fit statistics in the study. Increasing the number of items from 20 to 40 noticeably increased power. In general, as the proportion of aberrant items within a response string increased, so did power. However, in several cases, once *AbN* hit 20% and more than half of the response string was replaced with a longstring (*AbI* = 60%), power levels either plateaued or decreased. The $l_{z(p)}$ and $l^*_{z(p)}$ statistics had the most success under all conditions, especially under the condition where GGUM was fit to the GGUM data. When GPCM was fit to GPCM data, power to detect longstrings ranged from .37 to .69 under the 20-item conditions (.46 to .85 when 40 items) using $l^*_{z(p)}$. When GGUM was appropriately fit to GGUM data, power to detect longstrings using $l^*_{z(p)}$ was comparatively low, ranging from .28 to .44 under the 20-item conditions (.34 to .53 under the 40-item conditions). When GGUM was fit to GPCM, the $l_{z(p)}$ and $l^*_{z(p)}$ statistics demonstrated slightly higher power than they did when GPCM was appropriately used for the GPCM data, however power was moderate. Power under the cross-fitting condition of GPCM fit to GGUM data revealed higher detection rates for parametric than nonparametric person-fit statistics compared to when the appropriate model was used, again, at the cost of inflated Type I error.

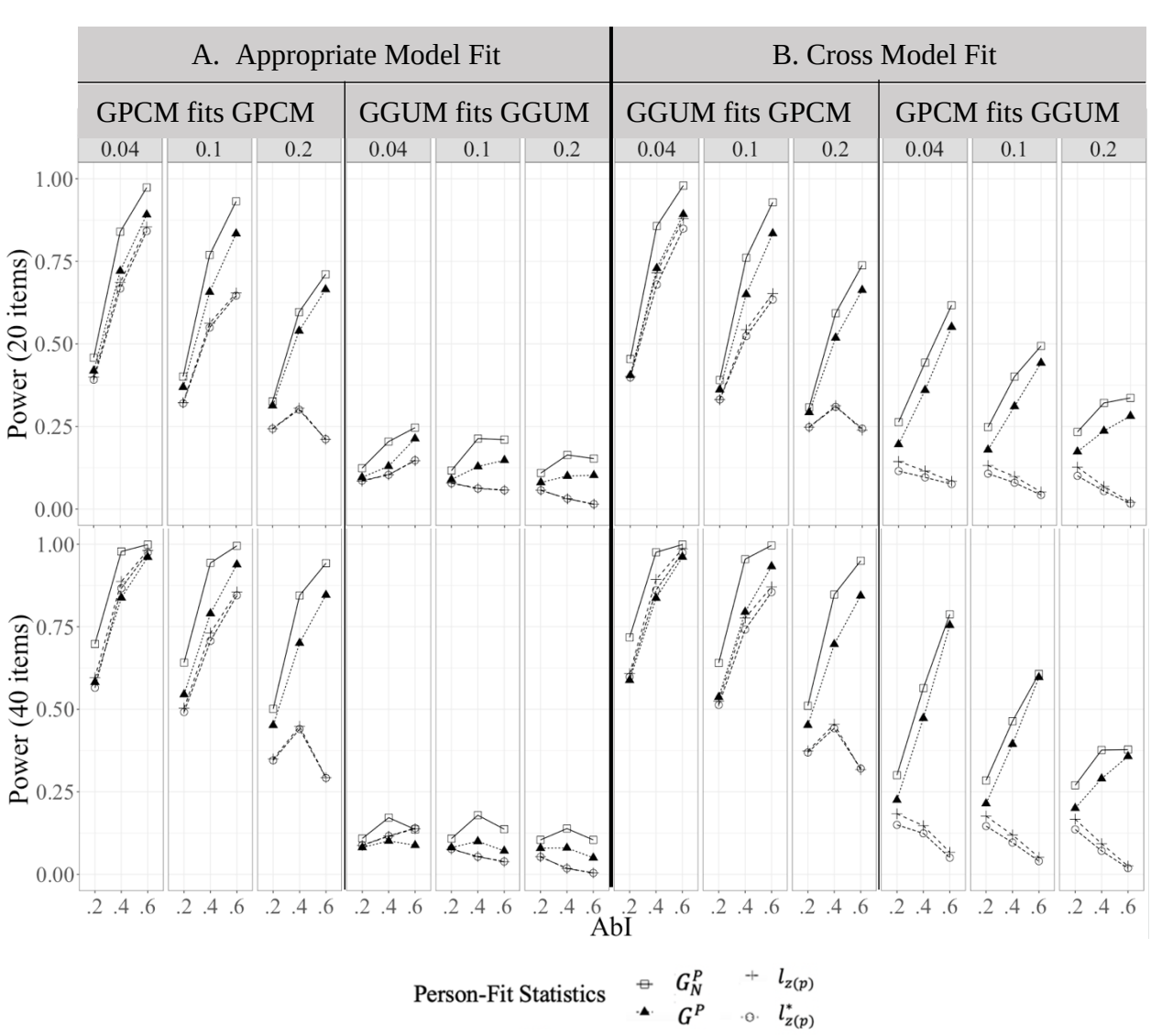
Figure 24. Average Power for Detecting Longstrings



Detection of ERS

Figure 25 shows the power rates for all person-fit statistics used to detect ERS. In the dominance setting (when GPCM was appropriately fit to the GPCM data), ERS was most easily detected by G_N^P with power reaching 100% in the 40-item datasets when *AbN* = 4% and *AbI* = 60%. The parametric person-fit statistics had the most trouble detecting ERS in all model-data fit conditions. Within the conditions where the appropriate model was fit to the 20-item datasets, in the dominance context power using the $l_{z(p)}$ and $l_{z(p)}^*$ statistics ranged from .21 to .85, and in the ideal point context power ranged from .02 to .15. For the nonparametric person-fit statistics used on the GPCM data, increasing *AbI* and the number of items, while decreasing *AbN*, generally resulted in higher power. ERS was very difficult to detect in the ideal point setting. When GGUM was fit to GGUM data, G_N^P had the highest detection rates, topping out at 25% for 20-item (18% for 40-item) datasets when *AbN* = 4% and *AbI* = 60%. Contrary to the results under the other model fitting conditions, adding items slightly decreased the detection rates for ERS when GGUM was fit to GGUM data. When aberrant conditions were highest (*AbN* = 20% and *AbI* ≥ 40%), the $l_{z(p)}$ and $l_{z(p)}^*$ statistics had the most trouble detecting ERS within GGUM data. Regarding the model cross-fit conditions, when GGUM was fit to GPCM data, detection rates were very similar to when the appropriate model was used for the GPCM data. When GPCM was fit to GGUM data, power increased for the nonparametric person-fit statistics, however caution is again provided as type I error was inflated for this condition. The lower detection rates using the $l_{z(p)}$ and $l_{z(p)}^*$ statistics, which depend on the parameter estimates from the inappropriate model fit of GPCM to GGUM data, became more evident (as low as 2% in some cases).

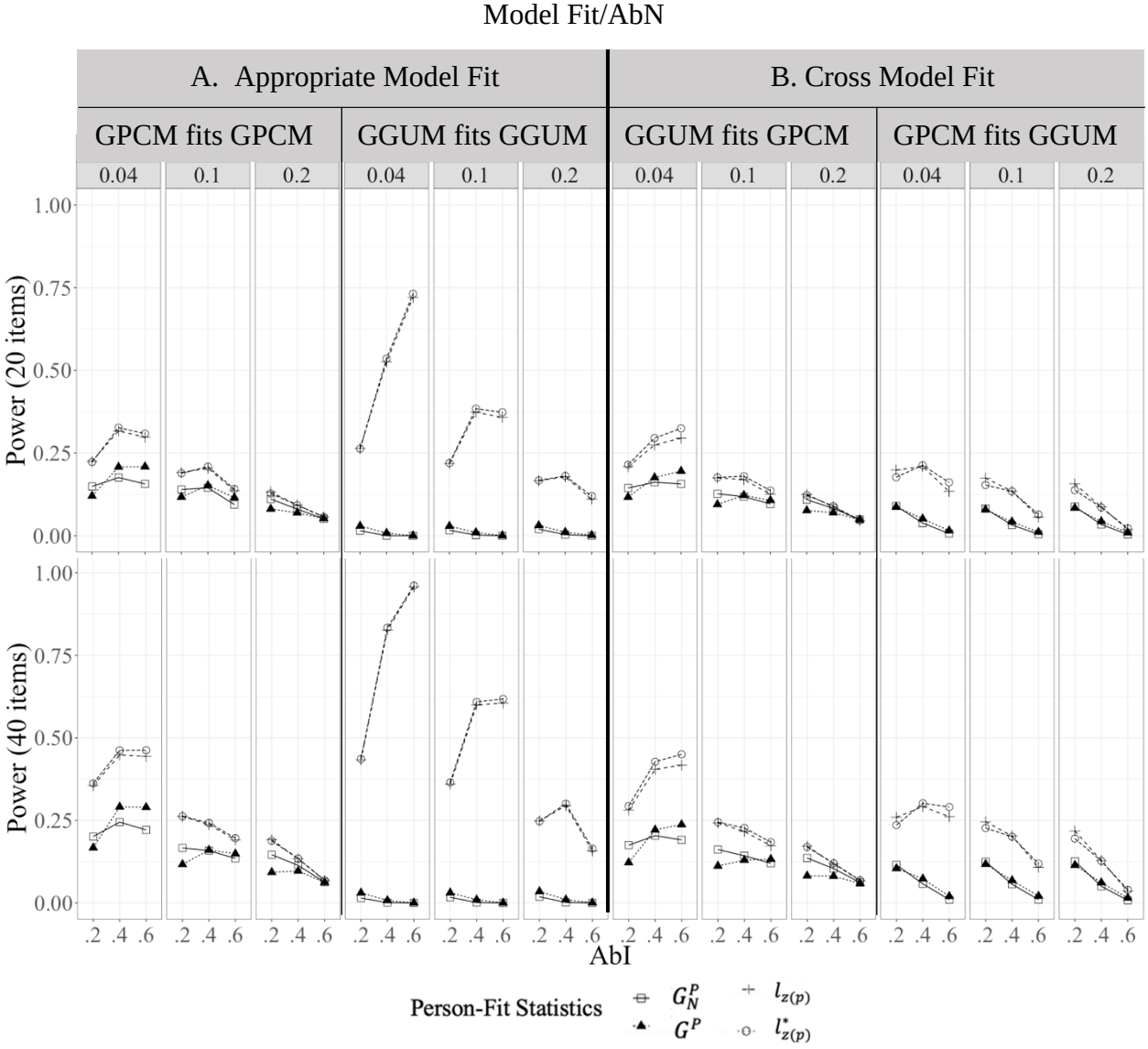
Figure 25. Average Power for Detecting Extreme Response Style (ERS)



Detection of MRS

When data reflected a dominance framework, MRS seemed to be the most difficult type of aberrant responding to detect (Figure 26). When data were generated using GPCM and the GPCM was used to fit the data, detection rates ranged from 5 to 33% across all person-fit statistics used in the study for the 20-item datasets and from 6 to 46% for the 40-item datasets. In the cleaner GPCM datasets, $l_{z(p)}$ and $l^*_{z(p)}$ had the highest detection rates, but once 10% of the sample demonstrated MRS ($AbN \geq .10$), the differences in detection rates across the different person-fit statistics became very small. Detection of MRS under the ideal point setting with the appropriate model fit to 40-item datasets had a very broad range depending on the condition (0 to 96%). When 40 items were used, *AbN* = .04, and *AbI* = .60, $l_{z(p)}$ and $l^*_{z(p)}$ were both able to detect MRS at a rate of 96%. However, using the same person-fit statistics under the same conditions except increasing *AbN* from .04 to .20, detection rates declined to 16%. When the appropriate model was fit to the GGUM data, the $l_{z(p)}$ and $l^*_{z(p)}$ statistics demonstrated a clear advantage over the nonparametric person-fit statistics in detecting MRS in the ideal point setting. In the cross-fitting conditions, power never reached above 32% for 20-item datasets and 45% for 40-item datasets. Using the highest performing person-fit statistics in each model-data fit condition, power generally increased with increasing levels of *AbI* until *AbI* reached 60%, at which point power tended to decrease. In other words, once artificial middle responses made up over half of the responses within a response vector, they often became more difficult to detect.

Figure 26. Average Power for Detecting Midpoint Response Style (MRS)

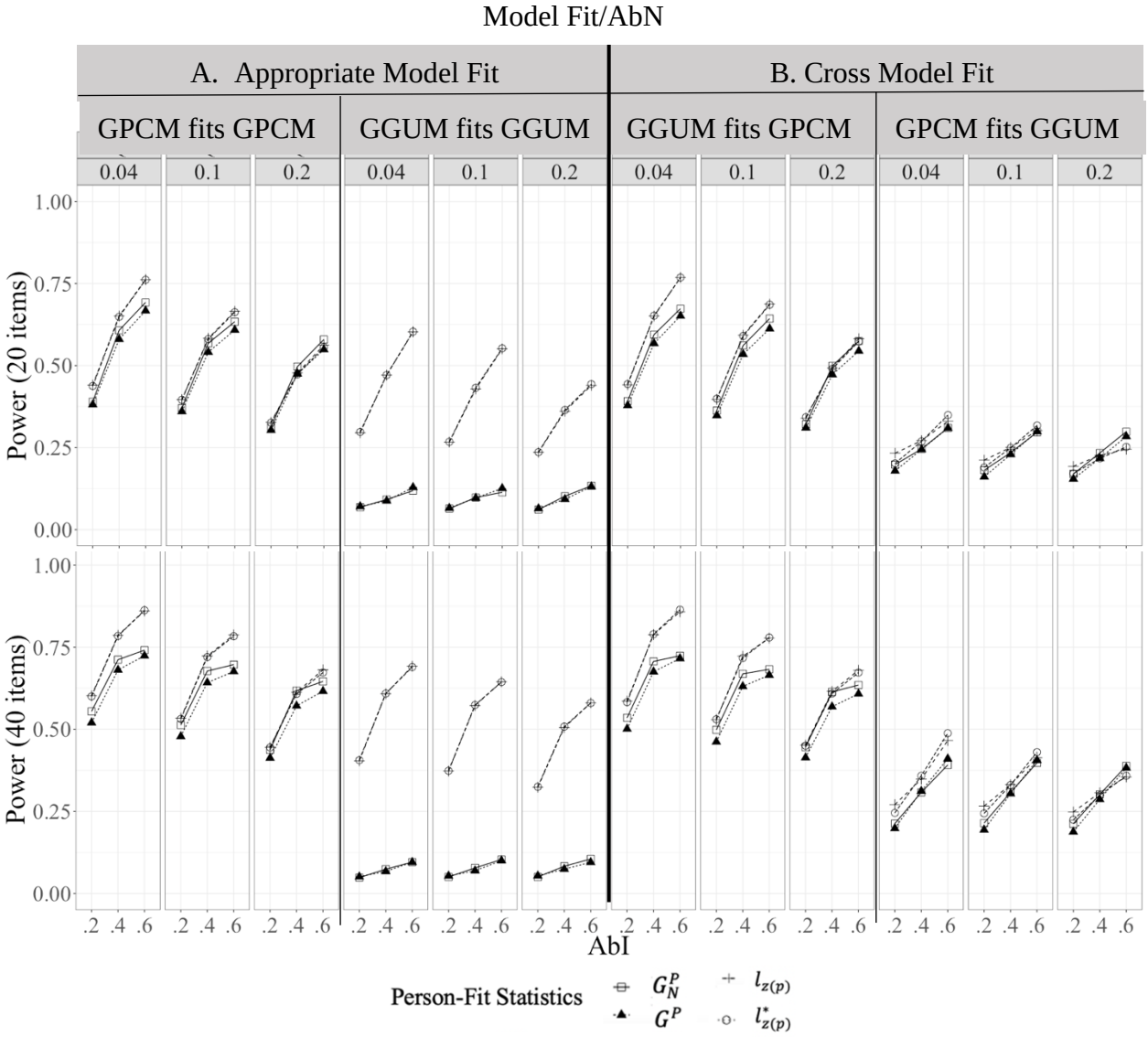


Detection of Mixed Aberrant Responding

The mixed aberrant responding condition was used to mimic what data may look like in a real testing scenario, where different respondents in the sample may demonstrate different types of aberrant tendencies. When the sample consisted of all four types of aberrant respondents (random responders, longstrings, MRS and ERS), the parametric person-fit statistics generally had higher detection rates compared the nonparametric person-fit statistics used in the study (see

Figure 27). This superiority was most notable for GGUM datasets when the appropriate model was used to estimate parameters. In the dominance data setting, the non-normed G^P demonstrated slightly lower power on average, a trend more noticeable in the 40-item datasets. When the appropriate model was fit to the data, $l_{z(p)}$ and $l^*_{z(p)}$ had slightly higher power in the dominance setting compared to the ideal point setting. For example, in the 40-item datasets with 4% *AbN* and 60% *AbI*, $l_{z(p)}$ and $l^*_{z(p)}$ both detected aberrant responding at an average rate of 86% in the GPCM datasets and 69% in the GGUM datasets. Patterns in detection rates across the different conditions for mixed aberrant responding illustrated an aggregate upward trend when *AbI* increased, the number of items increased, and when *AbN* decreased.

Figure 27. Average Power for Detecting Mixed Aberrant Responding



Discussion

Many researchers suggest that the quality of data be inspected before interpreting analytical results (e.g., DeSimone et al., 2018; Huang et al., 2012; Kline, 2009) as the presence of aberrant data can impact conclusions made from statistical analyses. As discussed in this paper, many approaches used to detect aberrant responding have been studied under a dominance

context. The nonparametric person-fit statistics used in the study are popular due to their ease of computation, not being based on strict model assumptions, and previous work that has shown their potential in detecting certain types of aberrant responding. However, these procedures have not been well-studied with ideal point response data, and the published research work regarding parametric person-fit statistics is scarce under an ideal point context (Tendeiro, 2017). Further, it was unclear how these nonparametric and parametric person-fit statistics would perform when dominance or ideal point response data are misspecified. Researchers who use nonparametric person-fit statistics may do so without assessing model fit if their primary analyses do not require the fit of an IRT model. Without knowing which type of response process is more appropriate for a set of data, researchers may be using person-fit statistics that are not effective for their data and the type of aberrant behavior present. Further, selecting cutoff criteria using a procedure that does not represent the response process of the empirical data may also have a significant impact on aberrant data detection. In this study, two nonparametric and two parametric person-fit statistics that have been shown to work well in the dominance setting were applied to ideal point data in a simulated design so that full control of factors could allow for comparison between the different approaches across various settings.

Under model misspecification, results were consistent with expectations that the Type I error would be most inflated when GPCM was inappropriately fit to data generated according to GGUM. Several studies, including the current study, suggest that unfolding models may be able to adequately fit dominance data, while the same may not be true for dominance models (e.g., GPCM) fit to unfolding data (e.g., Stark et al., 2006; Tay et al., 2011). Type I error rates reached as high as 31% when GPCM was used to fit GGUM data with 40 items. This is partially due to how the cutoff criteria was obtained. As previously mentioned, under the condition where

GPCM was fit to GGUM data, it was assumed that the researcher would also use GPCM to generate the model-fitting data in the process of obtaining the cut-off values for the person-fit statistics. For instance, for the nonparametric person-fit statistic, G^P, it can be assumed that the statistic values would be lower (indicating better person-fit) in the case for GPCM model-fitting data, where invariant item ordering and the Guttman scale is better upheld compared to a GGUM setting. In this case, the median cut-off would then be lower too. When this cutoff value was applied to the GGUM data (where G^P values were higher due to the nature of unfolding data), more vectors would be flagged, even though they may not be aberrant, resulting in high Type I error rates. Under the second cross fit condition (GGUM fit to GPCM data), Type I error rates were much closer to the 5% nominal rate, an indication that data generated according to GGUM may be an adequate approximation of GPCM data under the conditions used in the study.

To the researchers' knowledge, there is no published work on parametric person-fit analysis detecting random responding and longstrings in GGUM data. Random responding in the GGUM datasets was more prone to detection than any other type of aberrant responding in the study (longstrings, ERS, MRS and Mixed aberrant responding). The parametric person-fit statistics had the most success in detecting random responding in the unfolding setting. Longstrings were relatively more difficult to detect than random responding in both GPCM and GGUM datasets. Compared to when GPCM was appropriately fit to GPCM data, cross-fitting GGUM to GPCM data resulted in power to detect random and longstring responses that was equally as high or higher in some cases. However, when cross-fitting GPCM to GGUM data, detection rates were low for longstring and random responding, while Type I error rates were inflated under this condition.

Researchers have cautioned that when respondents within a sample are inconsistent in exhibiting response styles (such as ERS or MRS), it can be very difficult to compare participants' test scores and the validity of conclusions drawn from the data is threatened (Baumgartner & Steenkamp, 2001; de Jong et al., 2008; van Herk et al., 2004). In the presence of ERS, G_N^P had the highest detection rates across all data-model fit conditions in the study. Emons (2008) also found $U3^P$ and G_N^P to outperform $l_{z(p)}$ for detecting ERS when the dominance graded response model (GRM) was used. The $l_{z(p)}$ and $l_{z(p)}^*$ statistics had the most difficulty detecting ERS within the GGUM data when the proportion of the sample demonstrating ERS was higher (e.g., *AbN* = 20%). Considering the assumptions of nonparametric person-fit statistics, it was assumed that their detection rates would be lower in the ideal point data conditions of the study. This was the case for all GGUM data except when data included vectors exhibiting ERS. The only time nonparametric person-fit statistics outperformed the parametric person-fit statistics in the unfolding context was in the presence of ERS. Tendeiro (2017) also found ERS to be more difficult to detect than MRS in the unfolding data context using the $l_{z(p)}$ and $l_{z(p)}^*$ statistics. The overall, lower power rates for detecting ERS in the unfolding context using the parametric person-fit statistics in this study were comparable to Tendeiro (2017). Liu and Wang (2019) demonstrated how fitting standard unfolding IRT models that assume no response style results in biased parameter estimates and suggest using a general unfolding model combined with a softmax function to accommodate various response styles via scoring functions. Thus, future research could investigate the performance of the $l_{z(p)}$ and $l_{z(p)}^*$ statistics in the ideal point setting when the parameter estimates are obtained using Liu and Wang's suggested general unfolding model for multiple response styles.

Aberrant responding due to MRS was the most difficult type of aberrant responding to detect in the dominance setting. Compared to the other person-fit statistics used in the study, the $l_{z(p)}$ and $l^*_{z(p)}$ statistics had the most success detecting MRS in both dominance and ideal point settings. In the ideal point data conditions, MRS was more effectively detected than ERS by the $l_{z(p)}$ and $l^*_{z(p)}$ statistics, which is also a reported finding on similar conditions in Tendeiro's (2017) study. The parametric person-fit statistics demonstrated a broad range of detection rates for MRS in the ideal point settings, where detection rates were reasonably high unless the proportion of the sample with MRS (*AbN*) reached 20%. The detection rates for ERS and MRS in the unfolding and dominance settings may be affected by the proportion of extreme and midpoint item scores present in the randomly generated GGUM and GPCM data. For example, the ratio of extreme scores (0, 5) to midpoint scores (3, 4) in the clean GGUM datasets ranged from 1.40 to 2.08 (similar to what was reported in Tendeiro's [2017] paper where the range was 1.40 to 1.89). If more extreme scores naturally existed in the data, then it makes sense that extreme response style would be more difficult to detect. Conversely, it is reasonable to postulate that MRS may be relatively easier to detect in this setting.

Results also demonstrated how none of the person-fit statistics in the study were robust to model misspecification of the GPCM to GGUM data as demonstrated by the inflated type I error. These results serve as an important caveat to be careful when classifying aberrant responding within the context of ideal point data. Even when a dominance model may have adequate fit as deemed by some statistics, the person-fit statistic Type I error may be greatly inflated if the underlying response process assumptions are violated. Future research could involve different methods for obtaining cutoff criteria which may help alleviate this potential problem. On the

other hand, when GGUM was misspecified to GPCM data, both Type I error and power were comparable to results when the appropriate model (GPCM) was used.

Limitations

There are several limitations of the study that warrant attention in the interpretation of results. Firstly, while the purpose of using a simulation study was to have full control over factors, and these factors were chosen to approximate realistic settings, the results may be limited in their generalizability to other factors and levels of the factors not used in the study. For example, the number of response categories and item discrimination were not factors in the study but have been found to affect the performance of person-fit statistics (Tendeiro, 2017; Tendeiro & Meijer, 2014). Another limitation concerns possible effects of methodological choices regarding how aberrant data were simulated. That is, an assumption was made that the creation of aberrant data accurately reflects, or at least approximates real response behavior that is considered aberrant. Along these lines, the definition of aberrant data could be different depending on the purpose and nature of the assessment. Further, the procedures used for determining cut-off criteria may have a significant effect on the results reported in this study, as the use of data replication processes for identifying cut-score heuristics would provide different results than pre-set criterion values. Finally, it should be realized that a confounding effect resulting from the impact of aberrant data on model fit and parameter estimates likely affected the performance of the person-fit statistics. St-Onge et al. (2011) report that the accuracy of person-fit statistics may increase to a certain point, and then decrease with continued increases in the amount of aberrant responses in the data. Thus, the operationalization of the different types of aberrant behavior may have systematically biased parameter estimates, which in turn could be the reason for the decline in person-fit statistic performance in some conditions.

Conclusions

Practitioners should carefully choose the person-fit indices that make the most sense for their analyses. This study evaluated several parametric and nonparametric person-fit statistics for detecting random, longstring, ERS, MRS and mixed aberrant responding present in both dominance (GPCM) and ideal point (GGUM) data. The simulation results show that nonparametric person-fit statistics have a very limited ability to detect aberrant responding in the ideal point context. Additionally, extreme response style may be very difficult to detect in data generated according to GGUM, while MRS is difficult to detect in data generated according to the GPCM. Future research could improve on the detection of these types of aberrant responding under the two different response process assumptions.

Prior to this study, it was unclear how the person-fit statistics would perform in detecting longstring and random responding in ideal point settings. The parametric person-fit statistics had high power for detecting random responding in the ideal point setting, and comparable power to results in the dominance data for longstrings. It is recommended that researchers be particularly careful when applying person-fit statistics to unfolding data as model fit and choice of statistics could greatly impact results.

CHAPTER 7

OVERALL CONCLUSIONS

The three studies presented in this dissertation explore the model fit, impacts of aberrant responding, and detection of aberrant responding under dominance and unfolding model contexts while varying conditions of test length, proportion of aberrant responders in the sample, proportion of aberrant responses within a response vector, types of aberrant responding, and model-data generation and specification. The investigation began with Study 1, where the impacts of aberrant responding on model fit for two parametric IRT models (GPCM and GGUM) based on different underlying response process assumptions were examined. One of the first endeavors of an empirical researcher often involves testing model assumptions and ultimately model fit, though recommendations for assessing dimensionality under an unfolding framework (compared to dominance) are far less apparent in published literature. Results from Study 1 for the clean (non-aberrant) GGUM datasets coincided with other studies that have found an additional spurious factor appearing for unfolding data (Tay et al., 2011; Tay & Drasgow, 2012; Williams, 2015). Aberrant responding using longstrings tended to have a greater impact on dimensionality assessment compared to the other types of aberrant responding in the study for both GPCM and GGUM datasets, with additional factors being retained from the parallel analyses as the occurrences of longstrings increased. Results also suggested that GGUM was able to fit GPCM data reasonably well, however the same was not true for GPCM fit to the GGUM data, a finding consistent with the literature (Chernyshenko et al., 2007; Stark et al., 2006). A valid concern for the practical researcher involves the unknowingness of whether model-data misfit is a result of poor-quality data or if it is due to model misspecification. Study 1 results emphasize the importance of carefully examining the quality of data before making

conclusions about model misspecification. It was also a critical precursor to Studies 2 and 3 because model fit can impact the detection of aberrant responding which was the focus of these studies.

The detection of aberrant responding under a dominance framework (using GPCM) has been widely researched. However, the studies in this dissertation included combinations of types and amounts of aberrant responding that add to the current literature in this setting. Random responding and ERS were relatively easier to detect than longstrings, and MRS was actually very difficult to detect. In the GPCM datasets with less aberrant data, $l_{z(p)}$ and $l^*_{z(p)}$ had the highest detection rates for MRS compared to the other types of person-fit statistics, but once 10% of the sample demonstrated MRS ($AbN \geq .10$), the H^T statistic usually had higher detection rates. The parametric person-fit statistics, $l_{z(p)}$ and $l^*_{z(p)}$, also had higher detection rates than the other studied person-fit statistics for longstrings and random responding (along with G^P). However, detection of ERS in the dominance data was highest using the normed nonparametric person-fit statistics, $U3^P$ and G^P_N, similar to results from Emons (2008).

Within the conditions involving unfolding data, the detection of aberrant responding was reasonably effective using the parametric person-fit statistics, $l_{z(p)}$ and $l^*_{z(p)}$, though power was slightly lower in most conditions compared to the power of these two statistics in the dominance setting. Similar to the results for the dominance data, the parametric person-fit statistics had the most success in detecting random responding in the unfolding data. However, these statistics had the most trouble detecting ERS. The difficulty in detecting ERS was also a finding for Tendeiro (2017), where the author notes this could be a consequence of the higher proportion of extreme responses generated in GGUM data, making ERS less likely to be identified as abnormal. Generally, all nonparametric person-fit statistics had a clear disadvantage in detecting aberrant

responding in GGUM data, except in the case of ERS, where $U3^P$ and G_N^P outperformed the parametric person-fit statistics. Nonetheless, power never reached above 25% for the nonparametric statistics applied to GGUM data when using the appropriate model to generate cutoff criteria. Currently, no known nonparametric person-fit statistics exist that have been shown to detect aberrant responding reasonably well in the unfolding context.

The simulations in Study 3 also explored the effects of model misspecification on the performance of person-fit statistics under various conditions of aberrant responding. Results demonstrated how type I error rates when GPCM was misspecified to GGUM data were inflated compared to when the correct model was specified and compared to when GGUM was misspecified to GPCM data. Results also suggest that detection rates are affected by model misspecification when using the resampling-based method described in the study to obtain cutoff criteria. Again, this was most notable for the misspecification condition where GPCM was fit to GGUM data. Based on the findings from the study, type I error and detection rates were comparable for the conditions where the correct model was used and when GGUM was fit to GPCM data. Hong et al. (2020) encourage researchers to consider specifying relatively more complex alternative models when computing person-fit statistics. The current study supports this recommendation if the appropriate model is not known. As GGUM is considered the more complex model in the study, it may be used to fit GGUM and GPCM data to produce person-fit statistics with relatively similar results compared to when the most appropriate model is used. However, fitting GPCM (less complex) to GGUM data resulted in severely inflated type I error.

Limitations

A few principal limitations noted within each of the three studies are worth highlighting here. The leading purpose behind using a simulation design for the studies was to have full

control over the factors which would not have been possible using empirical data. While conditions were carefully chosen to reflect realistic conditions when using real data, the generalizability of results to real data may be limited to settings where conditions closely follow those used in the study. These studies used only two IRT models, two test lengths, and fixed sample sizes and number of response categories. In reality, results may be different for the numerous other possible levels of these conditions. Additionally, method effects regarding how the aberrant data were simulated could have impacted results. It was assumed that the way aberrant data (e.g., ERS, MRS) was generated mimics the behavior of respondents who exhibit this behavior in reality. Further, it was noted that data generated according to GGUM had a larger ratio of extreme to middle responses. The predominance of extreme responses within the unfolding data may have been a contributing factor to the low power to detect ERS since extreme responses would not be as abnormal in these datasets. Lastly, only 100 replications were conducted for the studies due to the restrictive completion times needed per condition. Ideally, more replications would be run to improve the reliability of the results.

Future Research

In order to further research in this area, studies could focus on using different levels of various conditions including test length, sample size, and number of response categories. The use of other dominance and unfolding models (other than GPCM and GGUM) in the process of data generation and model fit would also contribute to the literature. Item discrimination parameters have also been found to influence detection rates using person-fit statistics (Tendeiro & Meijer, 2012) and could be manipulated to explore the effect on detection of aberrant responding in the unfolding context.

Because GGUM is a unidimensional IRT model, researchers must test the assumption of unidimensionality before applying the model in an IRT analysis (Chernyshenko et al., 2007). As previously mentioned, a substantial amount of published work can be found on dimensionality assessment under a dominance framework. However, literature is lacking with regards to dimensionality assessment under an unfolding framework. Future work could focus on establishing methodology used to assess dimensionality under the unfolding model framework or expand and validate the work of other recent advances in this area (Fan, 2020; Joo et al., 2021).

Although this dissertation was an attempt to contribute to sparse literature regarding the detection of aberrant responding in an unfolding model context, many questions still remain, and additional research is needed. As noted in the limitations, to the best of the author's knowledge, no nonparametric person-fit statistic has been developed for detecting aberrant responding in an unfolding model context. The use of nonparametric unfolding models (e.g., MUDFOLD; Balafas, 2016) could assist in the development of a nonparametric person-fit statistic. It is unclear how other person-fit statistics not included in the current dissertation perform in an unfolding model context. Additionally, the way cutoff criteria were obtained for each person-fit statistic used in the study could have influenced results. A separate study could investigate how cutoff criteria methodology impacts detection of aberrant responding under various conditions for dominance and unfolding contexts, as well as model misspecification. As research has indicated that the use of ideal point response models may provide more flexibility in applications to unidimensional empirical data than dominance models, continuing research to identify person-fit statistics that function effectively for unfolding data is recommended.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Andrich, D. (1988). The application of an unfolding model of the pirt type to the measurement of attitude. *Applied Psychological Measurement*, *12*(1), 33–51. https://doi.org/10.1177/014662168801200105

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology*, *49*(2), 347–365. https://doi.org/10.1111/j.2044-8317.1996.tb01093.x

Andrich, D. & Guanzhong Luo. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, *17*(3), 253–276. https://doi.org/10.1177/014662169301700307

Armstrong, R. D., & Shi, M. (2009a). Model-free CUSUM methods for person fit. *Journal of Educational Measurement*, *46*(4), 408–428. https://doi.org/10.1111/j.1745-3984.2009.00090.x

Armstrong, R. D., & Shi, M. (2009b). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, *33*(5), 391–410. https://doi.org/10.1177/0146621609331961

Armstrong, R. D., Stoumbos, Z. G., T., K. M., & Shi, M. (2007). On the performance of the lz person-fit statistic. *Practical Assessment, Research, and Evaluation*, *12*. https://doi.org/10.7275/XZ5D-7J62

Artner, R. (2016). A simulation study of person-fit in the Rasch model. *Psychological Test and Assessment Modeling*, *58*(3), 531–563.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *The Public Opinion Quarterly*, *48*(2), 491–509. JSTOR.

Baker, F. (2001). *The basics of item response theory* (Second). ERIC Clearinghouse on Assessment and Evaluation. https://files.eric.ed.gov/fulltext/ED458219.pdf

Balafas, S. E. (2020). Mudfold: A nonparametric item response theory model for unidimensional unfolding. *The R Journal*, *12*(1), 115.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Beck, M. F., Albano, A. D., & Smith, W. M. (2019). Person-fit as an index of inattentive responding: A comparison of methods using polytomous survey data. *Applied Psychological Measurement*, *43*(5), 374–387. https://doi.org/10.1177/0146621618798666

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340–345. https://doi.org/10.1037/1040-3590.4.3.340

Bortolotti, S. L. V., Tezza, R., de Andrade, D. F., Bornia, A. C., & de Sousa Júnior, A. F. (2013). Relevance and advantages of using the item response theory. *Quality & Quantity*, *47*(4), 2341–2360. https://doi.org/10.1007/s11135-012-9684-5

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*(2), 218–229. https://doi.org/10.1037/pspp0000085

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI Work Satisfaction scale. *Personality and Individual Differences*, *49*(7), 743–748. https://doi.org/10.1016/j.paid.2010.06.019

Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M.-C., & Delgado, K. M. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: How theoretically appropriate measurement makes an empirical difference. *Journal of Applied Psychology*, *99*(4), 564–586. https://doi.org/10.1037/a0034688

Cavalini, P. M. (1992). *It's an ill wind that brings no good. Studies on odour annoyance and the dispersion of odorant concentrations from industries.* [University of Groningen]. https://core.ac.uk/reader/148291930

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(1), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, *6*(3), 170–175. https://doi.org/10.1111/j.1467-9280.1995.tb00327.x

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523–562. https://doi.org/10.1207/S15327906MBR3604_03

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*(1), 88–106. https://doi.org/10.1037/1040-3590.19.1.88

Cizek, G. J., & Wollack, J. A. (2016). *Handbook of quantitative methods for detecting cheating on tests*. Taylor & Francis.

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back random responding: Effectiveness of MMPI-2 and Personality Assessment Inventory validity indices. *Psychological Assessment*, *15*(2), 223–234. https://doi.org/10.1037/1040-3590.15.2.223

Conijn, J. M., Emons, W. H. M., & Sijtsma, K. (2014). Statistic lz-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement*, *38*(2), 122–136. https://doi.org/10.1177/0146621613497568

Conn, S. R., & Rieke, M. L. (1994). *The 16PF fifth edition technical manual*. Inst for Personality & Ability Testing.

Coombs, C. H. (1964). *A theory of data* (pp. xviii, 585). Wiley.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612. https://doi.org/10.1177/0013164410366686

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

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*(4), 429–449. https://doi.org/10.1111/j.1745-3984.2009.00091.x

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Cyr, L. (2000). *Insufficient effort responding on mturk surveys: Evidence-based quality control for organizational research* [Portland State University]. https://doi.org/10.15760/etd.6337

Dalal, D. K., Carter, N. T., & Lake, C. J. (2014). Middle response scale options are inappropriate for ideal point scales. *Journal of Business and Psychology*, *29*(3), 463–478. JSTOR.

de Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*(1), 104–115.

Degner, L. F., Sloan, J. A., & Venkatesh, P. (1997). The control preferences scale. *Canadian Journal of Nursing Research Archive*, 21–44.

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology*, *67*(2), 309–338. https://doi.org/10.1111/apps.12117

DeSimone, J. A., & Harms, P. D. (2017). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology*, *33*(5), 559–577. https://doi.org/10.1007/s10869-017-9514-9

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171–181. https://doi.org/10.1002/job.1962

Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied Measurement*, *7*(2), 170–183.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, *3*(4), 465–476. https://doi.org/10.1111/j.1754-9434.2010.01273.x

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In *Handbook of industrial and organizational psychology, Vol. 1, 2nd ed* (pp. 577–636). Consulting Psychologists Press.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*(2), 143–166. https://doi.org/10.1177/014662169501900203

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*(3), 224–247. https://doi.org/10.1177/0146621607302479

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Applied Psychological Measurement*, *33*(8), 599–619. https://doi.org/10.1177/0146621609334378

Emons, W. H. M., Glas, C. A. W., Meijer, R. R., & Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement*, *27*(6), 459–478. https://doi.org/10.1177/0146621603259270

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research*, *39*(1), 1–35. https://doi.org/10.1207/s15327906mbr3901_1

Fan, Y. (n.d.). *A method for assessing the dimensionality of unfolding responses* [Ph.D., University of Georgia]. Retrieved March 31, 2022, from https://www.proquest.com/docview/2489638918/abstract/4E7CA061BBAB43E7PQ/1

Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research*, *42*(3), 481–507.

Ferrando, P. J. (2009). Multidimensional factor-analysis-based procedures for assessing scalability in personality measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(1), 109–133. https://doi.org/10.1080/10705510802561352

Ferrando, P. J. (2010). Some statistics for assessing person-fit based on continuous-response models. *Applied Psychological Measurement*, *34*(4), 219–237. https://doi.org/10.1177/0146621609343288

Freund, P. A., & Lohbeck, A. (2021). Modeling self-determination theory motivation data by using unfolding IRT. *European Journal of Psychological Assessment*, *37*(5), 388–396. https://doi.org/10.1027/1015-5759/a000629

Gessaroli, M. E., & De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement*, *33*(2), 157–179.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for irt models for polytomous items. *Psychometrika*, *72*(2), 159–180. https://doi.org/10.1007/s11336-003-1081-5

Glas, C. A. W., & Meijer, R. R. (2003). A bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*(3), 217–233. https://doi.org/10.1177/0146621603027003003

Greenleaf, E. A. (1992). Measuring extreme response style. *The Public Opinion Quarterly*, *56*(3), 328–351. JSTOR.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*(2), 139–150. JSTOR. https://doi.org/10.2307/2086306

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and Applications*. Springer Science & Business Media.

Harris-Watson, A., Mcmillan, J., & Carter, N. (2020). Test-taker reactions to ideal point measures of personality. *Journal of Business and Psychology*. https://doi.org/10.1007/s10869-020-09682-8

Hendrawan, I., Glas, C. A. W., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement*, *29*(1), 26–44. https://doi.org/10.1177/0146621604270902

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement*, *15*(2), 153–169. https://doi.org/10.1177/014662169101500205

Hong, M., Lin, L., & Cheng, Y. (2021). Asymptotically corrected person fit statistics for multidimensional constructs with simple structure and mixed item types. *Psychometrika*, *86*(2), 464–488. https://doi.org/10.1007/s11336-021-09756-3

Hong, S. E., Monroe, S., & Falk, C. F. (2020). Performance of person-fit statistics under model misspecification. *Journal of Educational Measurement*, *57*(3), 423–442. https://doi.org/10.1111/jedm.12207

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

Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. *The Public Opinion Quarterly*, *49*(2), 253–260.

Javaras, K. N., & Ripley, B. D. (2007). An "unfolding" latent variable model for likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association*, *102*(478), 454–463. https://doi.org/10.1198/016214506000000960

Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (2018). Mixture item response models for inattentive responding behavior. *Organizational Research Methods*, *21*(1), 197–225. https://doi.org/10.1177/1094428117725792

Jin, K.-Y., & Wang, W.-C. (2014). Generalized irt models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 116–138. https://doi.org/10.1177/0013164413498876

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Johnson, M. S., & Junker, B. W. (2003). Using data augmentation and markov chain monte carlo for the estimation of unfolding response models. *Journal of Educational and Behavioral Statistics*, *28*(3), 195–230. JSTOR.

Joo, S.-H., Chun, S., Stark, S., & Chernyshenko, O. S. (2019). Item parameter estimation with the general hyperbolic cosine ideal point IRT model. *Applied Psychological Measurement*, *43*(1), 18–33. https://doi.org/10.1177/0146621618758697

Joo, S.-H., Lee, P., Park, J. Y., & Stark, S. (2021). Assessing dimensionality of the ideal point item response theory model using posterior predictive model checking. *Organizational Research Methods*, 10944281211050608. https://doi.org/10.1177/10944281211050609

Joo, S.-H., Lee, P., & Stark, S. (2017). Evaluating anchor-item designs for concurrent calibration with the ggum. *Applied Psychological Measurement*, *41*(2), 83–96. https://doi.org/10.1177/0146621616673997

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298. https://doi.org/10.1207/S15324818AME1604_2

Kartal, S., & Dirlik, E. M. (2021). Examining the dimensionality and monotonicity of an attitude dataset based on the item response theory models. *International Journal of Assessment Tools in Education*, *8*(2), 296–309. https://doi.org/10.21449/ijate.728362

Kartal, S. K., & Kutlu, O. (2020). Analyzing the dimensionality of academic motivation scale based on the item response theory models. *Eurasian Journal of Educational Research*, 18.

Krimpen-Stoop, E. M. L. A. van, & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*(2), 199–217. JSTOR.

Kung, F. Y. H., Kwok, N., & Brown, D. J. (2018). Are attention check questions a threat to scale validity? *Applied Psychology*, *67*(2), 264–283. https://doi.org/10.1111/apps.12108

Lahuis, D. M., & Clark, P. (2009). *An examination of item response theory item fit indices for the graded response model. Manuscript submitted for publication*.

LaPalme, M., Tay, L., & Wang, W. (2018). A within-person examination of the ideal-point response process. *Psychological Assessment*, *30*(5), 567–581.

Lee, P., Stark, S., & Chernyshenko, O. S. (2014). Detecting aberrant responding on unidimensional pairwise preference tests: An application of lz based on the Zinnes–Griggs ideal point IRT model. *Applied Psychological Measurement*, *38*(5), 391–403. https://doi.org/10.1177/0146621614526636

Levine, M. (1984). *An introduction to multilinear formula score theory.* (Office of Naval Research No. 84–4; Measurement Series). Personnel and Training Research Programs. https://apps.dtic.mil/dtic/tr/fulltext/u2/a145696.pdf

Li, M. F., & Olejnik, S. (1997). The power of rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, *21*(3), 215–231. https://doi.org/10.1177/01466216970213002

Ligtvoet, R., Ark, L. A. V. D., Marvelde, J. M. T., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*(4), 578–595.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22 140*, 55–55.

Liu, C.-W., & Wang, W.-C. (2019). A general unfolding irt model for multiple response styles. *Applied Psychological Measurement*, *43*(3), 195–210. https://doi.org/10.1177/0146621618762743

Liu, J., & Zhang, J. (2020). An item-level analysis for detecting faking on personality tests: Appropriateness of ideal point item response theory models. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.03090

Liu, T., Sun, Y., Li, Z., & Xin, T. (2019). The impact of aberrant response on reliability and validity. *Measurement*, *17*(3), 133–142. https://doi.org/10.1080/15366367.2019.1584848

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, *33*(8), 579–598. https://doi.org/10.1177/0146621609331960

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's lz* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *37*(1), 57–81. JSTOR.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology*, *23*(4), 498–509. https://doi.org/10.1177/0022022192234006

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450–470. https://doi.org/10.1037/a0019216

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

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3. https://doi.org/10.1207/s15324818ame0901_2

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement*, *21*(2), 99–113. https://doi.org/10.1177/01466216970212001

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, *9*(3), 354–368. https://doi.org/10.1037/1082-989X.9.3.354

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*(2), 111–120. https://doi.org/10.1177/014662169401800202

Meijer, R. R., Muijtjens, A., & van der Vlueten, C. (1996). Nonparametric person-fit research: Some theoretical issues an empirical example. *Applied Measurement in Education*, *9*(1), 77. https://doi.org/10.1207/s15324818ame0901_7

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135. https://doi.org/10.1177/01466210122031957

Meijer, R. R., & Tendeiro, J. N. (2012). The use of the lz and lz* person-fit Statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics*, *37*(6), 758–766.

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. Mouton.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item response theory* (pp. 351–367). Springer. https://doi.org/10.1007/978-1-4757-2691-6_20

Molenaar, I. W. (1990). *A Weighted Loevinger H-coefficient Extending Mokken Scaling to Multicategory Items*. Psychologische Instituten der Rijksuniversiteit Groningen.

Molenaar, I. W. (1991). *A weighted Loevinger h-coefficient extending Mokken scaling to multicategory items*. Psychologische Instituten der Rijksuniversiteit Groningen.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*(1), 75–106. https://doi.org/10.1007/BF02294745

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Murphy, M. K. (1996). *Psychological aspects of survey methodology: Experiments on the response process*. University of London.

Nandakumar, R., Hotchkiss, L., & Roberts, J. S. (2002). *Attitudinal data: Dimensionality and start values for estimating item parameters*. 30.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*, *22*(1), 53–69. https://doi.org/10.1177/01466216980221004

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Nye, C. D., Joo, S.-H., Zhang, B., & Stark, S. (2020). Advancing and evaluating irt model data fit indices in organizational research. *Organizational Research Methods*, *23*(3), 457–486. https://doi.org/10.1177/1094428119833158

Paulhus, D. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609.

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, *33*(1), 36–48. https://doi.org/10.1111/emip.12023

Polak, M., de Rooij, M., & Heiser, W. J. (2012). A model-free diagnostic for single-peakedness of item responses using ordered conditional means. *Multivariate Behavioral Research*, *47*(5), 743–770. https://doi.org/10.1080/00273171.2012.715563

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests* (pp. xiii, 184). Nielsen & Lydiche.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, *14*(2), 95–101. https://doi.org/10.1111/j.0963-7214.2005.00342.x

Roberts, J. S., & Cui, W. (2004). *GGUM2004 Windows User's Guide*.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. (1999). *Estimating parameters in the generalized graded unfolding model: Sensitivity to the prior distribution assumption and the number of quadrature points used*. National Council n Measurement in Education, Montreal, Quebec, Canada.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*(1), 3–32. https://doi.org/10.1177/01466216000241001

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement*, *26*(2), 192–207. https://doi.org/10.1177/01421602026002006

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item Response Model for Unfolding Responses from a graded disagree-agree response scale. *Applied Psychological Measurement*, *20*(3), 231–255. https://doi.org/10.1177/014662169602000305

Roberts, J. S., & Thompson, V. M. (2011). Marginal maximum a posteriori item parameter estimation for the generalized graded unfolding model. *Applied Psychological Measurement*, *35*(4), 259–279. https://doi.org/10.1177/0146621610392565

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement*, *20*(3), 207–219. JSTOR.

Rupp, A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*(1), 3–38.

Santos, S., Julian, C., Bortolotti, S., Andrade, D., Slater, B., Assis, M. A., Kafatos, A., Henauw, S., Gottrand, F., Androutsos, O., Kersting, M., Sjostrom, M., Forsner, M., & Moreno, L. (2021). A new measure of health motivation influencing food choices and its association with food intakes and nutritional biomarkers in european adolescents. *Public Health Nutrition*, *24*, 1–11. https://doi.org/10.1017/S1368980019004658

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Şengül Avşar, A. (2021). Aberrant individuals' effects on fit indices both of confirmatory factor analysis and polytomous irt models. *Current Psychology*. https://doi.org/10.1007/s12144-021-01563-4

Seo, D. G., & Weiss, D. J. (2013). Lz person-fit index to identify misfit students with achievement test data. *Educational and Psychological Measurement*, *73*(6), 994–1016. https://doi.org/10.1177/0013164413497015

Sgammato, A. N. (2009). *An application of unfolding and cumulative item response theory models for noncognitive scaling: Examining the assumptions and applicability of the generalized graded unfolding model* [Ph.D., The University of North Carolina at Chapel Hill]. https://search.proquest.com/docview/304958643/abstract/53911F8A59904130PQ/1

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitative Methoden: Nieuwsbrief Voor Toegepaste Statistiek En Operationele Research*, *7*(22), 131–145.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*(2), 149–157. https://doi.org/10.1177/014662169201600204

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*(2), 191–207. https://doi.org/10.1007/BF02294835

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. SAGE.

Sinharay, S. (2016a). Asymptotic corrections of standardized extended caution indices. *Applied Psychological Measurement, 40*(6), 418–433. https://doi.org/10.1177/0146621616649963

Sinharay, S. (2016b). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika, 81*(4), 992–1013. https://doi.org/10.1007/s11336-015-9465-x

Sinharay, S. (2017). Are the nonparametric person-fit statistics more powerful than their parametric counterparts? Revisiting the simulations in karabatsos (2003). *Applied Measurement in Education, 30*(4), 314–328. https://doi.org/10.1080/08957347.2017.1353990

Sinharay, S. (2021). Latent-variable approaches utilizing both item scores and response times to detect test fraud. *Open Education Studies, 3*(1), 1–16. https://doi.org/10.1515/edu-2020-0137

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*(3), 331–342. https://doi.org/10.1007/BF02294437

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*(1), 25–39. https://doi.org/10.1037/0021-9010.91.1.25

Stening, B. W., & Everett, J. E. (1984). Response styles in a cross-cultural managerial study. *The Journal of Social Psychology, 122*(2), 151–156. https://doi.org/10.1080/00224545.1984.9713475

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2009). A Monte Carlo study of the effect of item characteristic curve estimation on the accuracy of three person-fit statistics. *Applied Psychological Measurement, 33*(4), 307–324. https://doi.org/10.1177/0146621608329503

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement*, *35*(6), 419–432. https://doi.org/10.1177/0146621610391777

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617. https://doi.org/10.1007/BF02294821

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*(1), 95–110. https://doi.org/10.1007/BF02294208

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting irt models to dichotomous and polytomous data: Assessing the relative model–data fit of ideal point and dominance models. *Applied Psychological Measurement*, *35*(4), 280–295. https://doi.org/10.1177/0146621610390674

Tay, L., & Drasgow, F. (2012). Theoretical, statistical, and substantive issues in the assessment of construct dimensionality: Accounting for the item response process. *Organizational Research Methods*, *15*(3), 363–384. https://doi.org/10.1177/1094428112439709

Tay, L., & Ng, V. (2018). Ideal point modeling of non-cognitive constructs: Review and recommendations for research. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02423

Tendeiro, J. N. (2017). The lz(p)* person-fit statistic in an unfolding model context. *Applied Psychological Measurement*, *41*(1), 44–59. https://doi.org/10.1177/0146621616669336

Tendeiro, J. N., & Castro-Alvarez, S. (2019). GGUM: An R package for fitting the generalized graded unfolding model. *Applied Psychological Measurement*, *43*(2), 172–173. https://doi.org/10.1177/0146621618772290

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, *51*(3), 239–259. https://doi.org/10.1111/jedm.12046

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529–554. JSTOR.

Torre, J. D. L., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, *45*(2), 159–177. https://doi.org/10.1111/j.1745-3984.2008.00058.x

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*(3), 299–314.

Turner, M. L. (2018). *The detection and impact of low cognitive effort survey responses* [Ph.D., University of Colorado at Boulder]. https://search.proquest.com/docview/2161217862/abstract/E996E275A600402CPQ/1

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19. https://doi.org/doi: 10.18637/jss.v020.i11

Van der Flier, H. (1980). Vergelijkbaarheid van individuele testprestatie [Comparability of individual test performance]. *Lisse, The Netherlands: Swets & Zeitlinger*.

Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross Cultural Psychology*, *13*(3), 267–298.

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: evidence of method bias in data from six EU Countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346–360. https://doi.org/10.1177/0022022104264126

van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts, and what model to use instead. *Applied Psychological Measurement*, *18*(2), 97–110. https://doi.org/10.1177/014662169401800201

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a Dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment*, *24*(1), 65–77. https://doi.org/10.1027/1015-5759.24.1.65

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Weiss, B., Crowe, M., Carter, N., Lynam, D., Watts, A., Lilienfeld, S., Skeem, J., & Miller, J. (2018). *Examining hypothesized curvilinear and interactive relations between psychopathic traits and externalizing problems in an offender sample using item response-based analysis*. https://doi.org/10.31234/osf.io/tqdrp

Wilgus, S., & Travis, J. (2019). *A comparison of ideal-point and dominance response processes with a Trust In Science Thurstone scale* (pp. 415–428). https://doi.org/10.1007/978-3-030-01310-3_36

Williams, E. (2015). *Dimensionality assessment of proximity-based data in unfolding model applications*. Georgia Institute of Technology.

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples.

Social Psychological and Personality Science, *8*(4), 454–464. https://doi.org/10.1177/1948550617703168

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. https://doi.org/10.1007/s10862-005-9004-7

Xia, Y., & Zheng, Y. (2018). Asymptotically normally distributed person fit indices for detecting spuriously high scores on difficult items. *Applied Psychological Measurement*, *42*(5), 343–358. https://doi.org/10.1177/0146621617730391

Zampetakis, L. A. (2010). Unfolding the measurement of the creative personality. *The Journal of Creative Behavior*, *44*(2), 105–123. https://doi.org/10.1002/j.2162-6057.2010.tb01328.x

Zampetakis, L. A., Lerakis, M., Kafetsios, K., & Moustakis, V. (2015). Using item response theory to investigate the structure of anticipated affect: Do self-reports about future affective reactions conform to typical or maximal models? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01438

Zhang, B., & Walker, C. M. (2008). Impact of missing data on person—Model fit and person trait estimation. *Applied Psychological Measurement*, *32*(6), 466–479. https://doi.org/10.1177/0146621607307692

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

Zwick, W., & Velicer, W. (1984). *A comparison of five rules for determining the number of components in data sets*. Annual Meeting of the American Psychological Association, Toronto, Ontario. https://files.eric.ed.gov/fulltext/ED251510.pdf

APPENDICES

Appendix A

R Code for Condition of GGUM Fit to GGUM Data

```
#Much of the below code is inspired by and adapted from Tendeiro's work on OSF:
https://www.jorgetendeiro.com/publication/tendeiro_2017/


# 0. Prepare environment ----
rm(list=ls())
if (!is.null(dev.list())) dev.off(dev.list()["RStudioGD"])
library(Rcpp)
library(psych)
library(fastGHQuad)
library(abind)
#Sys.setenv(JAVA_HOME='C:/Users/jreimers/Documents/Jreimers/jdk1.8.0_232')
#install.packages("rJava", type="binary")
#install.packages("GGUM", type="binary")
library(rJava)
library(GGUM)
#Parallel Analysis
library(paran)
#install.packages("Hmisc",dependecies=T)
library(Hmisc)
#PF
#install.packages("PerFit", type="binary")
#install.packages("RCurl", type="binary")
library(RCurl)
library(PerFit)
library(mokken)
# Parallel processing:
library(doParallel)
library(foreach)

#1. Conditions and fixed parameters
n_items.vec   <- c( 20, 40)
AbI.vec <- c(.20, .40, .60)
AbN.vec <- c(.04, .10, .20)
AbType.vec   <- c(  "Random_Responders", "Longstringers",   "ERS", "MRS", "Mixed")
parameters          <- expand.grid(n_items.vec, AbI.vec, AbN.vec, AbType.vec)
colnames(parameters) <- c("n_items.vec", "AbI.vec", "AbN.vec", "AbType.vec")
rm(n_items.vec, AbI.vec, AbN.vec, AbType.vec)
# Fixed parameters:
n       <- 1000
```

```
N       <- 1000
cats    <- 6
#detectCores()


#  Setup parallel backend to use 39 processors (cores):
cl <- makeCluster(39, setup_strategy = "sequential")
registerDoParallel(cl, cores = 39)
# END SECTION

# 3. Run the simulation ----

#
start.time <- Sys.time()
print(start.time)

outcome.simulation <- foreach(i=1:90) %:% # i=27 rep=1 c(17, 18, 35, 6, 24) i=1 rep=1
  foreach(rep=1:100, .packages=c("psych", "fastGHQuad", "abind", "GGUM", "mokken",
"PerFit", "paran", "Hmisc")) %dopar% {
    set.seed(1000*i+rep)
    # Specify varying parameters for cell i:i=1
    n_items    <- parameters[i, 1]
    AbI        <- parameters[i, 2]
    AbN        <- parameters[i, 3]
    AbType     <- parameters[i, 4]
    n_Abitems<-(n_items*AbI)
    n_Cleanitems<-n_items-n_Abitems




######START: STEP 1 Generate CLEAN Data & Estimate Parameters ###############
##############################################################################
##############################################################################
    # Generate item scores according to the GGUM:
    gendata <- GenData.GGUM(n, n_items, cats-1, seed=(1000*i+rep))
    data.ext<-gendata[[5]]
    write.table(data.ext, file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/1.GGUMGGUM_clean/GGUM_condition_",
i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)

    # Estimate item and person parameters BEFORE aberrant behaviour:
    IP.est     <- GGUM(data.ext, cats-1)

    ##correct for negative delta and thus opposite signed thetas from true
    for (item in 1:n_items) {
```

```
    if ((gendata[[2]][item]<0 && IP.est$delta[item]>0)||(gendata[[2]][item]>0 &&
IP.est$delta[item]<0)){
      IP.est$delta[item]<-IP.est$delta[item]*(-1)
    }
  }

  Th.est     <- GGUM::Theta.EAP(IP.est)
  Th.est.ext  <- as.vector(Th.est [,2])



  # Compare generated and estimated parameters, save results:
  MAD.alpha  <- round(sum(abs(IP.est$alpha - gendata[[1]])) / n_items, 4)
  BIAS.alpha <- round(sum(    IP.est$alpha - gendata[[1]])  / n_items, 4)
  cor.alpha  <- round(cor(    IP.est$alpha, gendata[[1]])      , 4)
  MAD.delta  <- round(sum(abs(IP.est$delta - gendata[[2]])) / n_items, 4)
  BIAS.delta <- round(sum(    IP.est$delta - gendata[[2]])  / n_items, 4)
  cor.delta  <- round(cor(    IP.est$delta, gendata[[2]])      , 4)
  MAD.taus   <- round(sum(abs(IP.est$taus[, 1:cats-1] - gendata[[3]][, 1:cats-1])) / (n_items *
cats-1), 4)
  BIAS.taus  <- round(sum(    IP.est$taus[, 1:cats-1] - gendata[[3]][, 1:cats-1]) / (n_items * cats-
1), 4)
  cor.taus   <- round(cor(  c(IP.est$taus[, 1:cats-1]), c(gendata[[3]][, 1:5])), 4)
  MAD.th     <- round(sum(abs(Th.est.ext  - gendata[[4]]), na.rm=TRUE) / n, 4)
  BIAS.th    <- round(sum(    Th.est.ext  - gendata[[4]], na.rm=TRUE) / n, 4)
  cor.th     <- round(cor(    Th.est.ext , gendata[[4]], use = 'complete.obs'), 4)

 ######END: STEP 1 Generate CLEAN Data & Estimate Parameters ###############
################################################################################
################################################################################



 ###### START: STEP 2 Create ABERRANT data & Estimate Parameters ###############
################################################################################
################################################################################


 ########################Generate aberrant datasets#############################
 ############(1 & 2) ERS and MRS############
  if (AbType == "MRS") {
    N.middle  <- which(rowSums((gendata[[5]][, 1:n_items] >= 1) * (gendata[[5]][, 1:n_items]
<=  4)) >= ceiling(AbI * n_items))
    N.extreme <- which(rowSums((gendata[[5]][, 1:n_items] == 0|gendata[[5]][, 1:n_items] ==1)
+ (gendata[[5]][, 1:n_items] == 4|gendata[[5]][, 1:n_items] ==5)) >= ceiling(AbI * n_items))
```

```
    if (length(N.extreme) > (n * AbN)) #if n_ppl with sufficient extreme scores is greater tha
n*AbN then sample n*AbN ppl from them...otherwise just use as many as available.
    {subs.middle <- sort(sample(N.extreme, (n * AbN), replace = FALSE))} else {
     subs.middle <- N.extreme}

    scrs.middle  <- gendata[[5]][subs.middle,  1:n_items, drop = FALSE]
    scrs.middle  <- t(apply(scrs.middle, 1, function(vec)
    {
     extreme.scrs <- which((vec == 0|vec == 1|vec == 4|vec ==5)  == 1) #Take the extreme
scores and below will sample from them to eventually replace with middle scores to mimic MRS
     extreme.scrs <- sort(sample(extreme.scrs, ceiling(n_items * AbI), replace = FALSE))
     if (length(extreme.scrs) > 0) {vec[extreme.scrs] <- sapply(vec[extreme.scrs], function(x)
     {
      if (x == 0) {2}
      else if (x == 1) {2}
      else 3})
     }
     vec
    }))

    data.aberrant<-gendata[[5]]
    data.aberrant[subs.middle, 1:n_items]  <- scrs.middle
    subs <- sort(c(subs.middle))
   }


   if (AbType == "ERS") {
    N.middle  <- which(rowSums((gendata[[5]][, 1:n_items] >= 1) * (gendata[[5]][, 1:n_items]
<=  4)) >= ceiling(AbI * n_items))
    N.extreme <- which(rowSums((gendata[[5]][, 1:n_items] == 0|gendata[[5]][, 1:n_items] ==1)
+ (gendata[[5]][, 1:n_items] == 4|gendata[[5]][, 1:n_items] ==5)) >= ceiling(AbI * n_items))

    if (length(N.middle) > (n * AbN)) #if n_ppl with sufficient middle scores is greater tha
n*AbN then sample n*AbN ppl from them...otherwise just use as many as available.
    {subs.extreme <- sort(sample(N.middle, (n * AbN), replace = FALSE))} else {
     subs.extreme <- N.middle}


    scrs.extreme <- gendata[[5]][subs.extreme, 1:n_items, drop = FALSE] #as of now this is just
the clean scores generated for the random sample of to-be aberrant responders
    scrs.extreme <- t(apply(scrs.extreme, 1, function(vec) #apply the function to scrs.extreme
rows
    {
     middle.scrs <- which(((vec >= 1) * (vec <= 4)) == 1) #take the middle scores (1,2,3,4) and
below will sample n_items*AbI from them
```

```
      middle.scrs <- sort(sample(middle.scrs, ceiling(n_items * AbI), replace = FALSE))
      if (length(middle.scrs) > 0) {vec[middle.scrs] <- sapply(vec[middle.scrs], function(x)
      {
       if (x == 1) {0}
       else if (x == 2) {0}
       else 5})
      }
      vec
     }))

     data.aberrant<-gendata[[5]]
     data.aberrant[subs.extreme, 1:n_items] <- scrs.extreme
     subs <- sort(c(subs.extreme))
    }


    #############(3)Longstringers#############

    #JR# The following is a redo of longstring simulation where value and initial index were first
found for
    #JR#  each ab respondee and then repeated so that string was consecutive
    if (AbType == "Longstringers") {
     data.aberrant<-as.data.frame(gendata[[5]])
     subs.longstring <- sort(sample(1:1000, (n * AbN), replace = FALSE))

     #Find Longstring Value
     for (p in 1:1000){
      longstringvalue<-sample(0:(cats-1),1)
      data.aberrant$Longstringvalue[p] <-longstringvalue
     }
     #Find initial position to start longstring
     data.aberrant<-as.data.frame(data.aberrant)
     for (p in 1:1000){
      initial.position<-sample(1:(n_items-(AbI*n_items)+1),1)
      data.aberrant$Longstringinitial[p] <-initial.position
     }

     scrs.longstring <- data.aberrant[subs.longstring, 1:(n_items+2), drop = FALSE] #as of now
this is just the clean scores generated for the random sample of to-be aberrant responders

     for (p in 1:nrow(scrs.longstring)){

scrs.longstring[p,(scrs.longstring$Longstringinitial[p]:(scrs.longstring$Longstringinitial[p]+(AbI
*n_items)-1))] <- scrs.longstring$Longstringvalue[p]
     }
```

```
    #remove last two columns specifying Longstring value and initial position
    scrs.longstring<-as.matrix(scrs.longstring[, 1:n_items])

    data.aberrant<-gendata[[5]]
    data.aberrant[subs.longstring, 1:n_items] <- scrs.longstring
    subs <- sort(c(subs.longstring))
   }

   #############(4)Random Responders############

   if (AbType == "Random_Responders") {
    data.aberrant<-gendata[[5]]
    subs.random <- sort(sample(1:1000, (n * AbN), replace = FALSE))

    scrs.random <- gendata[[5]][subs.random, 1:n_items, drop = FALSE] #as of now this is just
the clean scores generated for the random sample of to-be aberrant responders
    scrs.random <- t(apply(scrs.random, 1, function(vec)
    {
     random.scrs      <- sort(sample(1:n_items, ceiling(n_items * AbI), replace = FALSE))
     vec[random.scrs] <- sample(0:(cats-1), ceiling(n_items * AbI), replace = TRUE)
     vec
    }))

    data.aberrant[subs.random, 1:n_items] <- scrs.random
    subs <- sort(c(subs.random))
   }


   #############(5) Mixed - includes all 4 types of aberrant responders############
AbType<-"Mixed"
   if (AbType == "Mixed") {
    #Index of respondents who have enough middle and extreme responses to be changed for the
condition of MRS and ERS
    N.middle  <- which(rowSums((gendata[[5]][, 1:n_items] >= 1) * (gendata[[5]][, 1:n_items]
<=  4)) >= ceiling(AbI * n_items))
    N.extreme <- which(rowSums((gendata[[5]][, 1:n_items] == 0|gendata[[5]][, 1:n_items] ==1)
+ (gendata[[5]][, 1:n_items] == 4|gendata[[5]][, 1:n_items] ==5)) >= ceiling(AbI * n_items))

    #First sample from simulees for each AbType and make sure they do not overlap


    if (length(N.middle) <= length(N.extreme))
    {
     if (length(N.middle) > ((n * AbN) / 4))
     {subs.extreme  <- sort(sample(N.middle, (n * AbN) / 4, replace = FALSE))} else {
      subs.extreme  <- N.middle}
```

```
      if (length(setdiff(N.extreme, subs.extreme)) > ((n * AbN) / 4))
      {subs.middle <- sort(sample(setdiff(N.extreme, subs.extreme), (n * AbN) / 4, replace =
FALSE))} else {
        subs.middle <- setdiff(N.extreme, subs.extreme)}
    }
    if (length(N.extreme) < length(N.middle))
    {
      if (length(N.extreme) > ((n * AbN) / 4))
      {subs.middle  <- sort(sample(N.extreme, (n * AbN) / 4, replace = FALSE))} else {
        subs.middle  <- N.extreme}
      if (length(setdiff(N.middle, subs.middle)) > ((n * AbN) / 4))
      {subs.extreme <- sort(sample(setdiff(N.middle, subs.middle), (n * AbN) / 4, replace =
FALSE))} else {
        subs.extreme <- setdiff(N.middle, subs.middle)}
    }

    subs.random <-  sort(sample(setdiff(1:n, c(subs.middle, subs.extreme)), (n * AbN) / 4,
replace = FALSE))
    subs.longstring <- sort(sample(setdiff(1:n, c(subs.middle, subs.extreme, subs.random)), (n *
AbN) / 4, replace = FALSE))

    #Once all subs.AbType have been created, substitute their response data accordinginly
    scrs.extreme <- gendata[[5]][subs.extreme, 1:n_items, drop = FALSE]
    scrs.extreme <- t(apply(scrs.extreme, 1, function(vec)
    {
      middle.scrs <- which(((vec >= 1) * (vec <= 4)) == 1)
      middle.scrs <- sort(sample(middle.scrs, ceiling(n_items * AbI), replace = FALSE))
      if (length(middle.scrs) > 0) {vec[middle.scrs] <- sapply(vec[middle.scrs], function(x)
      {
        if (x == 1) {0}
        else if (x == 2) {0}
        else 5})
      }
      vec
    }))
    #
    scrs.middle  <- gendata[[5]][subs.middle,  1:n_items, drop = FALSE]
    scrs.middle  <- t(apply(scrs.middle, 1, function(vec)
    {
      extreme.scrs <- which((vec == 0|vec == 1|vec == 4|vec ==5)  == 1)
      extreme.scrs <- sort(sample(extreme.scrs, ceiling(n_items * AbI), replace = FALSE))
      if (length(extreme.scrs) > 0) {vec[extreme.scrs] <- sapply(vec[extreme.scrs], function(x)
      { #if x is a 0 or 1, replace with middle option of 2. If x is a 4 or 5, replace with middle
option of 3.
        if (x == 0) {2}
        else if (x == 1) {2}
```

```
    else 3})
  }
  vec
}))



#scores for longstringers
data.aberrant<-as.data.frame(gendata[[5]])
#Find Longstring Value
for (p in 1:1000){
  longstringvalue<-sample(0:(cats-1),1)
  data.aberrant$Longstringvalue[p] <-longstringvalue
}
#Find initial position to start longstring
data.aberrant<-as.data.frame(data.aberrant)
for (p in 1:1000){
  initial.position<-sample(1:(n_items-(AbI*n_items)+1),1)
  data.aberrant$Longstringinitial[p] <-initial.position
}

scrs.longstring <- data.aberrant[subs.longstring, 1:(n_items+2), drop = FALSE] #as of now
this is just the clean scores generated for the random sample of to-be aberrant responders

for (p in 1:nrow(scrs.longstring)){

scrs.longstring[p,(scrs.longstring$Longstringinitial[p]:(scrs.longstring$Longstringinitial[p]+(AbI
*n_items)-1))] <- scrs.longstring$Longstringvalue[p]
}

#remove last two columns specifying Longstring value and initial position
scrs.longstring<-as.matrix(scrs.longstring[, 1:n_items])



#scores for random responders
scrs.random <- gendata[[5]][subs.random, 1:n_items, drop = FALSE] #as of now this is just
the clean scores generated for the random sample of to-be aberrant responders
scrs.random <- t(apply(scrs.random, 1, function(vec)
{
  random.scrs      <- sort(sample(1:n_items, ceiling(n_items * AbI), replace = FALSE))
  vec[random.scrs] <- sample(0:(cats-1), ceiling(n_items * AbI), replace = TRUE)
  vec
}))
```

```
    #substitute subs scores with ab subs scores from above
    data.aberrant                <- gendata[[5]]
    data.aberrant[subs.middle, ]     <- scrs.middle
    data.aberrant[subs.extreme, ]    <- scrs.extreme
    data.aberrant[subs.random, ]     <- scrs.random
    data.aberrant[subs.longstring, ] <- scrs.longstring


    subs <- sort(c(subs.extreme, subs.middle, subs.longstring, subs.random))
    subs.tbl <- matrix(c(subs.extreme, subs.middle, subs.longstring, subs.random),
nrow=(AbN*n/4))
    colnames(subs.tbl)=c("subs.extreme", "subs.middle", "subs.longstring", "subs.random")
    write.table(subs.tbl, file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/Subs_Data_Folder/GGUMab_mixedsubs_condition_",
i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)

  }



  write.table(data.aberrant, file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/1.GGUMGGUM_ab/GGUMab_condition_",
i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)
  write.table(subs, file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/Subs_Data_Folder/GGUMab_subs_condition_",
i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)

#########################################################################
 ##now that the ab dataset is created, next fit the model to the data####################

  # Estimate item and person parameters (aberrant):
  IP.est.aber     <- GGUM(data.aberrant, cats-1)

  for (item in 1:n_items) {
    if ((gendata[[2]][item]<0 && IP.est.aber$delta[item]>0)||(gendata[[2]][item]>0 &&
IP.est.aber$delta[item]<0)){
      IP.est.aber$delta[item]<-IP.est.aber$delta[item]*(-1)
    }
  }

  Th.est.aber    <- GGUM::Theta.EAP(IP.est.aber)
  Th.est.ext.aber <- as.vector(Th.est.aber[,2])


  #export IPs for both clean and aberrant datasets
```

```
   write.table(IP.est.aber[["alpha"]], file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/1.GGUMGGUM_ab_IP/GGUMab_IP.alpha_conditio
n_", i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)
   write.table(IP.est[["alpha"]], file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/1.GGUMGGUM_clean_IP/GGUM_IP.alpha_conditio
n_", i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)
   write.table(IP.est.aber[["delta"]], file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/1.GGUMGGUM_ab_IP/GGUMab_IP.delta_condition
_", i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)
   write.table(IP.est[["delta"]], file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/1.GGUMGGUM_clean_IP/GGUM_IP.delta_conditio
n_", i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)
   write.table(IP.est.aber[["taus"]], file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/1.GGUMGGUM_ab_IP/GGUMab_IP.taus_condition
_", i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)
   write.table(IP.est[["taus"]], file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/1.GGUMGGUM_clean_IP/GGUM_IP.taus_condition
_", i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)
   write.table(Th.est.aber, file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/1.GGUMGGUM_ab_IP/GGUMab_IP.theta_condition
_", i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)
   write.table(Th.est, file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/Datasets/1.GGUMGGUM_clean_IP/GGUM_IP.theta_conditio
n_", i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)


   # Compare generated and estimated parameters, save results:
   MAD.alpha.aber  <- round(sum(abs(IP.est.aber$alpha - gendata[[1]])) / n_items, 4)
   BIAS.alpha.aber <- round(sum(    IP.est.aber$alpha - gendata[[1]])  / n_items, 4)
   cor.alpha.aber  <- round(cor(    IP.est.aber$alpha, gendata[[1]])      , 4)
   MAD.delta.aber  <- round(sum(abs(IP.est.aber$delta - gendata[[2]])) / n_items, 4)
   BIAS.delta.aber <- round(sum(    IP.est.aber$delta - gendata[[2]])  / n_items, 4)
   cor.delta.aber  <- round(cor(    IP.est.aber$delta, gendata[[2]])      , 4)
   MAD.taus.aber   <- round(sum(abs(IP.est.aber$taus[, 1:cats-1] - gendata[[3]][, 1:cats-1])) /
(n_items * cats-1), 4)
   BIAS.taus.aber  <- round(sum(    IP.est.aber$taus[, 1:cats-1] - gendata[[3]][, 1:cats-1]) /
(n_items * cats-1), 4)
   cor.taus.aber   <- round(cor(  c(IP.est.aber$taus[, 1:cats-1]), c(gendata[[3]][, 1:cats-1])), 4)
   MAD.th.aber     <- round(sum(abs(Th.est.ext.aber - gendata[[4]]), na.rm=TRUE) / n, 4)
   BIAS.th.aber    <- round(sum(    Th.est.ext.aber - gendata[[4]], na.rm=TRUE)  / n, 4)
   cor.th.aber     <- round(cor(    Th.est.ext.aber, gendata[[4]], use = 'complete.obs'), 4)
   #
   MAD.th.aber.fit     <- round(sum(abs(Th.est.ext.aber[-subs] - gendata[[4]][-subs]),
na.rm=TRUE) / length((1:n)[-subs]), 4)
   BIAS.th.aber.fit    <- round(sum(    Th.est.ext.aber[-subs] - gendata[[4]][-subs], na.rm=TRUE)
/ length((1:n)[-subs]), 4)
```

```
   cor.th.aber.fit     <- round(cor(    Th.est.ext.aber[-subs], gendata[[4]][-subs], use =
'complete.obs'), 4)
   MAD.th.aber.misfit  <- round(sum(abs(Th.est.ext.aber[subs] - gendata[[4]][subs]),
na.rm=TRUE) / length((1:n)[subs]), 4)
   BIAS.th.aber.misfit <- round(sum(    Th.est.ext.aber[subs] - gendata[[4]][subs], na.rm=TRUE)
/ length((1:n)[subs]), 4)
   cor.th.aber.misfit  <- round(cor(    Th.est.ext.aber[subs], gendata[[4]][subs], use =
'complete.obs'), 4)


##############################################################################
##############################################################################
###############END: STEP 2 Create ABERRANT data & Estimate Parameters
##############################################################################



##############################################################################
##############################################################################
################START: STEP 3 MODEL FIT & Reliability ########################


   ################## Model Fit, Item Fit, and Dimensionality###################

   # Compute chisq/df ratios for single, pairs, and triples of items (Drasgow et al., 1995):
   MODFIT.res <- MODFIT(IP.est)
   MODFIT.res.aber <- MODFIT(IP.est.aber)

   perc.item.flagged<-rowSums(MODFIT.res[[4]][, 4:7]) / rowSums(MODFIT.res[[4]][, 1:7])
   perc.item.flagged.ab<-rowSums(MODFIT.res.aber[[4]][, 4:7]) /
rowSums(MODFIT.res.aber[[4]][, 1:7])

   Mean.SD.chsqr<-MODFIT.res[[4]][, 8:9]

   Mean.SD.chsqr.ab<-MODFIT.res.aber[[4]][, 8:9]


   #Likert-scaled data processing as done in Tay, et. al, 2011 & Tay & Drasgow, 2012

  #Clean data
   #First Reverse code lower 30% items
   #Note: items are ordered by delta either inc or dec (arbitrary), thus use if then statement to
determine which end the negative deltas are
   Items.Neg.Deltas<-which(gendata[["delta.gen"]]<0)
   Nitems.revcode<- .3*n_items
```

```
  if (IP.est$delta[1]<0){
    Items.revcode<-Items.Neg.Deltas[1:Nitems.revcode]
  } else if (IP.est$delta[1]>0){
    Items.revcode<-(n_items-Nitems.revcode+1):n_items}
  data.ext.reverse1<-data.ext
  data.ext.reverse1[,Items.revcode]<- (cats-1)-data.ext[,Items.revcode]
  #### Next, item-total biserial correlation
  #### and further reverse scoreing for items with negative item-total correlations
  interitemstats<-psych::alpha(data.ext.reverse1, check.keys = TRUE)
  keys<-unlist(interitemstats["keys"])
  data.ext.reverse2<-reverse.code(keys, data.ext.reverse1)
  item_array<- paste0("item", 1:n_items)
  colnames(data.ext.reverse2)<-item_array
  #Parallel Analysis for clean data
  parallel<-paran(data.ext.reverse2, cfa=FALSE)
  factors_retained.paran<-parallel[[1]]
  #eigs for clean data
  data.cor<-cor(data.ext.reverse2, use="complete.obs")
  eig<-eigen(data.cor)
  eig<- eig$values



  #ab data
  #First Reverse code lower 30% items
  #Note: items are ordered by delta either inc or dec (arbitrary), thus use if then statement to
determine which end the negative deltas are
  Items.Neg.Deltas<-which(IP.est.aber$delta<0)
  Nitems.revcode<- .3*n_items
  if (IP.est.aber$delta[1]<0){
    Items.revcode<-Items.Neg.Deltas[1:Nitems.revcode]
  } else if (IP.est.aber$delta[1]>0){
    Items.revcode<-(n_items-Nitems.revcode+1):n_items}
  data.ab.reverse1<-data.aberrant
  data.ab.reverse1[,Items.revcode]<- (cats-1)-data.aberrant[,Items.revcode]
  #### Next, item-total biserial correlation
  #### and further reverse scoreing for items with negative item-total correlations
  interitemstats<-psych::alpha(data.ab.reverse1, check.keys = TRUE)
  keys<-unlist(interitemstats["keys"])
  data.ab.reverse2<-data.ab.reverse1
  data.ab.reverse2<-reverse.code(keys, data.ab.reverse2)
  item_array<- paste0("item", 1:n_items)
  colnames(data.ab.reverse2)<-item_array
  #Parallel Analysis for ab data
  parallel.ab<-paran(data.ab.reverse2, cfa=FALSE)
  factors_retained.paran.ab<-parallel.ab[[1]]
```

```
    #eigs for ab data
    data.cor.ab<-cor(data.ab.reverse2, use="complete.obs")
    eig.ab<-eigen(data.cor.ab)
    eig.ab<- eig.ab$values



    #Information Criterion
    AIC<-IP.est[[12]][3]
    BIC<-IP.est[[12]][4]

    AIC.aber<-IP.est.aber[[12]][3]
    BIC.aber<-IP.est.aber[[12]][4]



    ##Prep dataset for IIO and HT analyses
    #remove invariant response strings from clean data
    find.invariants <- apply(data.ext.reverse2, 1, function(vec) max(vec) - min(vec))
    pos.invariants  <- which(find.invariants == 0)
    H_data          <- if (length(pos.invariants) > 0) data.ext.reverse2[-pos.invariants,] + 1 else
data.ext.reverse2 + 1 #JT# Not sure why we need the '+1'! Can you tell me?...

    #remove invariant response strings from ab data
    find.invariants <- apply(data.ab.reverse2, 1, function(vec) max(vec) - min(vec))
    pos.invariants  <- which(find.invariants == 0)
    H_data.ab          <- if (length(pos.invariants) > 0) data.ab.reverse2[-pos.invariants,] + 1 else
data.ab.reverse2 + 1 #JT# Not sure why we need the '+1'! Can you tell me?...

    ###Check item ordering with coef h
    H.list<-coefH(H_data)
    H<-as.numeric(H.list$H[[1]])
    H_SE<-gsub("\\(", "", (H.list$H[[2]]))
    H_SE<-as.numeric(gsub("\\)", "", (H_SE)))

    H.list.aber<-coefH(H_data.ab)
    H.aber<-as.numeric(H.list.aber$H[[1]])
    H_SE.aber<-gsub("\\(", "", (H.list.aber$H[[2]]))
    H_SE.aber<-as.numeric(gsub("\\)", "", (H_SE.aber)))

    # Reliability using Molenaar Sijtsma MS statistic.
    reliab.MS.aber<-check.reliability(data.ab.reverse2)[[1]]
    reliab.MS<-check.reliability(data.ext.reverse2)[[1]]
    reliab.alpha.aber<-check.reliability(data.ab.reverse2)[[2]]
    reliab.alpha<-check.reliability(data.ext.reverse2)[[2]]
    reliab.lambda2.aber<-check.reliability(data.ab.reverse2)[[3]]
```

```
reliab.lambda2<-check.reliability(data.ext.reverse2)[[3]]



################################################################################
################################################################################
##############################END: STEP 3 MODEL FIT & Reliability ################

################################################################################
########################START: STEP 4 OBTAINING CUTOFFS
################################################################################

   #Step 1: Generate model-fitting item score vectors based on GGUM using the person and item parameters estimated from the aberrant data.
   #Step 2: Estimate item parameters for modelfitting data from step 1
   #Step 3: Reverse code items with negative deltas (Likert scaling technique; Tay et al. 2011) for use of nonparametric PFS (not neccessary for parametric lz and lzstar); We thought this is what a researcher would do in realistic setting when data is ideal point and trying to use nonparametric pfs
   #Step 4: Compute PFSs for modelfitting data and obtain the upper or lower 5% value as cutoff depending on the PFS
   #Step 5: Run steps 1-4 in loop for multiple replications and take the median cutoff value to be used **Will only be able to run 20 replications due to time (see step 2 ~ 20minutes)


   ##PART A: Steps 1-5 using Parm ests from aberrant data


   # Generate data based on estimated model parameters (aber):
   #install.packages("Hmisc")
   ##Step 1: Generate model-fitting item score vectors based on GGUM using the person and item parameters estimated from the aberrant data.
   ##Start loop to generate data, compute PFSs, and take median to get the cutoff to be used

   Nreps=20
   N=1000
   Th.est.aber.NA <- which(is.na(Th.est.ext.aber))
   PFS.cutoffs <- matrix(NA, Nreps, 6)
   for (r in 1:Nreps){ #r=1
     set.seed(1000*i+1000*rep+r)
     I<-n_items
     probs.array <- array(NA, dim = c(N, I, cats))
     C<-cats-1
     for (z in 0:C)
     {
```

```
    probs.array[, , z + 1] <- GGUM:::P.GGUM(z, IP.est.aber$alpha, IP.est.aber$delta,
IP.est.aber$taus,Th.est.ext.aber,  C)
  }


  if  (length(Th.est.aber.NA) >0 ){
    modelfitting.data <- apply(probs.array[-Th.est.aber.NA, , ], 1:2, function(vec) which(
rmultinom(1, 1, vec) == 1) - 1) #JT# 997x20, so the three NA simulees were removed.
  } else{
    modelfitting.data <- apply(probs.array, 1:2, function(vec) which( rmultinom(1, 1, vec) ==
1) - 1) #JT# 997x20, so the three NA simulees were removed.
  }
  IP.est.modelfitting    <- GGUM(modelfitting.data, cats-1) #Step 2


  Th.est.gendata     <- GGUM:::Theta.EAP(IP.est.modelfitting)
  Th.est.ext.gendata <- as.vector(Th.est.gendata[,2])

  #Reverse code before computing PFSs
  #Note: items are ordered by delta either inc or dec (arbitrary), thus use if then statement to
determine which end the negative deltas are
  Items.Neg.Deltas<-which(IP.est.modelfitting$delta<0)
  Nitems.revcode<- .3*n_items
  if (IP.est.modelfitting$delta[1]<0){
    Items.revcode<-Items.Neg.Deltas[1:Nitems.revcode]
  } else if (IP.est.modelfitting$delta[1]>0){
    Items.revcode<-(n_items-Nitems.revcode+1):n_items}
  modelfitting.data.reverse1<-modelfitting.data
  modelfitting.data.reverse1[,Items.revcode]<- (cats-1)-modelfitting.data[,Items.revcode]
  #### Next, item-total biserial correlation
  #### and further reverse scoreing for items with negative item-total correlations
  interitemstats<-psych::alpha(modelfitting.data.reverse1, check.keys = TRUE)
  keys<-unlist(interitemstats["keys"])
  modelfitting.data.reverse2<-modelfitting.data.reverse1
  modelfitting.data.reverse2<-reverse.code(keys, modelfitting.data.reverse2)
  item_array<- paste0("item", 1:n_items)
  colnames(modelfitting.data.reverse2)<-item_array



  #compute PFSs for the generated data for the purpose of creating the cutoffs (median over
replications)

  ##Prep dataset for IIO and HT analyses
  find.invariants <- apply( modelfitting.data.reverse2, 1, function(vec) max(vec) - min(vec))
  pos.invariants  <- which(find.invariants == 0)
```

```
    H_data.cutoff   <- if (length(pos.invariants) > 0)  modelfitting.data.reverse2[-pos.invariants,]
+ 1 else  modelfitting.data.reverse2 + 1


    #compute PFs
    #U3poly
    U3poly.res.list<-U3poly( modelfitting.data.reverse2 , cats)
    U3poly.res.modfit<-U3poly.res.list[[1]]

    #Gpoly
    Gpoly.res.list<-Gpoly( modelfitting.data.reverse2 , cats)
    Gpoly.res.modfit<-Gpoly.res.list[[1]]


    #Gnormed.poly
    Gnormed.poly.res.list<-Gnormed.poly( modelfitting.data.reverse2 , cats)
    Gnormed.poly.res.modfit<-Gnormed.poly.res.list[[1]]
    #HT
    #first ranspose the data matrix so that persons are columns and items are rows
    data.aberrant.t<-as.matrix(t(H_data.cutoff))
    # Using CoefH in Mokken to calculate the Ht for persons instead of items.
    HT.list<-coefH(data.aberrant.t, se=FALSE)
    HT.res.modfit<-as.data.frame(HT.list[[2]])


    #lz
    lzpoly.res.list<-lzpoly.mixed(matrix=as.matrix(modelfitting.data), IP= IP.est.modelfitting
,Ability= Th.est.ext.gendata, C=5)
    lzpoly.res.modfit<- as.data.frame(lzpoly.res.list)

    #lzstar
    lzstar.res.list<-lzstarpoly.mixed(matrix=as.matrix(modelfitting.data), IP=IP.est.modelfitting,
C=5, Ability= Th.est.ext.gendata)
    lzstar.res.modfit<- as.data.frame(lzstar.res.list)



    #Compute cutoff (used median)
    U3poly.modfit.cut       <- round(quantile(U3poly.res.modfit$PFscores, probs = .95), 4)
    Gpoly.modfit.cut        <- round(quantile(Gpoly.res.modfit$PFscores, probs = .95), 4)
    Gnormed.poly.modfit.cut  <- round(quantile(Gnormed.poly.res.modfit$PFscores, probs =
.95), 4)
    HT.modfit.cut           <- round(quantile( HT.res.modfit, probs = .05, na.rm=TRUE), 4)
    lzpoly.modfit.cut       <- round(quantile(lzpoly.res.modfit, probs = .05, na.rm=TRUE), 4)
    lzstar.modfit.cut       <- round(quantile(lzstar.res.modfit, probs = .05, na.rm=TRUE), 4)
```

```
    PFS.cutoffs[r, 1]<-  U3poly.modfit.cut
    PFS.cutoffs[r, 2]<-  Gpoly.modfit.cut
    PFS.cutoffs[r, 3]<-  Gnormed.poly.modfit.cut
    PFS.cutoffs[r, 4]<-  HT.modfit.cut
    PFS.cutoffs[r, 5]<-  lzpoly.modfit.cut
    PFS.cutoffs[r, 6]<-  lzstar.modfit.cut
    names(PFS.cutoffs)<- c("U3poly", "Gpoly", "Gnormed.poly", "HT", "lzpoly", "lzstar")


  }
  set.seed(1000*i+rep)


  cutoff.use.U3poly <- round(median( PFS.cutoffs[, 1]), 4)
  cutoff.SE.U3poly  <- round(sd(PFS.cutoffs[, 1]), 4)

  cutoff.use.Gpoly <- round(median( PFS.cutoffs[, 2]), 4)
  cutoff.SE.Gpoly  <- round(sd(PFS.cutoffs[, 2]), 4)

  cutoff.use.Gnormed.poly <- round(median( PFS.cutoffs[, 3]), 4)
  cutoff.SE.Gnormed.poly  <- round(sd(PFS.cutoffs[, 3]), 4)

  cutoff.use.HT <- round(median( PFS.cutoffs[, 4]), 4)
  cutoff.SE.HT  <- round(sd(PFS.cutoffs[, 4]), 4)

  cutoff.use.lzpoly <- round(median( PFS.cutoffs[, 5]), 4)
  cutoff.SE.lzpoly  <- round(sd(PFS.cutoffs[, 5]), 4)

  cutoff.use.lzstar <- round(median( PFS.cutoffs[, 6]), 4)
  cutoff.SE.lzstar  <- round(sd(PFS.cutoffs[, 6]), 4)


  ##PART B: Using clean data

  #compute PFSs for the generated data for the purpose of creating the cutoffs (median over
replications)

  ##Prep dataset for IIO and HT analyses
  find.invariants <- apply( data.ext.reverse2, 1, function(vec) max(vec) - min(vec))
  pos.invariants  <- which(find.invariants == 0)
  H_data.cutoff   <- if (length(pos.invariants) > 0)  data.ext.reverse2[-pos.invariants,] + 1 else
data.ext.reverse2 + 1


  #compute PFs
```

```
#U3poly
U3poly.res.list<-U3poly( data.ext.reverse2 , cats)
U3poly.res.modfit<-U3poly.res.list[[1]]

#Gpoly
Gpoly.res.list<-Gpoly( data.ext.reverse2 , cats)
Gpoly.res.modfit<-Gpoly.res.list[[1]]


#Gnormed.poly
Gnormed.poly.res.list<-Gnormed.poly( data.ext.reverse2 , cats)
Gnormed.poly.res.modfit<-Gnormed.poly.res.list[[1]]
#HT
#first ranspose the data matrix so that persons are columns and items are rows
data.aberrant.t<-as.matrix(t(H_data.cutoff))
# Using CoefH in Mokken to calculate the Ht for persons instead of items.
HT.list<-coefH(data.aberrant.t, se=FALSE)
HT.res.modfit<-as.data.frame(HT.list[[2]])



#lz
lzpoly.res.list<-lzpoly.mixed(matrix=data.ext, IP= IP.est , Ability= Th.est.ext, C=5)
lzpoly.res.modfit<- as.data.frame(lzpoly.res.list)

#lzstar
lzstar.res.list<-lzstarpoly.mixed(matrix=data.ext, IP=IP.est, C=5, Ability= Th.est.ext)
lzstar.res.modfit<- as.data.frame(lzstar.res.list)



#Compute cutoff (used median)
U3poly.modfit.cut       <- round(quantile(U3poly.res.modfit$PFscores, probs = .95), 4)
Gpoly.modfit.cut        <- round(quantile(Gpoly.res.modfit$PFscores, probs = .95), 4)
Gnormed.poly.modfit.cut  <- round(quantile(Gnormed.poly.res.modfit$PFscores, probs = .95),
4)
HT.modfit.cut           <- round(quantile( HT.res.modfit, probs = .05, na.rm=TRUE), 4)
lzpoly.modfit.cut       <- round(quantile(lzpoly.res.modfit, probs = .05, na.rm=TRUE), 4)
lzstar.modfit.cut       <- round(quantile(lzstar.res.modfit, probs = .05, na.rm=TRUE), 4)

PFS.cutoffs<- data.frame(matrix(ncol = 6, nrow = Nreps))
PFS.cutoffs[1, 1]<-  U3poly.modfit.cut
PFS.cutoffs[1, 2]<-  Gpoly.modfit.cut
PFS.cutoffs[1, 3]<-  Gnormed.poly.modfit.cut
PFS.cutoffs[1, 4]<-  HT.modfit.cut
PFS.cutoffs[1, 5]<-  lzpoly.modfit.cut
```

```
PFS.cutoffs[1, 6]<-  lzstar.modfit.cut
names(PFS.cutoffs)<- c("U3poly", "Gpoly", "Gnormed.poly", "HT", "lzpoly", "lzstar")

cutoff.use.U3poly.clean <- round( PFS.cutoffs[1, 1], 4)

cutoff.use.Gpoly.clean <- round( PFS.cutoffs[1, 2], 4)

cutoff.use.Gnormed.poly.clean <- round( PFS.cutoffs[1, 3], 4)

cutoff.use.HT.clean <- round( PFS.cutoffs[1, 4], 4)

cutoff.use.lzpoly.clean <- round( PFS.cutoffs[1, 5], 4)

cutoff.use.lzstar.clean <- round( PFS.cutoffs[1, 6], 4)


################################################################################
################################################################################
######################END: STEP 4 OBTAINING CUTOFFS
################################################################################

################################################################################
################################################################################
#############START: STEP 5 Type I error and power rates
################################################################################



#PART A: Compute PFS for the aberrant datasets
#First Prep data for HT using Mokken package



#compute PFs
#U3poly
U3poly.res.list<-U3poly(data.ab.reverse2, cats)
U3poly.res<-U3poly.res.list[[1]]

#Gpoly
Gpoly.res.list<-Gpoly(data.ab.reverse2, cats)
Gpoly.res<-Gpoly.res.list[[1]]
#Gnormed.poly
Gnormed.poly.res.list<-Gnormed.poly(data.ab.reverse2, cats)
Gnormed.poly.res<-Gnormed.poly.res.list[[1]]
#HT
#first ranspose the data matrix so that persons are columns and items are rows
```

```
   data.aberrant.t<-as.matrix(t(H_data.ab))
   # Using CoefH in Mokken to calculate the Ht for persons instead of items.
   HT.list<-coefH(data.aberrant.t, se=FALSE)
   HT.res<-as.data.frame(HT.list[[2]])



   #lz

   lzpoly.res.list<-lzpoly.mixed(matrix=as.matrix(data.aberrant), IP=IP.est.aber, Ability=
Th.est.ext.aber, C=5)
   lzpoly.res<- as.data.frame(lzpoly.res.list)
   Avglzpoly<-mean(na.omit(as.matrix(lzpoly.res)))
   SDlzpoly<-sd(na.omit(as.matrix(lzpoly.res)))

   #lzstar
   lzstar.res.list<-lzstarpoly.mixed(matrix=as.matrix(data.aberrant), IP=IP.est.aber, C=5,
Ability= Th.est.ext.aber)
   lzstar.res<- as.data.frame(lzstar.res.list)
   Cor_lzandlzstar<-cor(lzpoly.res, lzstar.res, use = 'complete')
   Avglzstar<-mean(na.omit(as.matrix(lzstar.res)))
   SDlzstar<-sd(na.omit(as.matrix(lzstar.res)))


   #PART B: Compute Power and Type I error Using cuttoffs computed with Aberrant parms
   # Type I error and power rates:
   #U3poly
   TypeIerror.U3poly    <- round(mean(na.omit(U3poly.res[[1]][-subs]                >
cutoff.use.U3poly)), 4)
   Power.U3poly <- round(mean(na.omit(U3poly.res[[1]][subs]                        >
cutoff.use.U3poly)), 4)

   #Gpoly
   TypeIerror.Gpoly    <- round(mean(na.omit(Gpoly.res[[1]][-subs]                 >
cutoff.use.Gpoly)), 4)
   Power.Gpoly <- round(mean(na.omit(Gpoly.res[[1]][subs]                          >
cutoff.use.Gpoly)), 4)


   #Gnormed.poly
   TypeIerror.Gnormed.poly    <- round(mean(na.omit(Gnormed.poly.res[[1]][-subs]     >
cutoff.use.Gnormed.poly)), 4)
   Power.Gnormed.poly <- round(mean(na.omit(Gnormed.poly.res[[1]][subs]             >
cutoff.use.Gnormed.poly)), 4)

   #HT
```

```
   TypeIerror.HT    <- round(mean(na.omit(HT.res[[1]][-subs]                    < cutoff.use.HT)),
4)
   Power.HT <- round(mean(na.omit(HT.res[[1]][subs]                            < cutoff.use.HT)),
4)



   #lzpoly
   TypeIerror.lzpoly    <- round(mean(na.omit(lzpoly.res[[1]][-subs]              <
cutoff.use.lzpoly)), 4)
   Power.lzpoly <- round(mean(na.omit(lzpoly.res[[1]][subs]                      <
cutoff.use.lzpoly)), 4)

   #lzstar
   TypeIerror.lzstar   <- round(mean(na.omit(lzstar.res[[1]][-subs]              <
cutoff.use.lzstar)), 4)
   Power.lzstar<- round(mean(na.omit(lzstar.res[[1]][subs]                       <  cutoff.use.lzstar)),
4)



   ######################################################################
   ############ACCURACY####################
   #
   Accuracy.HT<-((Power.HT*length(subs))+((1-TypeIerror.HT)*(n-length(subs))))/n
   Accuracy.Gnormed.poly<-((Power.Gnormed.poly*length(subs))+((1-
TypeIerror.Gnormed.poly)*(n-length(subs))))/n
   Accuracy.Gpoly<-((Power.Gpoly*length(subs))+((1-TypeIerror.Gpoly)*(n-length(subs))))/n
   Accuracy.U3poly<-((Power.U3poly*length(subs))+((1-TypeIerror.U3poly)*(n-
length(subs))))/n
   Accuracy.lzpoly<-((Power.lzpoly*length(subs))+((1-TypeIerror.lzpoly)*(n-length(subs))))/n
   Accuracy.lzstar<-((Power.lzstar*length(subs))+((1-TypeIerror.lzstar)*(n-length(subs))))/n

   #Again, these (below) were not used in the dissertation. I was just curious to see the difference
   Accuracy.HT.clean<-((Power.HT.clean*length(subs))+((1-TypeIerror.HT.clean)*(n-
length(subs))))/n
   Accuracy.Gnormed.poly.clean<-((Power.Gnormed.poly.clean*length(subs))+((1-
TypeIerror.Gnormed.poly.clean)*(n-length(subs))))/n
   Accuracy.Gpoly.clean<-((Power.Gpoly.clean*length(subs))+((1-TypeIerror.Gpoly.clean)*(n-
length(subs))))/n
   Accuracy.U3poly.clean<-((Power.U3poly.clean*length(subs))+((1-
TypeIerror.U3poly.clean)*(n-length(subs))))/n
   Accuracy.lzpoly.clean<-((Power.lzpoly.clean*length(subs))+((1-TypeIerror.lzpoly.clean)*(n-
length(subs))))/n
   Accuracy.lzstar.clean<-((Power.lzstar.clean*length(subs))+((1-TypeIerror.lzstar.clean)*(n-
length(subs))))/n
```

```
#########################################################################
##################################  RESULTS  ##################################


   result<- list(
     c(i,  n_items, AbI, AbN, AbType,
       #
       factors_retained.paran, factors_retained.paran.ab,eig[1], eig[2], eig[3], eig[4], eig[5], eig[6],
eig[7], eig[8], eig[9], eig[10],eig.ab[1], eig.ab[2], eig.ab[3], eig.ab[4], eig.ab[5], eig.ab[6],
eig.ab[7], eig.ab[8], eig.ab[9], eig.ab[10],
       #
       AIC, BIC, AIC.aber, BIC.aber,
       #MODFIT
       perc.item.flagged[1], perc.item.flagged[2], perc.item.flagged[3],  Mean.SD.chsqr[1,1],
Mean.SD.chsqr[1,2],Mean.SD.chsqr[2,1],Mean.SD.chsqr[2,2], Mean.SD.chsqr[3,1],
Mean.SD.chsqr[3,2],
       perc.item.flagged.ab[1], perc.item.flagged.ab[2], perc.item.flagged.ab[3],
Mean.SD.chsqr.ab[1,1], Mean.SD.chsqr.ab[1,2],Mean.SD.chsqr.ab[2,1],Mean.SD.chsqr.ab[2,2],
Mean.SD.chsqr.ab[3,1], Mean.SD.chsqr.ab[3,2],
       #
       length(subs),

       #
       MAD.alpha, BIAS.alpha, cor.alpha,
       MAD.delta, BIAS.delta, cor.delta,
       MAD.taus , BIAS.taus , cor.taus ,
       MAD.th   , BIAS.th   , cor.th   ,
       #
       MAD.alpha.aber    , BIAS.alpha.aber    , cor.alpha.aber,
       MAD.delta.aber    , BIAS.delta.aber    , cor.delta.aber,
       MAD.taus.aber     , BIAS.taus.aber     , cor.taus.aber ,
       MAD.th.aber       , BIAS.th.aber       , cor.th.aber   ,
       #
       MAD.th.aber.fit   , BIAS.th.aber.fit   , cor.th.aber.fit,
       MAD.th.aber.misfit, BIAS.th.aber.misfit, cor.th.aber.misfit,
       #
       H, H_SE, H.aber, H_SE.aber,
       reliab.MS, reliab.MS.aber, reliab.alpha, reliab.alpha.aber, reliab.lambda2,
reliab.lambda2.aber,
       #
       Power.U3poly, Power.U3poly.clean, Power.Gpoly, Power.Gpoly.clean,
Power.Gnormed.poly,Power.Gnormed.poly.clean, Power.HT, Power.HT.clean, Power.lzpoly,
Power.lzpoly.clean,Power.lzstar, Power.lzstar.clean,
```

```
      TypeIerror.U3poly, TypeIerror.Gpoly, TypeIerror.Gnormed.poly,
TypeIerror.HT,TypeIerror.lzpoly, TypeIerror.lzstar,
      TypeIerror.U3poly.clean, TypeIerror.Gpoly.clean, TypeIerror.Gnormed.poly.clean,
TypeIerror.HT.clean,TypeIerror.lzpoly.clean, TypeIerror.lzstar.clean,

      #
      Accuracy.U3poly, Accuracy.Gpoly, Accuracy.Gnormed.poly,
Accuracy.HT,Accuracy.lzpoly,Accuracy.lzstar,
      Accuracy.U3poly.clean, Accuracy.Gpoly.clean, Accuracy.Gnormed.poly.clean,
Accuracy.HT.clean,Accuracy.lzpoly.clean,Accuracy.lzstar.clean,
      #
      # Avglzpoly.rev, Avglzpoly, SDlzpoly.rev, SDlzpoly, Avglzstar.rev, Avglzstar,
SDlzstar.rev, SDlzstar,
      #cutoffs
      cutoff.use.U3poly.clean, cutoff.use.U3poly,
      cutoff.use.Gpoly.clean, cutoff.use.Gpoly,
      cutoff.use.Gnormed.poly.clean, cutoff.use.Gnormed.poly,
      cutoff.use.HT.clean, cutoff.use.HT,
      cutoff.use.lzpoly.clean, cutoff.use.lzpoly,
      cutoff.use.lzstar.clean, cutoff.use.lzstar,
      cutoff.SE.U3poly,
      cutoff.SE.Gpoly,
      cutoff.SE.Gnormed.poly,
      cutoff.SE.HT,
      cutoff.SE.lzpoly,
      cutoff.SE.lzstar

    )
  )


  #Clean memory:


  rm(data.ext, gendata, data.aberrant, IP.est, IP.est.aber,factors_retained.paran.ab,
factors_retained.paran,
     N.extreme, N.middle, subs, subs.extreme, subs.middle, subs.random, subs.longstring,
     #ratio.median.extreme.to.middle, Median.I.extreme.scores, Median.I.middle.scores,eig,
eig.ab
     Th.est, Th.est.ext, Th.est.aber, Th.est.ext.aber,eig, eig.ab,
     n_items, AbI, AbN,
     #
     scrs.extreme, scrs.longstring, scrs.middle, scrs.random,
     #
     MAD.alpha, BIAS.alpha, cor.alpha,
     MAD.delta, BIAS.delta, cor.delta,
```

MAD.taus , BIAS.taus , cor.taus ,
MAD.th , BIAS.th , cor.th ,
MAD.alpha.aber , BIAS.alpha.aber , cor.alpha.aber,
MAD.delta.aber , BIAS.delta.aber , cor.delta.aber,
MAD.taus.aber , BIAS.taus.aber , cor.taus.aber ,
MAD.th.aber , BIAS.th.aber , cor.th.aber ,
MAD.th.aber.fit , BIAS.th.aber.fit , cor.th.aber.fit,
MAD.th.aber.misfit, BIAS.th.aber.misfit, cor.th.aber.misfit,
MODFIT.res, MODFIT.res.aber, AIC, BIC, AIC.aber, BIC.aber, data.cor,eig,
eig_gte1,data.cor.aber,eig.aber, eig_gte1.aber,
H, H_SE, H.aber, H_SE.aber,
reliab.MS, reliab.MS.aber, reliab.alpha, reliab.alpha.aber, reliab.lambda2,
reliab.lambda2.aber,
#
#TypeIerror.U3poly, Power.U3poly, TypeIerror.Gpoly,
Power.Gpoly,TypeIerror.Gnormed.poly, Power.Gnormed.poly,TypeIerror.HT, Power.HT,
#MODFIT
perc.item.flagged, perc.item.flagged.ab, Mean.SD.chsqr, Mean.SD.chsqr.ab,
#power/typeIerror
Power.U3poly, Power.Gpoly, Power.Gnormed.poly, Power.HT, TypeIerror.U3poly,
TypeIerror.Gpoly, TypeIerror.Gnormed.poly, TypeIerror.HT, Power.lzpoly,
Power.lzpoly.clean,Power.lzstar, Power.lzstar.clean,
Power.U3poly.clean, Power.Gpoly.clean, Power.Gnormed.poly.clean, Power.HT.clean,
TypeIerror.U3poly.clean, TypeIerror.Gpoly.clean, TypeIerror.Gnormed.poly.clean,
TypeIerror.HT.clean,
TypeIerror.lzpoly, TypeIerror.lzstar,
#
Accuracy.U3poly, Accuracy.Gpoly, Accuracy.Gnormed.poly,
Accuracy.HT,Accuracy.lzpoly,Accuracy.lzstar,
Accuracy.U3poly.clean, Accuracy.Gpoly.clean, Accuracy.Gnormed.poly.clean,
Accuracy.HT.clean,Accuracy.lzpoly.clean,Accuracy.lzstar.clean,
#
Avglzpoly.rev, Avglzpoly, SDlzpoly.rev, SDlzpoly, Avglzstar.rev, Avglzstar,
SDlzstar.rev, SDlzstar,
#cutoffs
cutoff.use.U3poly.clean, cutoff.use.U3poly,
cutoff.use.Gpoly.clean, cutoff.use.Gpoly,
cutoff.use.Gnormed.poly.clean, cutoff.use.Gnormed.poly,
cutoff.use.HT.clean, cutoff.use.HT,
cutoff.use.lzpoly.clean, cutoff.use.lzpoly,
cutoff.use.lzstar.clean, cutoff.use.lzstar,
cutoff.SE.U3poly,
cutoff.SE.Gpoly,
cutoff.SE.Gnormed.poly,
cutoff.SE.HT,
cutoff.SE.lzpoly,

```
    cutoff.SE.lzstar,

    modelfitting.data, modelfitting.data.reverse1, modelfitting.data.reverse2



  )



  result.tbl<-as.data.frame(t(unlist(result)))

  colnames(result.tbl) <- c("i",  "n_items", "AbI", "AbN", "AbType",
                            #
                            "factors", "factors.aber","eig[1]", "eig[2]", "eig[3]", "eig[4]", "eig[5]",
"eig[6]", "eig[7]", "eig[8]", "eig[9]", "eig[10]","eig.ab[1]", "eig.ab[2]", "eig.ab[3]","eig.ab[4]",
"eig.ab[5]", "eig.ab[6]", "eig.ab[7]", "eig.ab[8]", "eig.ab[9]", "eig.ab[10]",
                            #
                            #
                            "AIC", "BIC", "AIC.aber", "BIC.aber",
                            #MODFIT
                            "perc.sing.flagged", "perc.dbls.flagged", "perc.trpls.flagged",
"mean.chisqr.sing", "SD.chisqr.sing", "mean.chisqr.dbls", "SD.chisqr.dbls", "mean.chisqr.trpls",
"SD.chisqr.trpls",
                            "perc.sing.flagged.ab", "perc.dbls.flagged.ab", "perc.trpls.flagged.ab",
"mean.chisqr.sing.ab", "SD.chisqr.sing.ab", "mean.chisqr.dbls.ab", "SD.chisqr.dbls.ab",
"mean.chisqr.trpls.ab", "SD.chisqr.trpls.ab",
                            #
                            "length(subs)",
                            #ratio.median.extreme.to.middle, Median.I.extreme.scores,
Median.I.middle.scores,
                            #
                            "MAD.alpha", "BIAS.alpha", "cor.alpha",
                            "MAD.delta", "BIAS.delta", "cor.delta",
                            "MAD.taus" , "BIAS.taus" , "cor.taus" ,
                            "MAD.th"   , "BIAS.th"   , "cor.th"   ,
                            #
                            "MAD.alpha.aber"    , "BIAS.alpha.aber"    , "cor.alpha.aber",
                            "MAD.delta.aber"    , "BIAS.delta.aber"    , "cor.delta.aber",
                            "MAD.taus.aber"     , "BIAS.taus.aber "    , "cor.taus.aber" ,
                            "MAD.th.aber"       , "BIAS.th.aber"       , "cor.th.aber"   ,
                            #
                            "MAD.th.aber.fit"   , "BIAS.th.aber.fit"   , "cor.th.aber.fit",
                            "MAD.th.aber.misfit", "BIAS.th.aber.misfit", "cor.th.aber.misfit",
                            #
                            "H", "H_SE", "H.aber", "H_SE.aber",
```

```
                    "reliab.MS", "reliab.MS.aber", "reliab.alpha", "reliab.alpha.aber",
"reliab.lambda2", "reliab.lambda2.aber",#74
                    #
                    "Power.U3poly","Power.U3poly.clean", "Power.Gpoly","Power.Gpoly.clean",
"Power.Gnormed.poly","Power.Gnormed.poly.clean", "Power.HT",
"Power.HT.clean","Power.lzpoly", "Power.lzpoly.clean","Power.lzstar", "Power.lzstar.clean",
                    "TypeIerror.U3poly", "TypeIerror.Gpoly", "TypeIerror.Gnormed.poly",
"TypeIerror.HT","TypeIerror.lzpoly", "TypeIerror.lzstar",
                    "TypeIerror.U3poly.clean", "TypeIerror.Gpoly.clean",
"TypeIerror.Gnormed.poly.clean", "TypeIerror.HT.clean","TypeIerror.lzpoly.clean",
"TypeIerror.lzstar.clean",
                    #
                    "Accuracy.U3poly", "Accuracy.Gpoly", "Accuracy.Gnormed.poly",
"Accuracy.HT","Accuracy.lzpoly","Accuracy.lzstar",
                    "Accuracy.U3poly.clean", "Accuracy.Gpoly.clean",
"Accuracy.Gnormed.poly.clean",
"Accuracy.HT.clean","Accuracy.lzpoly.clean","Accuracy.lzstar.clean",
                    #
                    #"Avglzpoly.rev",  "Avglzpoly",  "SDlzpoly.rev",  "SDlzpoly",
"Avglzstar.rev",  "Avglzstar",  "SDlzstar.rev",  "SDlzstar",
                    #cutoffs
                    "cutoff.use.U3poly.clean", "cutoff.use.U3poly",
                    "cutoff.use.Gpoly.clean", "cutoff.use.Gpoly",
                    "cutoff.use.Gnormed.poly.clean", "cutoff.use.Gnormed.poly",
                    "cutoff.use.HT.clean", "cutoff.use.HT",
                    "cutoff.use.lzpoly.clean", "cutoff.use.lzpoly",
                    "cutoff.use.lzstar.clean", "cutoff.use.lzstar",

                    "cutoff.SE.U3poly",
                    "cutoff.SE.Gpoly",
                    "cutoff.SE.Gnormed.poly",
                    "cutoff.SE.HT",
                    "cutoff.SE.lzpoly",
                    "cutoff.SE.lzstar"

  )
  result.df<-as.data.frame(result.tbl)
  result.df$rep<-rep
  result.df$modelcondition<-"1A"
  result.df$modelgen<-"GGUM"
  result.df$modelfit<- "GGUM"

  write.table(result.df, file=paste0("C:/Users/jreimers/Documents/Jreimers/NEW
DISSERTATION/SimOutcomes/1A_Results/GGUMGGUM_condition_",
i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)
```

```
  }

print(Sys.time())
print(Sys.time() - start.time)

# Stop parallel cluster:
stopCluster(cl)

# END SECTION
```

Appendix B

R Code for Condition of GPCM Fit to GPCM Data

```
### much of the below code is inspired by and adapted from Tendeiro's work on OSF:
https://www.jorgetendeiro.com/publication/tendeiro_2017/


# 0. Prepare environment ----
rm(list=ls())
if (!is.null(dev.list())) dev.off(dev.list()["RStudioGD"])
library(Rcpp)
library(psych)
library(fastGHQuad)
library(abind)
library(mirt)
library(stats4)
library(stats)
library(PP)
library(Matrix)
library(paran)

#Parallel Analysis
library(paran)
library(Hmisc)
#PF

library(RCurl)
library(PerFit)
library(mokken)
# Parallel processing:
library(doParallel)
library(foreach)

#1. Conditions and fixed parameters
n_items.vec   <- c( 20, 40)
AbI.vec <- c(.20, .40, .60)
AbN.vec <- c(.04, .10, .20)
AbType.vec   <- c(  "Random_Responders", "Longstringers",   "ERS", "MRS", "Mixed")
parameters          <- expand.grid(n_items.vec, AbI.vec, AbN.vec, AbType.vec)
colnames(parameters) <- c("n_items.vec", "AbI.vec", "AbN.vec", "AbType.vec")
rm(n_items.vec, AbI.vec, AbN.vec, AbType.vec)
# Fixed parameters:
n       <- 1000
N       <- 1000
cats    <- 6
```

```
#detectCores()


#  Setup parallel backend to use 3 processors (cores):
cl <- makeCluster(7, setup_strategy = "sequential")
registerDoParallel(cl, cores = 7)
# END SECTION


#Also run Modfit.gpcm and lzstarMix functions
#
#
#
# 3. Run the simulation ----

#
start.time <- Sys.time()
print(start.time)

outcome.simulation <- foreach(i=1:1) %:%
  foreach(rep=1:50, .packages=c("psych", "fastGHQuad", "abind", "mokken", "PerFit", "paran",
"Hmisc", "PP", "mirt", "Matrix")) %dopar% {
    set.seed(1000*i+rep)
    # Specify varying parameters for cell
    n_items    <- parameters[i, 1]
    AbI        <- parameters[i, 2]
    AbN        <- parameters[i, 3]
    AbType     <- parameters[i, 4]
    n_Abitems<-(n_items*AbI)
    n_Cleanitems<-n_items-n_Abitems


################################################################################
################################################################################
##############START: STEP 1 Generate CLEAN Data & Estimate Parameters############

    # Generate item scores according to the GPCM:
    ## discrimination parameters from random uniform distribution for each dataset (a_1, a_2,...)
    a <- matrix(runif(n=n_items*1, min=.5, max=2)) #create dataset for each rep
    data.a<- t(a)

    ###generating step difficulty parms: used website: https://rpubs.com/okanbulut/pcmsimulation
    zerovector<- (rep(0,n_items))
    #generate differences using cumsum
    difmatrix<-t(apply(matrix(runif(n_items*(cats-1),.3,1),n_items),1, cumsum))
    #transform differences to d_jk by making them relative to mean
```

```
   d_jk <- -(difmatrix-rowMeans(difmatrix))
   #generate item location parms b_j
   i.locations<- rnorm(n_items)
   #compute item category location parms
   True.bparms_gpcm<- (i.locations+d_jk)*-1
   thresholds<- rbind(t(zerovector), t(True.bparms_gpcm))


   ##Generate true thetas
   TrueThetasGPCM= rnorm(n, mean=0, sd=1)
   data.TrueThetasGPCM=as.matrix(TrueThetasGPCM)
   colnames(data.TrueThetasGPCM)<- c("TrueTheta")

   Taus_gpcm<-as.data.frame("NA")

   data.gpcm<- as.matrix(sim_gpcm(thres=thresholds, alpha=a, theta=TrueThetasGPCM))

   gendata<-list(alpha.gen=data.a, b.gen=True.bparms_gpcm, Taus_NA=Taus_gpcm,
theta.gen=data.TrueThetasGPCM, data=data.gpcm)

   data.ext <-gendata[[5]]

   item_array<- c(paste0("item", seq(1,n_items)))
   colnames(data.gpcm)<-item_array

   #write.table(data.ext, file=paste0("/Volumes/Backup Plus/NEW
DISSERTATION/Datasets/1C GPCMGPCM/3.GPCMGPCM_clean/GPCMGPCM_condition_",
i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)

   # Estimate item and person parameters BEFORE aberrant behavior:
   model.gpcm<-paste0("f=1-",n_items)
   results.gpcm<-mirt(data=data.gpcm, model=model.gpcm, itemtype="gpcm", SE=TRUE,
verbose=FALSE)
   IP.est<- as.data.frame(coef(results.gpcm, IRTpars=TRUE, simplify=TRUE))

   Th.est     <- fscores(results.gpcm, method="EAP", full.scores=TRUE)
   Th.est.ext <- as.vector(Th.est)



##############################################################################
##############################################################################
####################END: STEP 1 Generate CLEAN Data & Estimate Parameters
##############################################################################
```

```
################################################################################
################################################################################
######################START: STEP 4 OBTAINING CUTOFFS
################################################################################


   ##PART A: Using Parm ests from aberrant data
   Nreps=20
   Th.est.aber.NA <- which(is.na(Th.est.ext))
   PFS.cutoffs <- matrix(NA, Nreps, 6)
   for (r in 1:Nreps){ #r=1
     #Generate model-fitting item score vectors based on GPCM and the estimated person and
item parameters using the aberrant data.
     ####
     thresholds.modelfit<- as.data.frame(rbind(t(zerovector), t(IP.est[, 2:cats])))

     modelfitting.data <- as.matrix(sim_gpcm(thres=thresholds.modelfit, alpha=IP.est$items.a ,
theta=Th.est.ext))

     item_array<- c(paste0("item", seq(1,n_items)))
     colnames(modelfitting.data )<-item_array

     #estimate parms for study 3 with parametric PFS
     model.gpcm<-paste0("f=1-",n_items)
     results.gpcm.modelfit<-mirt(data=modelfitting.data, model=model.gpcm, itemtype="gpcm",
SE=TRUE, verbose=FALSE)
     IP.est.modelfitting <- as.data.frame(coef(results.gpcm.modelfit, IRTpars=TRUE,
simplify=TRUE))

     Th.est.gendata   <- fscores(results.gpcm.modelfit, method="EAP", full.scores=TRUE)
     Th.est.ext.gendata <- as.vector(Th.est.gendata)



   #compute PFSs for the generated data for the purpose of creating the cutoffs (median over
replications)

   ##Prep dataset for IIO and HT analyses
   find.invariants <- apply( modelfitting.data, 1, function(vec) max(vec) - min(vec))
   pos.invariants  <- which(find.invariants == 0)
   H_data.cutoff   <- if (length(pos.invariants) > 0)  modelfitting.data[-pos.invariants,] + 1 else
modelfitting.data + 1


   #compute PFs
   #U3poly
```

```
   U3poly.res.list<-U3poly( modelfitting.data , (cats))
   U3poly.res.modfit<-U3poly.res.list[[1]]

   #Gpoly
   Gpoly.res.list<-Gpoly( modelfitting.data , (cats))
   Gpoly.res.modfit<-Gpoly.res.list[[1]]


   #Gnormed.poly
   Gnormed.poly.res.list<-Gnormed.poly( modelfitting.data , (cats))
   Gnormed.poly.res.modfit<-Gnormed.poly.res.list[[1]]
   #HT
   #first ranspose the data matrix so that persons are columns and items are rows
   data.aberrant.t<-as.matrix(t(H_data.cutoff))
   # Using CoefH in Mokken to calculate the Ht for persons instead of items.
   HT.list<-coefH(data.aberrant.t, se=FALSE)
   HT.res.modfit<-as.data.frame(HT.list[[2]])

   #lzpoly and lzstar using Sandhip's lzstarMix function
   a.star<-as.numeric(unlist(IP.est.modelfitting[,1]))
   b.star<-matrix(as.numeric(unlist(IP.est.modelfitting[, 2:cats])), nrow=n_items)
   scores.1<- matrix(as.double(as.character(modelfitting.data)), nrow=n)

   lzstar.res.list<-lzstarMix(scores = scores.1,theta=Th.est.ext.gendata,a=a.star,b=b.star,
c=NULL, est="mle")
   lzstar.res.modfit<-lzstar.res.list[,2]
   lzpoly.res.modfit<-lzstar.res.list[,1]




   #Compute cutoff (used median)
   U3poly.modfit.cut       <- round(quantile(U3poly.res.modfit$PFscores, probs = .95), 4)
   Gpoly.modfit.cut        <- round(quantile(Gpoly.res.modfit$PFscores, probs = .95), 4)
   Gnormed.poly.modfit.cut  <- round(quantile(Gnormed.poly.res.modfit$PFscores, probs = .95),
4)
   HT.modfit.cut           <- round(quantile( HT.res.modfit, probs = .05, na.rm=TRUE), 4)
   lzpoly.modfit.cut       <- round(quantile(lzpoly.res.modfit, probs = .05, na.rm=TRUE), 4)
   lzstar.modfit.cut       <- round(quantile(lzstar.res.modfit, probs = .05, na.rm=TRUE), 4)

   #r=1

   PFS.cutoffs[r, 1]<-  U3poly.modfit.cut
   PFS.cutoffs[r, 2]<-  Gpoly.modfit.cut
   PFS.cutoffs[r, 3]<-  Gnormed.poly.modfit.cut
   PFS.cutoffs[r, 4]<-  HT.modfit.cut
```

```
PFS.cutoffs[r, 5]<-  lzpoly.modfit.cut
PFS.cutoffs[r, 6]<-  lzstar.modfit.cut
}
names(PFS.cutoffs)<- c("U3poly", "Gpoly", "Gnormed.poly", "HT", "lzpoly", "lzstar")



cutoff.use.U3poly <- round( median(PFS.cutoffs[, 1]), 4)
cutoff.SE.U3poly  <- round(sd(PFS.cutoffs[, 1]), 4)

cutoff.use.Gpoly <- round( median(PFS.cutoffs[, 2]), 4)
cutoff.SE.Gpoly  <- round(sd(PFS.cutoffs[, 2]), 4)

cutoff.use.Gnormed.poly <- round( median(PFS.cutoffs[, 3]), 4)
cutoff.SE.Gnormed.poly  <- round(sd(PFS.cutoffs[, 3]), 4)

cutoff.use.HT <- round( median(PFS.cutoffs[, 4]), 4)
cutoff.SE.HT  <- round(sd(PFS.cutoffs[, 4]), 4)

cutoff.use.lzpoly <- round( median(PFS.cutoffs[, 5]), 4)
cutoff.SE.lzpoly  <- round(sd(PFS.cutoffs[, 5]), 4)

cutoff.use.lzstar <- round( median(PFS.cutoffs[, 6]), 4)
cutoff.SE.lzstar  <- round(sd(PFS.cutoffs[, 6]), 4)




################################################################################
################################################################################
##############################END: STEP 4 OBTAINING CUTOFFS
################################################################################




######################START: STEP 5 Type I error and power rates ###################
################################################################################
################################################################################


#PART A: Compute PFS for the aberrant datasets
```

```
   #First Prep data for HT using mokken package

   #compute PFs
   #U3poly
   U3poly.res.list<-U3poly(data.ext, (cats))
   U3poly.res<-U3poly.res.list[[1]]

   #Gpoly
   Gpoly.res.list<-Gpoly(data.ext, (cats))
   Gpoly.res<-Gpoly.res.list[[1]]
   #Gnormed.poly
   Gnormed.poly.res.list<-Gnormed.poly(data.ext, (cats))
   Gnormed.poly.res<-Gnormed.poly.res.list[[1]]
   #HT
   #first ranspose the data matrix so that persons are columns and items are rows
   data.aberrant.t<-as.matrix(t(data.ext))
   # Using CoefH in Mokken to calculate the Ht for persons instead of items.
   HT.list<-coefH(data.aberrant.t, se=FALSE)
   HT.res<-as.data.frame(HT.list[[2]])



   #lzpoly and lzstar using Sandhip's lzstarMix function
   a.star<-as.numeric(unlist(IP.est[,1]))
   b.star<-matrix(as.numeric(unlist(IP.est[, 2:cats])), nrow=n_items)
   scores.1<- matrix(as.double(as.character(data.ext)), nrow=n)

   lzstar.res.list<-lzstarMix(scores = scores.1,theta=Th.est.ext,a=a.star,b=b.star, c=NULL,
est="mle")
   lzstar.res<-lzstar.res.list[,2]
   lzpoly.res<-lzstar.res.list[,1]


   #PART B: Compute Power and Type I error Using cuttoffs computed with Aberrant parms
   # Type I error and power rates:
   #U3poly
   TypeIerror.U3poly    <- round(mean(na.omit(U3poly.res[[1]]               >  cutoff.use.U3poly)),
4)

   #Gpoly
   TypeIerror.Gpoly    <- round(mean(na.omit(Gpoly.res[[1]]               > cutoff.use.Gpoly)), 4)


   #Gnormed.poly
   TypeIerror.Gnormed.poly    <- round(mean(na.omit(Gnormed.poly.res[[1]]      >
cutoff.use.Gnormed.poly)), 4)
```

```
   #HT
   TypeIerror.HT    <- round(mean(na.omit(HT.res[[1]]                    < cutoff.use.HT)), 4)



   #lzpoly
   TypeIerror.lzpoly    <- round(mean(na.omit(lzpoly.res        <  cutoff.use.lzpoly)), 4)

   # #lzstar
   TypeIerror.lzstar   <- round(mean(na.omit(lzstar.res                    <  cutoff.use.lzstar)), 4)



##############################################################################
##############################################################################
###################################RESULTS    ###############################
##############################################################################

   result<- list(
     c(  TypeIerror.U3poly, TypeIerror.Gpoly, TypeIerror.Gnormed.poly,
TypeIerror.HT,TypeIerror.lzpoly, TypeIerror.lzstar)
   )


   result.tbl<-as.data.frame(t(unlist(result)))

   colnames(result.tbl) <- c("TypeIerror.U3poly", "TypeIerror.Gpoly",
"TypeIerror.Gnormed.poly", "TypeIerror.HT","TypeIerror.lzpoly", "TypeIerror.lzstar")



   result.df<-as.data.frame(result.tbl)
   result.df$rep<-rep
   result.df$modelcondition<-"1C"
   result.df$modelgen<-"GPCM"
   result.df$modelfit<-"GPCM"

   write.table(result.df, file=paste0("/Volumes/Backup Plus/NEW
DISSERTATION/SimOutcomes/TypeIError/C/GPCMGPCM_condition_",
i,"rep",rep,".txt"),col.names=TRUE, row.names=FALSE, sep=", ", quote=FALSE)


  }
```

```
print(Sys.time())
print(Sys.time() - start.time)

# Stop parallel cluster:
stopCluster(cl)

# END SECTION
```

Appendix C

R Code for MODFIT function Using GPCM

```
MODFIT.gpcm <- function(data, model, C, IP.est, precision = 4)
{
 N     <- nrow(data)
 I     <- ncol(data)
 N.NAs <- N - colSums(is.na(data))

 if (I <= 10)
 {
  doublets <- t(combn(I, 2))
  triplets <- t(combn(I, 3))
 } else {
  # Find packets:
  obs.props   <- colMeans(data > 0, na.rm = TRUE)
  group.low   <- sort(order(obs.props)[1:ceiling(I / 3)])
  group.med   <- sort(order(obs.props)[(ceiling(I / 3) + 1) : ceiling(2*I/3)])
  group.high  <- sort(order(obs.props)[(ceiling(2*I / 3) + 1) : I])
  groups      <- cbind(group.low = group.low[1:floor(I / 3)],
                       group.med[1:floor(I / 3)],
                       group.high[1:floor(I / 3)])
  packets     <- lapply(seq_len(nrow(groups)), function(row) sort(groups[row, ]))
  if ((I %% 3) == 1) {packets[[1]][4] <- group.low[ceiling(I / 3)]}
  if ((I %% 3) == 2) {
   packets[[1]][4] <- group.low[ceiling(I / 3)]
   packets[[2]][4] <- group.med[ceiling(I / 3)]
  }

  #
  # singlets    <- 1:I
  doublets    <- matrix(unlist(lapply(packets, function(x) combn(x,2))),
                        ncol = 2, byrow = TRUE)
  triplets    <- matrix(unlist(lapply(packets, function(x) combn(x,3))),
                        ncol = 3, byrow = TRUE)
 }

 # NAs for doublets and triplets:
 N.NAs.doublets <- apply(doublets, 1,
                         function(vec) N - sum(rowSums(is.na(data[, vec])) > 0))
 N.NAs.triplets <- apply(triplets, 1,
                         function(vec) N - sum(rowSums(is.na(data[, vec])) > 0))

 # Nodes and weights:
 nodes.chi   <- seq(-3, 3, length.out = 61)
```

```
  N.nodes.chi <- length(nodes.chi)
  weights     <- dnorm(nodes.chi) / sum(dnorm(nodes.chi))


  # Singlets:
  probs.array.aber.drasgow <- array(NA, dim = c(N.nodes.chi, I, max(C) + 1))
  if (model == "GPCM")
  {
   probs.list.drasgow<-
     plink::gpcm(x=IP.est[1:(C+1)], cat=rep((C+1), I), theta=nodes.chi )
   probs.out.drasgow<-probs.list.drasgow@prob[,2:(I*(C+1)+1)]
   for (i in 1:I)
   {
    for (z in 0:C)
    {
     probs.array.aber.drasgow[, i, z + 1] <- probs.out.drasgow[,((z+1)+((i-1)*(C+1)))]

    }
   }
  }
  weights.arr             <- array(rep(weights, I * (max(C) + 1)),
                             c(N.nodes.chi, I, max(C) + 1))
  N.NAs.mat               <- matrix(rep(N.NAs, max(C) + 1), nrow = I,
                             byrow = FALSE)
  if (length(C) > 1) for (i in 1:I) N.NAs.mat[i, (C[i] + 1):(max(C) + 1)] <- NA
  expected.mat.drasgow    <- N.NAs.mat * apply((probs.array.aber.drasgow * weights.arr), 2:3,
sum)
  if (length(C) == 1)
  {
   observed.mat.drasgow    <- t(apply(data, 2,
                              function(vec) table(factor(vec, levels = 0:C))))
  } else
  {
   observed.mat.drasgow <- matrix(NA, nrow = I, ncol = max(C) + 1)
   for (i in 1:I)
   {
    observed.mat.drasgow[i, 1:(C[i] + 1)] <- table(factor(data[, i], levels = 0:C[i]))
   }
  }
  # Merge cells with expected frequencies < 5:
  expected.order              <- t(apply(expected.mat.drasgow, 1, order))
  expected.mat.drasgow        <- t(sapply(1:I, function(it) expected.mat.drasgow[it,
expected.order[it, ]]))
  observed.mat.drasgow        <- t(sapply(1:I, function(it) observed.mat.drasgow[it,
expected.order[it, ]]))
  expected.mat.drasgow.less5  <- rowSums(expected.mat.drasgow < 5, na.rm = TRUE)
```

```
  N.expected.mat.drasgow.less5   <- sum(expected.mat.drasgow.less5 > 0)
  pos.expected.mat.drasgow.less5 <- which(expected.mat.drasgow.less5 > 0)
  df                             <- if (length(C) == 1) rep(C, I) else C
  if (N.expected.mat.drasgow.less5 > 0)
  {
    sapply(1:N.expected.mat.drasgow.less5, function(it)
    {
      item    <- pos.expected.mat.drasgow.less5[it]
      pos.sum <- expected.mat.drasgow.less5[item]
      if (sum(expected.mat.drasgow[item, 1:pos.sum]) < 5) {pos.sum <- pos.sum + 1}
      expected.mat.drasgow[item, pos.sum]         <- sum(expected.mat.drasgow[item, 1:pos.sum])
      expected.mat.drasgow[item, 1:(pos.sum - 1)] <- 1
      observed.mat.drasgow[item, pos.sum]         <- sum(observed.mat.drasgow[item, 1:pos.sum])
      observed.mat.drasgow[item, 1:(pos.sum - 1)] <- 1
      df[item] <- if (length(C) == 1) (C + 1) - pos.sum else (C[item] + 1) - pos.sum
    })
  }
  # Compute (adjusted) chi squares (/df):
  chisq        <- rowSums(((observed.mat.drasgow - expected.mat.drasgow)^2) /
expected.mat.drasgow, na.rm = TRUE)
  chisq.df     <- chisq / df
  chisq.adj    <- sapply(1:I, function(it) max(0, 3000 * (chisq[it] - df[it]) / N.NAs[it] + df[it]))
  chisq.adj.df <- chisq.adj / df
  singlets.res <- cbind(Item = 1:I, N.NAs, df, chisq, chisq.df, chisq.adj, chisq.adj.df)

  # Doublets:
  doublets.NAs <- cbind(doublets, N.NAs.doublets)
  doublets.res <- t(apply(doublets.NAs, 1, function(vec)
  {
    item1   <- vec[1]
    item2   <- vec[2]
    N.NAs.d <- vec[3]
    probs.array.aber.drasgow.item1     <- probs.array.aber.drasgow[, item1, ]
    probs.array.aber.drasgow.item2     <- probs.array.aber.drasgow[, item2, ]
    probs.array.aber.drasgow.item1.arr <- array(rep(probs.array.aber.drasgow.item1, max(C) + 1),
c(N.nodes.chi, max(C) + 1, max(C) + 1))
    probs.array.aber.drasgow.item2.arr <- array(rep(probs.array.aber.drasgow.item2, max(C) + 1),
c(N.nodes.chi, max(C) + 1, max(C) + 1))
    probs.array.aber.drasgow.item2.arr <- aperm(probs.array.aber.drasgow.item2.arr, c(1, 3, 2))
    weights.arr2                       <- array(rep(weights.arr, (max(C) + 1) * (max(C) + 1)),
c(N.nodes.chi, max(C) + 1, max(C) + 1)) # For doublets
    expected.mat.drasgow.it1.it2       <- N.NAs.d * apply(probs.array.aber.drasgow.item1.arr *
probs.array.aber.drasgow.item2.arr * weights.arr2, 2:3, sum)
    if (length(C) > 1)
    {
      if (C[item1] < max(C)) expected.mat.drasgow.it1.it2[(C[item1] + 2):(max(C) + 1), ] <- NA
```

```
    if (C[item2] < max(C)) expected.mat.drasgow.it1.it2[, (C[item2] + 2):(max(C) + 1)] <- NA
   }
   if (length(C) == 1)
   {
    observed.mat.drasgow.it1.it2 <- table(factor(data[, item1], levels = 0:C), factor(data[, item2],
levels = 0:C)) # table(data[, item1], data[, item2])
   } else
   {
    observed.mat.drasgow.it1.it2 <- matrix(NA, nrow = max(C) + 1, ncol = max(C) + 1)
    observed.mat.drasgow.it1.it2[1:(C[item1] + 1), 1:(C[item2] + 1)] <- table(factor(data[,
item1], levels = 0:C[item1]), factor(data[, item2], levels = 0:C[item2]))
   }
   # Merge cells with expected frequencies < 5:
   expected.order.it1.it2           <- order(c(expected.mat.drasgow.it1.it2))
   expected.mat.drasgow.it1.it2     <- c(expected.mat.drasgow.it1.it2)[expected.order.it1.it2]
   observed.mat.drasgow.it1.it2     <- c(observed.mat.drasgow.it1.it2)[expected.order.it1.it2]
   expected.mat.drasgow.it1.it2.less5 <- expected.mat.drasgow.it1.it2 < 5
   if (length(C) == 1) {df.it1.it2 <- (C + 1)^2 - 1} else {df.it1.it2 <- (C[item1] + 1) * (C[item2] +
1) - 1}

   if (sum(expected.mat.drasgow.it1.it2.less5, na.rm = TRUE) > 0)
   {
    pos.sum <- max(which(expected.mat.drasgow.it1.it2.less5 == 1))
    if (sum(expected.mat.drasgow.it1.it2[1:pos.sum]) < 5) {pos.sum <- pos.sum + 1}
    expected.mat.drasgow.it1.it2[pos.sum] <- sum(expected.mat.drasgow.it1.it2[1:pos.sum])
    expected.mat.drasgow.it1.it2        <- expected.mat.drasgow.it1.it2[pos.sum : ((max(C) +
1)^2)]
    observed.mat.drasgow.it1.it2[pos.sum] <- sum(observed.mat.drasgow.it1.it2[1:pos.sum])
    observed.mat.drasgow.it1.it2        <- observed.mat.drasgow.it1.it2[pos.sum : ((max(C) +
1)^2)]
    if (length(C) == 1) {df.it1.it2 <- (C + 1)^2 - pos.sum} else {df.it1.it2 <- (C[item1] + 1) *
(C[item2] + 1) - pos.sum}
   }
   # Compute (adjusted) chi squares (/df):
   chisq       <- sum(((observed.mat.drasgow.it1.it2 - expected.mat.drasgow.it1.it2)^2) /
expected.mat.drasgow.it1.it2, na.rm = TRUE)
   chisq.df    <- chisq / df.it1.it2
   chisq.adj   <- max(0, 3000 * (chisq - df.it1.it2) / N.NAs.d + df.it1.it2)
   chisq.adj.df <- chisq.adj / df.it1.it2
   c(Item1 = item1, Item2 = item2, N = N.NAs.d, df = df.it1.it2, chisq = chisq, chisq.df =
chisq.df, chisq.adj = chisq.adj, chisq.adj.df = chisq.adj.df)
  }))
  doublets.res <- cbind(Doublet = 1:nrow(doublets), doublets.res)

  # Triplets:
  triplets.NAs <- cbind(triplets, N.NAs.triplets)
```

```
  triplets.res <-  t(apply(triplets.NAs, 1, function(vec)
  {
    item1   <- vec[1]
    item2   <- vec[2]
    item3   <- vec[3]
    N.NAs.t <- vec[4]
    probs.array.aber.drasgow.item1     <- probs.array.aber.drasgow[, item1, ]
    probs.array.aber.drasgow.item2     <- probs.array.aber.drasgow[, item2, ]
    probs.array.aber.drasgow.item3     <- probs.array.aber.drasgow[, item3, ]
    probs.array.aber.drasgow.item1.arr <- array(rep(probs.array.aber.drasgow.item1, (max(C) + 1)
* (max(C) + 1)), c(N.nodes.chi, max(C) + 1, max(C) + 1, max(C) + 1))
    probs.array.aber.drasgow.item2.arr <- array(rep(probs.array.aber.drasgow.item2, (max(C) + 1)
* (max(C) + 1)), c(N.nodes.chi, max(C) + 1, max(C) + 1, max(C) + 1))
    probs.array.aber.drasgow.item2.arr <- aperm(probs.array.aber.drasgow.item2.arr, c(1, 3, 2, 4))
    probs.array.aber.drasgow.item3.arr <- array(rep(probs.array.aber.drasgow.item3, (max(C) + 1)
* (max(C) + 1)), c(N.nodes.chi, max(C) + 1, max(C) + 1, max(C) + 1))
    probs.array.aber.drasgow.item3.arr <- aperm(probs.array.aber.drasgow.item3.arr, c(1, 4, 3, 2))
    weights.arr3                       <- array(rep(weights.arr, (max(C) + 1)^3), c(N.nodes.chi, max(C) +
1, max(C) + 1, max(C) + 1))
    expected.mat.drasgow.it1.it2.it3   <- N.NAs.t * apply(probs.array.aber.drasgow.item1.arr *
probs.array.aber.drasgow.item2.arr *
                                          probs.array.aber.drasgow.item3.arr * weights.arr3, 2:4,
sum)
    if (length(C) > 1)
    {
      if (C[item1] < max(C)) expected.mat.drasgow.it1.it2.it3[(C[item1] + 2):(max(C) + 1), , ] <-
NA
      if (C[item2] < max(C)) expected.mat.drasgow.it1.it2.it3[, (C[item2] + 2):(max(C) + 1), ] <-
NA
      if (C[item3] < max(C)) expected.mat.drasgow.it1.it2.it3[, , (C[item3] + 2):(max(C) + 1)] <-
NA
    }
    if (length(C) == 1)
    {
      observed.mat.drasgow.it1.it2.it3   <- table(factor(data[, item1], levels = 0:C), factor(data[,
item2], levels = 0:C), factor(data[, item3], levels = 0:C)) # table(data[, item1], data[, item2],
data[, item3])
    } else
    {
      observed.mat.drasgow.it1.it2.it3 <- array(NA, dim = c(max(C) + 1, max(C) + 1, max(C) +
1))
      observed.mat.drasgow.it1.it2.it3[1:(C[item1] + 1), 1:(C[item2] + 1), 1:(C[item3] + 1)] <-
        table(factor(data[, item1], levels = 0:C[item1]), factor(data[, item2], levels = 0:C[item2]),
factor(data[, item3], levels = 0:C[item3]))
    }
    # Merge cells with expected frequencies < 5:
```

```
    expected.order.it1.it2.it3            <- order(c(expected.mat.drasgow.it1.it2.it3))
    expected.mat.drasgow.it1.it2.it3      <-
c(expected.mat.drasgow.it1.it2.it3)[expected.order.it1.it2.it3]
    observed.mat.drasgow.it1.it2.it3      <-
c(observed.mat.drasgow.it1.it2.it3)[expected.order.it1.it2.it3]
    expected.mat.drasgow.it1.it2.it3.less5 <- expected.mat.drasgow.it1.it2.it3 < 5
    if (length(C) == 1) {df.it1.it2.it3 <- (C + 1)^3 - 1} else {df.it1.it2.it3 <- (C[item1] + 1) *
(C[item2] + 1) * (C[item3] + 1) - 1}
    if (sum(expected.mat.drasgow.it1.it2.it3.less5, na.rm = TRUE) > 0)
    {
      pos.sum <- max(which(expected.mat.drasgow.it1.it2.it3.less5 == 1))
      if (sum(expected.mat.drasgow.it1.it2.it3[1:pos.sum]) < 5) {pos.sum <- pos.sum + 1}
      expected.mat.drasgow.it1.it2.it3[pos.sum] <-
sum(expected.mat.drasgow.it1.it2.it3[1:pos.sum])
      expected.mat.drasgow.it1.it2.it3        <- expected.mat.drasgow.it1.it2.it3[pos.sum :
((max(C) + 1)^3)]
      observed.mat.drasgow.it1.it2.it3[pos.sum] <-
sum(observed.mat.drasgow.it1.it2.it3[1:pos.sum])
      observed.mat.drasgow.it1.it2.it3        <- observed.mat.drasgow.it1.it2.it3[pos.sum :
((max(C) + 1)^3)]
      if (length(C) == 1) {df.it1.it2.it3 <- (C + 1)^3 - pos.sum} else {df.it1.it2.it3 <- (C[item1] + 1)
* (C[item2] + 1) * (C[item3] + 1) - pos.sum}
    }
    # Compute (adjusted) chi squares (/df):
    chisq       <- sum(((observed.mat.drasgow.it1.it2.it3 - expected.mat.drasgow.it1.it2.it3)^2) /
expected.mat.drasgow.it1.it2.it3, na.rm = TRUE)
    chisq.df    <- chisq / df.it1.it2.it3
    chisq.adj   <- max(0, 3000 * (chisq - df.it1.it2.it3) / N.NAs.t + df.it1.it2.it3)
    chisq.adj.df <- chisq.adj / df.it1.it2.it3
    c(Item1 = item1, Item2 = item2, Item3 = item3, N = N.NAs.t, df = df.it1.it2.it3,
      chisq = chisq, chisq.df = chisq.df, chisq.adj = chisq.adj, chisq.adj.df = chisq.adj.df)
  }))
  triplets.res <- cbind(Triplet = 1:nrow(triplets), triplets.res)

  # Summarize results:
  f.int         <- function(x) {if (x < 1) 1 else (if (x < 2) 2 else (if (x < 3) 3 else (if (x < 4) 4 else (if
(x < 5) 5 else (if (x < 7) 6 else 7)))))}
  singlets.table <- c(table(factor(sapply(singlets.res[, 7], f.int), levels=1:7)),
round(mean(singlets.res[, 7]), 4), round(sd(singlets.res[, 7]), 4))
  doublets.table <- c(table(factor(sapply(doublets.res[, 9], f.int), levels=1:7)),
round(mean(doublets.res[, 9]), 4), round(sd(doublets.res[, 9]), 4))
  triplets.table <- c(table(factor(sapply(triplets.res[, 10], f.int), levels=1:7)),
round(mean(triplets.res[, 10]), 4), round(sd(triplets.res[, 10]), 4))
  all.table     <- rbind(singlets.table, doublets.table, triplets.table)
  rownames(all.table) <- c("Singlets", "Doublets", "Triplets")
```

```
  colnames(all.table) <- c("Less_1", "1_to_2", "2_to_3", "3_to_4",
"4_to_5","5_to_7","Larger_7", "Mean", "SD")

  res <- list(Singlets      = round(singlets.res, precision),
              Doublets      = round(doublets.res, precision),
              Triplets      = round(triplets.res, precision),
              Summary.table = round(all.table, precision))
  class(res) <- "MODFIT"
  return(res)
}

# Export data in MODFIT friendly format ----
Export.MODFIT <- function(data, C, IP, file.name = "MyData") {
  # Missing values: NA -> 9
  data[is.na(data)] <- 9
  write.xlsx2(data, paste0(file.name, "SCORES.xlsx"), col.names = FALSE, row.names =
FALSE)
  write.xlsx2(cbind(IP$alpha, IP$delta, IP$taus[, 1:C]), paste0(file.name, "IPs.xlsx"), col.names
= FALSE, row.names = FALSE)
}
```

Appendix D

R Code for Computing lz and lzstar Using GPCM

```
#Credit to Sandhip Sinharay who shared the following code via email to Jorge Tendeiro

# The function lzstatMx to compute lz and lzstar for a mixed-format test
  lzstarMix=function(scores,theta,a,b,c,est)
   {nitem=ncol(scores)
    mxscores=apply(b,1,f1)
    dich=which(mxscores %in% 1)
    poly=setdiff(1:nitem,dich)
    n3PL=length(dich)
    out=NULL
    for (j in 1:nrow(scores))
    {cnnum=0
    cnden=0
    s0num=0
    s0den=0
    if (n3PL>0)
    {pr=pr3PL(theta[j],a[dich],b[dich,1],c[dich])
    w=log(pr/(1-pr))
    e = exp(a[dich]*(theta[j]-b[dich,1]))
    p1 =  (1-c[dich])*a[dich]*e/((1+e)*(1+e))
    r = p1/(pr*(1-pr))
    cnden = sum(p1*r)
    cnnum = sum(p1*w)
    if (est=="wle") {p2 = p1*a[dich]*(1-e)/(1+e)
              s0num = sum(p2*r)
              s0den = 2*cnden}}
    if (n3PL<nitem)
    {for (i in poly)
    { probs=prGPCM(theta[j],a[i],b[i,1:mxscores[i]])
      der1=d1prGPCM(theta[j],a[i],b[i,1:mxscores[i]])
      cnnum=cnnum+sum(der1*log(probs))
      cnden=cnden+sum(der1**2/probs)}}
    cn=cnnum/cnden
    lznum=0
    lzden=0
    lzstden=0
    if (n3PL>0)
    {lznum=sum((scores[j,dich]-pr)*w)
    lzden=sum(w*w*pr*(1-pr))
    lzstden= sum(pr*(1.0-pr)*(w-cn*r)**2)}
    if (n3PL<nitem)
    {for (i in poly)
```

```
    { probs=prGPCM(theta[j],a[i],b[i,1:mxscores[i]])
      vmat=diag(probs)-probs%*%t(probs)
      lznum=lznum+log(probs[scores[j,i]+1])-sum(probs*log(probs))
      lzden=lzden+t(log(probs))%*%vmat%*%log(probs)
      der1=d1prGPCM(theta[j],a[i],b[i,1:mxscores[i]])
lzstden=lzstden+t(log(probs)-cn*der1/probs)%*%vmat%*%(log(probs)-cn*der1/probs)
      if (est=="wle")
         {scores=seq(0,mxscores[i])
          l1=sum(scores*probs)
          l2=sum(scores**2*probs)
          l3=sum(scores**3*probs)
          s0den=s0den+2*a[i]**2*(l2-l1**2)
          s0num=s0num+a[i]**3*(l3-3*l1*l2+2*l1**3)}}}
    lz = lznum/sqrt(lzden)
    lzstar = lznum/sqrt(lzstden)
    if (est=="wle") {lzstar = (lznum+cn*s0num/s0den)/sqrt(lzstden)}
    vrat=lzstden/lzden
    out=rbind(out,c(lz,lzstar))}
    return(out)}
# Some R functions required by lzstarMix are defined below
# Probability of a correct answer under the 3PL model
pr3PL=function(t,a,b,c){return(c+(1-c)/(1+exp(a*(b-t))))}
# Probabilities of different scores under GPCM
prGPCM=function(theta,a,b)#b for scores 1,2,...m for item scored on 0,1,...m.
  {probs=rep(exp(a*theta),length(b)+1)
   for (k in 2:length(probs))
     {probs[k]=exp(log(probs[k-1]) + a*(theta-b[k-1]))}
   return(probs/sum(probs))}
# Derivative of the P_ij for a GPCM item
d1prGPCM=function(theta,a,b)
{ probs=prGPCM(theta,a,b)
  scores=seq(0,length(b))
  return(a*probs*(scores-sum(scores*probs)))}
f1=function(x){return(sum(!is.na(x)))}
```

Appendix E

R Code for Computing lz_poly Using GGUM

```
# lzpoly.mixed under GGUM
#######Source: Tendeiro, Jorge. osf https://osf.io/jpmy2/##########

lzpoly.mixed <- function(matrix, IP, Ability, C)
{
  N <- dim(matrix)[1]
  I <- dim(matrix)[2]
  alpha <- IP$alpha
  delta <- IP$delta
  taus  <- IP$taus
  # Perfect response vectors allowed.
  P.CRF     <- array(NA, dim = c(N, I, max(C) + 1))
  for (z in 0:max(C)) {P.CRF[, , z + 1] <- GGUM:::P.GGUM(z, alpha, delta, taus, Ability, C)}
  log.P.CRF <- log(P.CRF)
  array.01  <- array(0, dim = c(N, I, max(C) + 1))
  for (n in 1:N) {array.01[n, , ][cbind(1:I, matrix[n, ] + 1)] <- 1}
  #
  l0p  <- apply(array.01 * log.P.CRF, 1, sum, na.rm = TRUE)
  El0p <- apply(P.CRF    * log.P.CRF, 1, sum, na.rm = TRUE)
  # Variance:
  var.arr <- array(NA, dim = c(N, I, max(C) + 1, max(C) + 1))
  for (z in 0:max(C))
  {
    P.CRF.slice          <- array(rep(P.CRF[, , z + 1], z + 1), dim = c(N, I, max(C) + 1))
    var.arr[, , , z + 1] <- P.CRF * P.CRF.slice * log.P.CRF * log(P.CRF / P.CRF.slice)
  }
  Vl0p <- apply(var.arr, 1, sum, na.rm = TRUE)
  res  <- (l0p - El0p) / sqrt(Vl0p)
  return(round(res, 4))
}
```

Appendix F

R Code for Computing lzstar_poly Using GGUM

```
# lzstarpoly.mixed from (https://www.jorgetendeiro.com/publication/tendeiro_2017/)

lzstarpoly.mixed <- function(matrix, IP, Ability, P.CRF, C)
{
 N <- nrow(matrix)
 I <- ncol(matrix)
 alpha <- IP$alpha
 delta <- IP$delta
 taus  <- IP$taus
 M<-2*C+1
 #   # Perfect response vectors allowed.
 P.CRF     <- array(NA, dim = c(N, I, C + 1))
 for (z in 0:C) {P.CRF[, , z + 1] <-  GGUM:::P.GGUM(z, alpha, delta, taus, Ability, C)}
 #
 d1P       <- dP.theta.arr(matrix, alpha, delta, Ability, taus, C)    # N x I x (C+1)
 ri        <- d1P / P.CRF
 array.01  <- array(0, dim = c(N, I, C + 1))
 for (n in 1:N) {array.01[n, , ][cbind(1:I, matrix[n, ] + 1)] <- 1}
 r0        <- -apply((array.01 - P.CRF) * ri, 1, sum, na.rm = TRUE)
 #
 wi        <- log(P.CRF)
 Wn        <- apply((array.01 - P.CRF)*wi, 1, sum, na.rm = TRUE)
 # Variance:
 var.arr <- array(NA, dim = c(N, I, C + 1, C + 1))
 for (z in 0:C)
 {
  P.CRF.slice          <- array(rep(P.CRF[, , z + 1], z + 1), dim = c(N, I, C + 1))
  var.arr[, , , z + 1] <- P.CRF * P.CRF.slice * wi * log(P.CRF / P.CRF.slice)
 }
 Vl0p     <- apply(var.arr, 1, sum, na.rm = TRUE)
 sigma2n  <- Vl0p / I
 cn       <- apply(d1P * wi, 1, sum, na.rm = TRUE) / apply(d1P * ri, 1, sum, na.rm = TRUE)
 wi.tilde <- wi - array(rep(cn, I * (C + 1)), dim = c(N, I, C + 1)) * ri
 #
 # Variance tau:
 V.tau <- c()
 D.arr <- array(NA, dim = c(N, I, C+1, C+1))
 tmp   <- P.CRF * (1 - P.CRF)
 for (z in 0:C)
 {
  P.CRF.2ndterm <- array(rep(P.CRF[, , z + 1], C+1), dim = c(N, I, C+1))
```

```
    D.arr[ , ,      , z + 1] <- - P.CRF * P.CRF.2ndterm
    D.arr[ , , z + 1, z + 1] <- tmp[, , z + 1]
  }
  for (n in 1:N)
  {
    sum <- 0
    for (i in 1:I)
    {
      sum <- sum + t(wi.tilde[n, i, ]) %*% D.arr[n, i, , ] %*% wi.tilde[n, i, ]
    }
    V.tau <- c(V.tau, sum)
  }
  tau2n    <- V.tau / I
  EWn      <- -cn * r0
  VWn      <- I * tau2n
  res      <- as.vector((Wn - EWn) / sqrt(VWn))
  return(round(res, 4))
}
```