USING MIXTURE, MULTI-PROCESS, AND OTHER MULTI-DIMENSIONAL IRT MODELS TO ACCOUNT FOR EXTREME AND MIDPOINT RESPONSE STYLE USE IN PERSONALITY ASSESSMENT

by

Michael J. Lucci

B.A., Saint Vincent College, 1986

M.A., University of Pittsburgh, 1988

M.A., University of Pittsburgh, 2003

Submitted to the Graduate Faculty of

the School of Education in partial f ulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Michael J. Lucci

It was defended on

November 3, 2017

and approved by

Clement Stone, PhD, Professor, Department of Psychology in Education

Suzanne Lane, PhD, Professor, Department of Psychology in Education

Feifei Ye, PhD, Assistant Professor, Department of Psychology in Education

Lauren Terhorst, PhD, Associate Professor, Department of Occupational Therapy

Dissertation Director: Clement Stone, PhD, Professor, Department of Psychology in Education

USING MIXTURE, MULTI-PROCESS, AND OTHER
MULTI-DIMENSIONAL IRT MODELS TO ACCOUNT FOR EXTREME AND
MIDPOINT RESPONSE STYLE USE IN PERSONALITY ASSESSMENT

Michael J. Lucci, PhD

University of Pittsburgh, 2017

The validity of interpreting questionnaire results is threatened by the possible overuse of extreme and midpoint response options. Since respondents may view the response options in different ways, accounting for midpoint (MRS) and extreme response style (ERS) use is important for accurate estimation of the latent trait. Biased sum scores provide poor trait estimates for two people with the same latent trait yet different response styles.

With the categorical view of response styles, respondents are seen as having a certain response style or not and are classified into different groups. The mixture graded response and mixture partial credit models were compared in this study. With the continuous view of response styles, respondents are seen as having varying degrees of different response style traits. A multidimensional model estimates substantive and response style trait levels for each person. A Multi-process model (M-PM) was used in this study to break down the response process into two and three subprocesses used in completing a five point Likert scale. The Multidimensional Partial Credit (MPCM) and Multidimensional Nominal Response (MNRM) models with substantive and response style scoring functions were also explored.

This study used an existing data set to investigate how the five different IRT models for addressing ERS and MRS performed for three different personality subscales (Anxiety, Openness to Experience Feelings, and Compliance) from the German version of Costa and McCrae's *NEO Personality Inventory-Revised*.

Each subscale illustrated different relationships with and uses of ERS and MRS traits. The response process traits of the M-PM differed from response style traits of the other models. The two and three class mixture models, the two and three dimensional MNRM and MPCM, and the two process model for intensity ERS and direction fit better than standard IRT models. ERS accounted for more item response variability than MRS. The MPCM is suggested to account for ERS and MRS due to the number of estimated parameters and amount of explained variability in item responses. The results are compared with each other and to results from a previous study. Limitations of this study and ideas for future research are presented.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Many thanks are expressed to Professor Clement Stone, my advisor and dissertation director, for everything you have done to help with my doctoral program and this project. You suggested the initial idea and many other important elements and revisions as the study developed and final document emerged. My sincere appreciation is also given to Professors Suzanne Lane, Feifei Ye, and Lauren Terhorst (other committee members) for your time, feedback, service, and helpful insights. The four of you are outstanding scholars and persons. Thank you so much for the privilege to be able to take courses with you and to work with the four of you. I am very grateful for all of our conversations and discussions.

You also understand very well how so many different things can happen in life as we complete our daily tasks. We have mourned the passing of Professor Kevin H. Kim and some of our family members. You have also shown much empathy to me with the various challenges that I have endured during my graduate program and I will always remember that.

A sincere, deep amount of heartfelt gratitude is expressed to Professor Fritz Ostendorf for allowing use of the data, which he and his late colleague Professor Alois Angleitner collected after writing the German version of the *NEO Personality Inventory-Revised*.

I also thank Professor Li Cai for providing much help with software guidance and understanding and estimating the multidimensional models with *flexMIRT*, Dr. Linda K. Muthén for answering questions regarding use of *Mplus* and mixture models, Professors Ulf Böckenholt, Daniel Bolt, David Thissen, and Eunike Wetzel for providing feedback about their research, Dr. Carl Falk for examples and suggestions, and the *flexMIRT* help desk for other software assistance and support. You made this project possible by the software programs and your technical support.

I thank my mother, Deanne R. Wargo Lucci, my brother, Mark, and sister, Maria, family, friends, and colleagues whose love, food, kindness, and support truly helped with completing and presenting these results in many very touching ways. In particular, I especially express much gratefulness to Gary D. Hart for his help with typing and formatting an initial draft of the text in Latex, generating some of the graphics, and help with Overleaf for the presentation pdf. I appreciate Lou Ann Sears for her thorough feedback regarding the bibliographic entries. I also thank Christy Kelsey Zigler for reading and commenting on a draft of the overview document. I thank Richard Hoover for help with Latex, the University of Pittsburgh librarians at the Greensburg and Oakland campuses, and the administrative assistants in the School of Education in Oakland and at the Pitt-Greensburg campus for their support. I am also appreciative of my helpful teaching assistants (Lauren, Virginia, Taylor, Josh, Shannon, Trey, Mickey, Alex, and Marcus) and other empathic students (Gina, Joe, Linda, Pam, Shirley, Tang, and others) that I have been blessed to work over the last several years. I thank Liz Marciniak, J. Wynn, and other classmates, friends, and other colleagues for their many kindnesses, prayers, and encouragement. You all have made this journey memorable for so many reasons.

I thank God, who has made all things possible.

This work is dedicated to the memory of my father, Oswald M. Lucci and to the memories of my grandparents, aunts, uncles, and friends who have passed. Your spirits live in all of the lives you have touched and you will always be in my heart.

1.0 INTRODUCTION

Since the use of questionnaires to determine a person's latent trait level is widespread in psychology and education, it is essential that the trait estimate be as accurate as possible. Traditionally with questionnaire use, the trait level is determined with Likert's method of summed ratings from the items (Likert, 1932). Unfortunately, the summed score can be biased when some respondents use certain response options more often than others. When a person tends to respond to a set of items independent of the item content across situations and time, a response style occurs (Jackson & Messick, 1958; Van Herk, Poortinga, & Verhallen, 2004). A person's response style is a trait indicated by overuse of certain response options and is independent of the latent trait being measured. The presence of substantial response style traits contaminates measurement of the desired latent trait.

To find improved trait estimates, psychologists can use item response theory models to account for response style use. Unlike the traditional method of using summed scores of item responses, IRT uses an estimation (search) process to find the most likely trait level that explains a person's responses (Embretson & Reise, 2000). An IRT model models the probability of choosing a response option as a function of the underlying latent trait (Van Vaerenbergh & Thomas, 2013). An IRT model is directly linked to the person's response behavior since it includes a parameter for the trait estimate and parameters to describe different aspects of the items (e.g. difficulty level). By adding parameters to basic IRT models, researchers have developed different types of models (e.g. multi-dimensional, mixture, and random thresholds) to address one or more response styles. The models address the response heterogeneity by producing different estimates (e.g. degree of response style trait, latent group specific parameters, or variable threshold parameters). Comparing how three multi-dimensional models: the multi-process model (M-PM), the Multi-dimensional Nominal Response model (MNRM),

and the mixture graded response model (mixGRM) find improved trait estimates for the respondents who tend to overuse midpoints or extreme categories is the focus of this study. These models are proposed to provide better fit to the data than the multi-dimensional partial credit model (MPCM) and mixture partial credit model (mixPCM).

1.1 STATEMENT OF PROBLEM

1.1.1 Response Styles, Why They May Occur, And Why They Matter

Response styles have been studied for decades by numerous researchers. Some of the commonly researched response styles appear in Table 1 (Van Vaerenbergh & Thomas, 2013). These are Acquiescence, Disacquiescence, Extreme, Mild (Nonextreme), Midpoint, Net Acquiescence, Noncontingent, and Response Range (RR) responding. The table also indicates the consequences (i.e., how item statistics are distorted) if the response style use is present in the dataset.

Acquiescent, extreme, and midpoint responding are three commonly studied response styles in cross cultural and personality research due to their adverse effects on item and scale statistics (Baumgartner & Steenkamp, 2001; Chen, Lee, & Stevenson, 1995; A. Harzing, 2006; Hoffmann, Mai, & Cristescu, 2013; Hui & Triandis, 1989; Van Herk et al., 2004). Acquiescence response style (ARS) is the tendency to agree with items, regardless of content. ARS increases observed item means. Extreme response style (ERS) is the tendency to use one or both extreme options, regardless of content. ERS leads to an increase (decrease) in observed item means if the highest (lowest) extreme option is selected. Midpoint response style (MRS) is the tendency to overuse the middle category and this brings observed item means closer to the midpoint.

In some studies, response style use has also occurred due to the mode of survey administration. Jordan, Marcus, and Reeder (1980) found that respondents tended to omit responses or to give extreme or acquiescent responses to a higher degree when asked questions by telephone rather than in person. The telephone interviewer may not probe as deeply

Table 1: Definitions and Consequences of Common Response Styles

Response Style (RS)		Definition	Consequences
ARS:	Acquiescence	Tendency to agree with items, regardless of content	IOM, IMMVR
DRS:	Disacquiescence	Tendency to disagree with items, regardless of content	DOM, IMMVR
ERS:	Extreme	Tendency to use lowest or highest categories, regardless of content	DOM, IOM, IV, DMMVR
MLRS:	Mild (Nonextreme)	Tendency to avoid using extreme categories	BOM, DV, IMMVR
MRS:	Midpoint	Tendency to use the middle category, regardless of content	BOM, DV, IMMVR
NARS:	Net Acquiescence	Tendency to show more acquiescence than disacquiescence	IV, DOM if negative
NCR:	Noncontingent responding	Tendency to answer randomly, nonpurposefully, or carelessly	None can be specified a priori
RR:	Response Range (Standard Deviation)	Tendency to use wide or narrow category range around mean response	If large: IV, DMMVR

Note: IOM = Increases observed means, IMMVR = Increases magnitude of multivariate relationships, DV = Decreases variance, DMMVR = Decreases magnitude of multivariate relationships, IV = Increases variance, BOM = Brings observed means closer to midpoint, DOM = Decreases observed means. Adapted from Van Vaerenbergh and Thomas (2013).

as an interviewer in a face-to-face situation. Weijters, Schillewaert, and Geuens (2008) found that persons completing a survey by phone made less use of the midpoint option and more use of acquiescent responses than persons completing the survey online or in paper format. Persons in the online group were less likely to use extreme or disagree responses than persons in the paper-pencil and telephone groups. When the group means were examined without accounting for response styles, the groups appeared to differ in consumer trust of retail employees. The group differences in the trust measure were not significant when response style use was addressed.

Other researchers have identified how differences in ethnicity, culture, gender, education, or age invoke response style use. A study by A. Harzing (2006) revealed that students from Spanish speaking countries gave high levels of extreme responses while students from East Asian countries gave high levels of midpoint responses. Ayidiya and McClendon (1990) found minority groups had a tendency to agree or to give extreme responses. A. Harzing (2006) found females tended to give more midpoint responses while males tended to give more extreme responses, but that age did not affect response style use. In contrast, Weijters, Geuens, and Schillewaert (2010b) found that older persons gave more extreme and midpoint responses than younger persons and that females gave higher levels of extreme responses than males. Their study also showed that respondents with low education levels gave high levels of extreme and midpoint responses.

Other studies have found relationships between response styles and personality variables. Hamilton (1968) found that individuals with higher anxiety levels used extreme options more than less anxious persons. Austin, Deary, and Egan (2006) found that use of extreme responses had a positive correlation with extraversion and conscientiousness. Wetzel and Carstensen (2015) however found negligible correlations between extreme responses and extraversion and conscientiousness traits. They found small to moderate negative correlations between use of midpoints and openness to Experience Feelings fantasy and openness to Experience Feelings and between use of extreme options and modesty. Wetzel and Carstensen (2015) also found that most of the personality facets that they examined had stronger correlations with acquiesence, disacquience, or both ARS and DRS than with extreme or midpoint responding.

Thus, these studies indicate that many factors such as mode of administration and respondent demographic or personality variables may contribute to response style use. As indicated in Table 1, the use of response styles contributes to the distortion of item statistics and properties. In addition to adversely influencing the item means, any response style use, particularly extreme and midpoint responding, can affect the mean and variance of the summed scores and any multivariate relationships such as correlations between items. Finally, interpretation of group means in aggregate-level analyses (Greenleaf, 1992b; A. Harzing, 2006) and strength of measurement invariance (Wetzel, 2013) can be impacted.

The differential use of response styles can also increase the dimensionality of the measurement process (Johnson & Bolt, 2010; Rost, 2004; von Davier & Khorramdel, 2013). This implies that use of response style traits is measured in addition to the trait of interest and the instrument can fail to be unidimensional. This causes measurement problems since using Likert's method for a particular scale depends upon the scale items measuring one underlying construct (Wu & Huang, 2010). With any type of response style use present, using the summed score does not provide an accurate estimate of the desired trait. Therefore, determining the trait estimate in another way is needed if response style use has been detected.

1.1.2 Methods to Deal With Response Styles

Due to the problems caused by response styles, many methods have been developed to account for response style use. Van Vaerenbergh and Thomas (2013) provide an extensive list and a concise review. For example, some simple methods to detect use of particular response styles include determining the proportions of responses in the relevant item categories (e.g., extremes and midpoints to assess Extreme and Midpoint Responding). ERS can also be assessed indirectly by Response Range (RR) which is the standard deviation of an individual's responses across a set of items. ERS and RR may be related but are different measures since RR reflects the narrowness or broadness of a person's response pattern. Small values of RR only imply that a narrow range of responses is used (Peterson, Rhi-Perez, & Albaum, 2014) and not necessarily little ERS. Persons could still tend to use some extreme categories. If

ERS and RR are highly correlated, they can be averaged to form an overall ERS measure for inclusion in a model to detect and account for response styles (Baumgartner & Steenkamp, 2001). To assess acquiescent response style (ARS), the amount of agreement with positively and negatively worded items (before reverse-scoring of negatively worded items) in the same scale can be found. ARS is also assessed by finding the amount of agreement with many heterogeneous items over several unrelated scales (Baumgartner & Steenkamp, 2001; Martin, 1964).

To account for response style use, the response style measures can be used as covariates in analysis of covariance or linear regression models (cf., Greenleaf, 1992a; Reynolds & Smith, 2010). These models can be used to illustrate the importance of addressing response style use. Greenleaf (1992a) found that standard deviation (RR) bias affected the classification of persons into marketing segments by response patterns. When the bias was removed, with the adjustment for the response styles, the composition of the segments changed. The mean age increased and mean education levels decreased. Diamantopoulos, Reynolds, and Simintiras (2006) and Reynolds and Smith (2010) found that conclusions to cross-cultural group comparisons can be altered by including one or more response style measures in the analysis. There were less significant differences between cross-cultural groups on the substantive traits (Interpersonal Influence Susceptibility, Self-Esteem, and Service Quality Aspects) when response style effects were removed.

More complex methods involving multilevel regression or correlated factor structure models have also been developed (e.g.,Baumgartner & Steenkamp, 2001; Weijters, Geuens, & Schillewaert, 2010a; Weijters, Schillewaert, & Geuens, 2008). These models use additional heterogeneous items which serve as indicators for use of different response styles. The extra items are used to create the simple measures (such as sums or proportions) which estimate a person's degree of a particular response style use. The part of variance shared by items due to response styles is removed so that only shared content variance remains. For example, Weijters et al. (2008) illustrated using extra items to account for four different response styles (ARS, DRS, ERS, MRS) in a means and covariance structure model. The response style measurement model improved the latent trait estimate by correcting for the bias that occurred in the factor model which did not address response styles. Using a multilevel model,

Baumgartner and Steenkamp (2001) examined the influence of five response styles on 14 different scales in 11 countries. On average, noncontingent (careless, random) responding did not bias scale scores systematically, but ERS and MRS did affect variation in scale scores, particularly when the scale mean (on the response scale) differed the most from the scale midpoint. Additionally, the study found that using balanced scales helped to effectively account for 60-62% of the variance due to ARS and DRS. Balanced scales consisted of pairs of items measuring the same content yet one is negatively worded and the other is positively worded.

Researchers (e.g., De Jong, Steenkamp, Fox, & Baumgartner, 2008; Khorramdel & von Davier, 2014; Van Vaerenbergh & Thomas, 2013; Zettler, Lang, Hülsheger, & Hilbig, 2015) have noted disadvantages of these methods:

(a) Simple methods do not account for influence of a substantive trait. Some items possibly measure both the trait and response style (Khorramdel & von Davier, 2014).

(b) Complex methods may require adding items unrelated to the desired construct. These extra items may help to measure response styles; however, they lengthen the survey which increases respondent time to complete the items. This can lead to nonresponse due to fatigue. Additionally, the items may be difficult to find (Khorramdel & von Davier, 2014; Van Vaerenbergh & Thomas, 2013).

(c) The methods may have not been validated to show response style use is actually measured. For example, a sum of extreme responses to correct for extreme response style, might not be valid if the summed extreme score does not provide a unidimensional measure of extreme response (Khorramdel & von Davier, 2014). Scores from response style indicators may be assumed to measure response styles yet the items may not have been tested for unidimensionality and reliability. Many studies have not reported both results of factor and reliability analyses to support use of the extra items as response style indicators.

(d) The sum score (count) method for detecting ERS (or ARS, MRS) does not separate person and item effects since the sum score method gives equal weights to all items (De Jong et al., 2008). Persons are different in their tendencies to use different categories and items can evoke response styles to different degrees.

(e) The methods do not attempt to explain how persons select a specific category during the response process (Zettler et al., 2015). While the models may fit well statistically to the data, the model were not developed to link response given and test behavior.

Fortunately, Item Response Theory (IRT) methods exist to overcome many of these limitations. IRT methods provide a model to account for the influence of the desired substantive trait on the responses. The models, such as the multi-process IRT model (M-PM), do not require additional data to be collected (Plieninger & Meiser, 2014) as do methods that use heterogeneous items (e.g. Weijters, Schillewaert, & Geuens, 2008).

IRT methods are useful since they provide a way to address the multi-dimensionality arising from response style use. Thus, in addition to the trait of interest, the item response models are used to address different patterns in use of the response scale. To account for the heterogeneous response scale use and address ERS or MRS, many different kinds of models have been used such as multidimensional IRT (MIRT) models, mixture IRT models, models for random thresholds or models with a person parameter that affects the thresholds.

With random threshold parameter models, an existing IRT model is supplemented with "threshold related" parameters which reflect individual use (shrinkage or expansion) of the response scale and account for response styles. For example, Johnson (2003) extended the graded response model (Samejima, 1969) to include a symmetric vector of threshold parameters representing differences from a central point (Thissen-Roe & Thissen, 2013). The threshhold parameters define the lower and upper bounds (of the midpoint, if present) relative to the central location, and separate remaining categories by their order of extremity. The Proportional Threshold Model (Rossi, Gilula, & Allenby, 2001) is a related model that demands that the thresholds across persons be proportional. Model output yields a vector of thresholds applying to all persons and items and a person specific scale parameter which shrinks or expands the response scale.

Using a different approach, Jin and Wang (2014) extended the partial credit model (PCM, Masters, 1982) so that a person-specific weight parameter is added to the thresholds. This weight parameter accounts for a person's tendency to use ERS. Two limitations of these models are that they do not account for other types of response styles (Jin & Wang,

2014; Johnson, 2003) and they do not provide a conceptual idea about how persons choose a particular response option (Zettler et al., 2015). This latter limitation can be overcome with the M-PM, a MIRT model, described below.

With a MIRT model, response styles are viewed as continuous latent random variables that are distributed along their own trait dimensions. Each person shows response style traits to different degrees. The MIRT model provides an estimate for the latent trait of interest and for any response style trait addressed by the model. One example is a multidimensional nominal response model for ERS (Bolt & Johnson, 2009; Johnson & Bolt, 2010). This model was extended by Falk and Cai (2015) to address other response styles such as ARS, MRS, and SDR. Another example of a MIRT model is the M-PM (Böckenholt, 2012). The M-PM models the distinct processes in which a person engages when completing the items.

With a mixture IRT model, response styles are studied using a categorical latent variable. Each person has a set of probabilities that indicate the likelihood of being assigned to particular response style groups. Using the maximum probability, the mixture IRT model classifies each person into a group representing those persons with a specific response style and provides class-specific item parameters which reflect different item characteristics for the class. Each group uses the response scale differently and is inferred to reflect a certain response behavior (Zettler et al., 2015).

One example of a mixture IRT model is the mixture graded response model (Sawatzky, Ratner, Kopec, & Zumbo, 2012). Comparison of using the multi-process model, multidimensional PCM, multidimensional nominal response model, mixture PCM, and mixture GRM to provide trait estimates while accounting for ERS and MRS was the focus of this study since no such study has been done.

1.1.3 Multi-Process Modeling and Mixture Modeling of Response Styles

The Multi-Process IRT model (M-PM) is an example of a noncompensatory model. In a noncompensatory MIRT model, a unidimensional model is used for each separate trait needed to complete a questionnaire item. The product of the probabilities from the separate models gives the probability of a particular response. This implies that the probability of a

response is no higher than the largest probability for a given trait. Thus, the compensation of a high trait value for a low trait value is reduced. This differs from a compensatory MIRT model where traits combine additively. A high value on one trait can compensate for a low value on another trait. A MIRT model provides a profile of scores for each person. For example, the M-PM breaks down the response process into a series of subprocesses. Each score indicates a person's specific trait level or tendency to use the related process. The number of subprocesses modeled depends upon the number of response options and determines the number of estimated trait scores.

For a questionnaire item with J response options, there are at most $J-1$ processes, since often a fewer number can be used. For example, Plieninger and Meiser (2014) used a four-process model to analyze seven-point response format data. For the four successive processes, there is a tree-like structure (as in Figure 1). The four processes can be summarized as indifference (1, use of midpoint or not), direction (2, agree or disagree), intensity (3, extreme or not), and central tendency (4, somewhat agree/disagree or just agree/disagree). Figure 1 shows how the response of person n to an item i with a seven-point response scale can be modeled using these four subprocesses. The tree-like structure shows that the processes are sequential and nested. See Table 2 for a brief explanation of the response process.

The M-PM attempts to explain how individuals differ in the processes. Analysis of the model yields a set of person trait estimates for each process and a set of model parameter estimates for each item. The probability of a particular response to an item can be determined by computing a probability of activating each process and then multiplying these probabilities. This model has been effective in accounting for use of response styles in measurement of personality and other traits (e.g., Böckenholt, 2012; Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013).

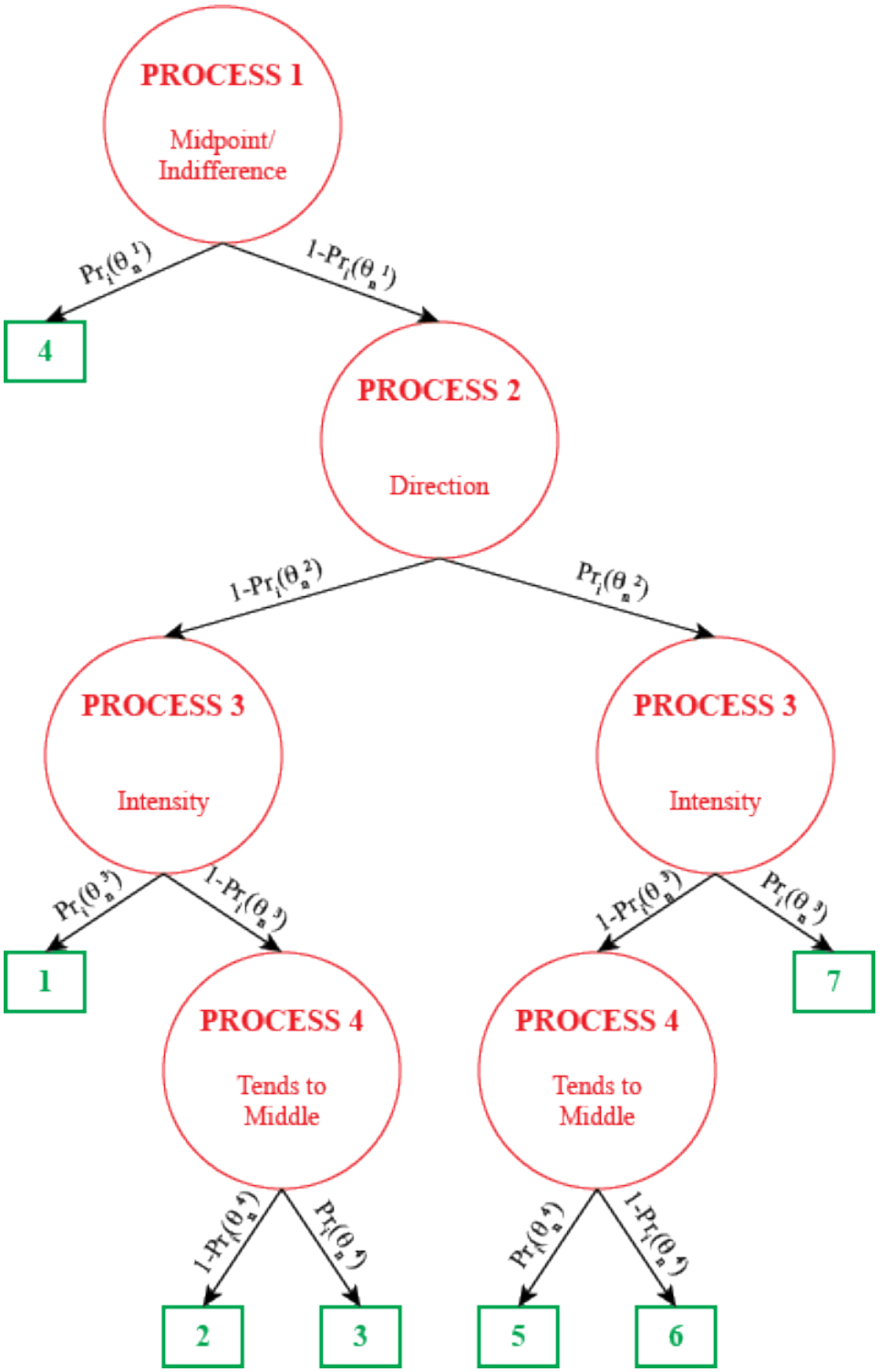
Use of a multi-process or other MIRT model is one way to account for response styles; a second way is to use a mixture IRT model. In a mixture IRT model, unknown population heterogeneity is explained by a categorical latent variable and the covariation of observed data within class is explained by continuous latent factors (G. Lubke & Neale, 2008; G. H. Lubke & Muthén, 2005; L. K. Muthén & Muthén, 1998-2012). This model involves a set of unobserved latent classes (subpopulations) and an IRT model for each class. The classes are

Table 2: Use of Four Latent Processes with a Seven-Point Response format

	Process	Description
1	Indifference	If a person does not have a distinct opinion about a given item's content, the person selects the middle category and the response process ends. The other processes are not invoked for the given item.
2	Direction	If the person has a well-defined opinion about the item content, the person chooses to agree or disagree with it.
3	Intensity	To express how strongly the opinion is held, the person chooses to select an extreme option or not.
4	Tendency to the Middle	If an extreme option is not chosen, then the person chooses to lean toward the midpoint or not.

Note: Although other interpretations for selecting categories are possible, the ideal interpretation here is that a person is honestly completing the questionnaire items by engaging in the four unique processes to different degrees. See Figure 1.

Figure 1: Tree-like structure of the four nested, sequential processes



Note: Four unobserved processes that are used to respond to a seven-point item. 1 = Completely Disagree, 2 = Disagree, 3 = Somewhat Disagree, 4 = Neutral, 5 = Somewhat Agree, 6 = Agree, 7 = Completely Agree, $Pr_i(\theta_n^h)$ = Probability person n with trait level θ_n^h uses process h to respond to item i. See Table 2.

not formed based on an observed variable (e.g., gender or age), but on a latent categorical variable. Different item response function parameters can exist for each class (Cho, 2013; G. Lubke & Neale, 2008; Rost, 1990; Sawatzky et al., 2012).

Polytomous mixture IRT models have been used in accounting for individual differences in rating scale use (Austin et al., 2006; Egberink, Meijer, & Veldkamp, 2010; Rost, 1991; Rost, Carstensen, & von Davier, 1997). The specific number of K latent classes is identified and class specific model parameters and latent trait person scores are estimated. Researchers have used such models to improve scaling of persons on physical health, personality, and other trait measures (e.g., Rost, Carstensen, & von Davier, 1997, Sawatzky, Ratner, Kopec, & Zumbo, 2012,Wetzel, Carstensen, & Böhnke, 2013).

1.2 PURPOSE OF STUDY

The purpose of this study was to compare how three types of IRT models: mixture models (mixPCM, mixGRM), multidimensional partial credit (MPCM) and nominal response (MNRM) models , and the multi-process model (M-PM), account for extreme and midpoint response styles in different personality domains. This study used a dataset which previously has been analyzed. The data set consists of responses to the 30 facets of the German version of the *NEO Personality Inventory-Revised (NEO PI-R)* (Ostendorf & Angleitner, 2004). The *NEO PI* was originally developed by Costa and McCrae (1992). There are 240 items measuring five dimensions of personality (Neuroticism, Extraversion, Openness to Experience Feelings, Agreeableness, and Conscientiousness). Each personality dimension has six lower-order facets measured by eight items with a five point Likert scale. Respondents rate themselves with the categories: 1 = strong disagreement, 2 = disagreement, 3 = neutral, 4 = agreement, 5 = strong agreement.

Since response style use can depend upon the personality trait measured, three of the 30 lower order facet subscales were chosen to illustrate and compare use of the IRT models in this study. The three facet scales were chosen since they were hypothesized to reflect use of ERS and MRS to different degrees. With two of the facets, persons were hypothesized

to exhibit use of either extreme or midpoint categories to a higher extent than for the other facets. One of the facets was assumed not illustrate a high use of ERS or MRS. Thus, the differences in strength of the relationship between the personality trait of interest and response style traits can provide diverse situations to compare how the three types of IRT models account for use of extreme and midpoint response styles.

1.3 SIGNIFICANCE / JUSTIFICATION OF STUDY

The main reason for conducting this study was that there are a limited number of studies with the multi-process model (M-PM). The M-PM is a response process model. This is believed to be useful since understanding what a model can say about the response process leads to more powerful uses of a model than just describing persons and data with parameters (Andrich, 1995). This is important since as expressed by Samejima (1979), a mathematical model's key role in psychology is to reasonably denote psychological reality (Ostini & Nering, 2006; Samejima, 1979). The M-PM hypothesizes that respondents make a series of decisions in selecting response options. The trait estimates work in a noncompensatory fashion. The multidimensional PCM does not hypothesize a response process. The trait estimates work together in a compensatory fashion.

There have been no studies comparing the M-PM with other multidimensional models such as the multidimensional partial credit model (MPCM) and multidimensional nominal response model (MNRM). This study examined how the M-PM, MPCM, and MNRM address ERS and MRS in an existing data set. Although the data have been previously analyzed with the mixture partial credit model(mixPCM) and MPCM (Wetzel, 2013), they have not been analyzed with the M-PM, the MNRM, and the mixture graded response model (mixGRM). Comparison of different IRT models to the same data allows measurement specialists to compare relative advantages of one model over another prior to choosing one to improve measurement (Swaminathan, Hambleton, & Rogers, 2007).

The second and third main reasons for this study were limited research on the mixture graded response model (mixGRM) and lack of research comparing the multi-process model (M-PM) with the mixGRM. In the literature with mixture models in questionnaire research, most of the studies used mixed Rasch models (e.g., Austin, Deary, & Egan, 2006; Eid & Raubner, 2000; Rost, 1991; Wetzel, 2013). Only a few studies that used the mixture graded response model exist (e.g., Egberink, Meijer, & Veldkamp, 2010; Sawatzky, Ratner, Kopec, & Zumbo, 2012). The current study contributes to the literature by illustrating the use of the mixture graded response model to account for extreme and midpoint response styles.

One recent study compared a multi-process model with a mixture partial credit model (Böckenholt & Meiser, 2017); however, no study which examined trait estimates from a multi-process model along with those from the mixture graded response models existed. This study contributes to the literature by providing a comparison between use of the multi-process model (M-PM), the MPCM, and MNRM, and use of the mixture models (mixGRM, mixPCM) to obtain trait estimates from the same data.

1.4 RESEARCH QUESTIONS

General Research Question A: Does modeling response styles with mixture, multidimensional, and multi-process models improve model-data fit for scales exhibiting Extreme (ERS) or Midpoint Response style (MRS) over the standard IRT models (Partial Credit (PCM) and Graded Response (GRM))?

It was hypothesized that for the facets showing the presence of MRS or ERS, the mixture Partial Credit and Graded Response models, the multi-process model, and multi-dimensional PCM and NRM would improve fit of the model to the data over the GRM and PCM.

General Research Question B1: How do the estimated latent correlations between the substantive and response style traits for each of the multidimensional IRT models and M-PM compare?

It was hypothesized that the facets will show different correlations between the latent substantive and response style trait estimates since the personality scales were chosen to illustrate differences in response style effects. Wetzel and Carstensen (2015) illustrated that traits such as Compliance and Openness to Experience Feelings had low latent correlations with either midpoint or extreme response style traits for the Multi-dimensional Partial Credit Model (MPCM). Traits such as Anxiety, Assertiveness, and Deliberation had negligible correlations with MRS and ERS traits. The Multi-Process Model may show correlations that are different from the MPCM since it is a response process model.

General Research Question B2: How do correlations between latent trait estimates based on the different IRT models compare with each other?

It was hypothesized that the models will provide substantive trait estimates that will correlate since the models account for response styles. The estimates are not expected to correlate perfectly since differences in estimates are likely due to the ways the models account for response styles.

General Research Question C: Which model, the mixture model (mixPCM or mixGRM), a multi-dimensional PCM or multi-dimensional NRM, or the multi-process model (M-PM), is best for addressing extreme and midpoint response styles?

The mixture models, the MPCM, the MNRM, and M-PM have not been directly compared in any previous study. By examining model output, and measures of fit, some conclusions and practical suggestions can be made regarding the scales examined.

2.0 LITERATURE REVIEW

In this chapter, the relevant background literature is presented. The chapter begins with a discussion of the two parameter logistic model and graded response model since these unidimensional IRT models are related to the models which are compared in the current study. IRT models are useful in test development for many assessment purposes (e.g. ability testing, test equating, performance assessment, and professional licensure or certification). Making distinctions among persons is also important in the areas of attitude and personality assessment where test developers also use IRT models. While some inventories contain dichotomously scored items (e.g., *Minnesota Multiphasic Personality Inventory-2*, Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), others contain polytomously scored items (e.g., *NEO Personality Inventory*, Costa & McCrae, 1992). The graded response model is useful in analyzing data for these latter cases due to the ordered response categories.

After the unidimensional models, their related assumptions, and parameters are discussed, the equivalence of IRT and CFA models is described since several examples in the literature are discussed in a CFA context. A brief summary of multidimensional IRT models follows this.

A concise discussion of survey research and response bias is then presented. This is followed by a discussion of response styles and methods to account for them. The IRT models to be compared in this study are then presented. First the multi-process model and some examples of using this model to account for response styles are described. Afterward the mixture IRT model is presented and studies which have used mixture IRT to account for response styles are then summarized.

2.1 ITEM RESPONSE THEORY MODELS AND ASSUMPTIONS

There are IRT models for both dichotomous and polytomous scoring of items used to measure one trait or many traits. To use an IRT model to analyze a set of data, there are at least two important assumptions about the data: dimensionality of the latent space and local independence (Hambleton & Swaminathan, 1985). The dimensionality of the latent space refers to the number of traits needed to describe a person's item responses. Usually a test is designed to measure one dominant trait even though there may be other secondary traits which influence examinee performance to a lesser degree. Unidimensional IRT models assume that one latent trait accounts for the performance and item responses. Local independence implies that given a specific trait level (i.e., controlling for trait level), the response to one item is not related to the reponse to another item (Embretson & Reise, 2000). These assumptions are interdependent since a set of data is unidimensional if item responses are locally independent based on one trait (Embretson & Reise, 2000).

Another assumption must also be made about the form of the item response model or item characteristic curve (ICC) used to predict the responses (given the person's trait level). This assumption depends on how the item is scored. For a dichotomous item, there is one monotone, increasing ICC. The ICC relates the examinees' performance on an item to the trait that underlies the performance (Hambleton, Swaminathan, & Rogers, 1991). With a polytomous item, at least one non-monotone ICC is needed with the monotone ICC. Although many possible curves and models exist, those relevant to this study are described next.

2.1.1 Unidimensional Models for binary scored items

For a test item that is dichotomously scored, there are two common IRT models. The one parameter model is based on the item difficulty and person trait level only. For the logistic form of the one parameter logistic model (1PLM), consider a randomly chosen person n with trait level θ_n responding to item i with difficulty parameter b_i. The probability of responding correctly with response $j = 1$ to the item can be expressed as in equation 2.1:

$$P_i(j = 1|\theta_n) = \frac{e^{a(\theta_n - b_i)}}{1 + e^{a(\theta_n - b_i)}} \tag{2.1}$$

where e is the base of the natural logarithm function. The a parameter is a slope or discrimination parameter that weights the difference between θ_n and b_i. For this form of the 1PLM, this slope is estimated and assumed to be common for all items.

In the model, the difficulty parameter, b_i, represents the point on the ability scale where a randomly chosen person with this trait level has probability of 0.5 in getting the item correct (Hambleton et al., 1991). Note that this point can also refer to the probability of endorsing a questionnaire item with two response options (e.g., agree, disagree). Persons with trait levels θ_n above (below) b_i endorse the item with probability greater (less) than 0.5. In the 1PLM, items are described by the different values of the difficulty parameter alone. The trait level of a person can be understood as a "threshold level" for item difficulty, since it correponds to the item difficulty level where a person is equally likely to endorse or not endorse such an item (Embretson & Reise, 2000).

In the 1PLM model, the items are seen as equivalent indicators for measuring the person's trait level and thus, the a parameter is constant. If the test items are seen as being unequally related to the trait level, then an item-level discrimination parameter, a_i, is added to equation 2.1 to give the two parameter logistic (2PLM) model (Embretson & Reise, 2000). This is shown in equation 2.2:

$$P_i(j = 1|\theta_n) = \frac{e^{a_i(\theta_n - b_i)}}{1 + e^{a_i(\theta_n - b_i)}} \tag{2.2}$$

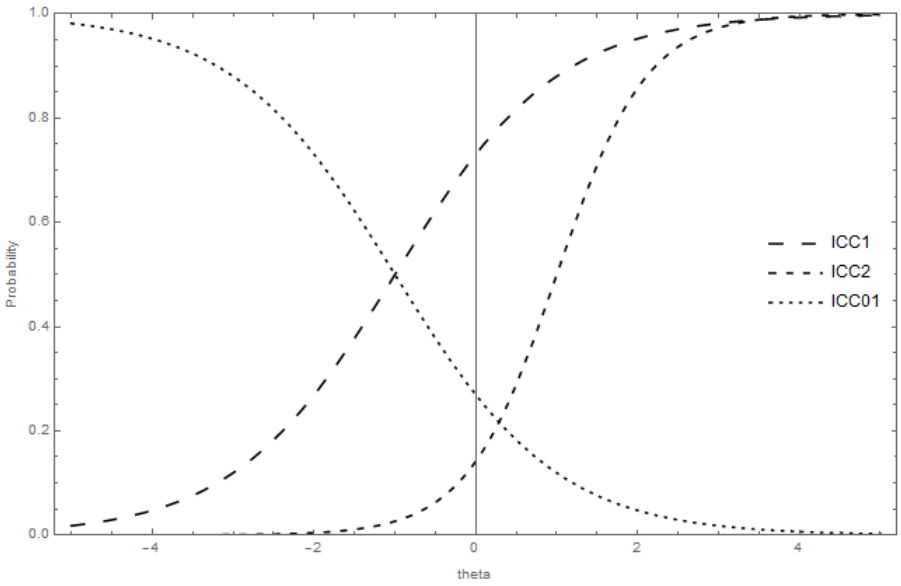
In this model, the discrimination parameter, a_i, indicates how the items differ in how well they distinguish between persons with different trait levels. Unlike the difficulty parameter which can be negative, the discrimination parameter of an item should be positive for the item to be useful. For a fixed trait level θ, the slope parameter can be viewed as an indicator of how well the persons near this trait levels are placed into the group of persons with trait estimates above the fixed θ level or in the group with trait level at or below the fixed θ level (Hambleton et al., 1991). Larger values of the slope parameter indicate items which are better at making distinctions between persons in a narrow range of trait levels while items with moderate slope values are better for distinguishing person performance over a wider range of trait levels.

A graph of the probability of getting a dichotomous item correct or endorsing the item is a monotonically increasing curve as the trait level increases. Usually only the probability of endorsing the item (selecting the positive category) is modeled since the probability function for not endorsing the item is easily found due to the complementary nature of the category functions (Ostini & Nering, 2006).

A graph showing the ICCs for two different items is presented in Figure 2. In the figure, the first curve rising from left to right ($ICC1$) represents the probability that a randomly chosen person with the corresponding trait level endorses item 1. In the figure, the curve that falls from left to right ($ICC01$) represents the complementary probability that such a randomly chosen person does *not* endorse item 1.

The two curves rising from left to right, $ICC1$ and $ICC2$, illustrate how the slope and difficulty parameters for the two items differ. The item on the left presents a less difficult and less discriminating item than the item on the right. These differences can also be described by the parameters for the items. The first item has $a_1 = 1.0$, $b_1 = -1.0$; while the second item has $a_2 = 1.8$, $b_2 = 1.0$.

Figure 2: Item Characteristic Curves for two items



Note: ICC1 $= P_1(j = 1|\theta)$, ICC2 $= P_2(j = 1|\theta)$, ICC01 $= P_1(j = 0|\theta)$.
See text for definitions of curves and parameters.

The item location is determined by the difficulty parameter b_i. As can be seen in Figure 2, the location (Inflection point) of the first (left) item is in the lower portion of the θ range while the location for the more difficult item on the right side is in the upper portion of the θ range which implies that a higher trait level is needed to solve or endorse the item to the right. What also can be seen in the figure is how the a_i parameter for the item on the right is larger than the item on the left. Thus, this second item is more discriminating than the first.

2.1.2 Graded Response Model

For items with an ordered response scale format with more than two categories, polytomous IRT models have been developed. Although many polytomous models allow the number of response categories to differ for each item, the model described here assumes the same number of response categories for each item since the number of response categories is often constant for questionnaire items. The number of categories here is J and are labeled $j = 0, 1, \cdots, M$. The $J = M + 1$ categories have M boundaries or thresholds between the categories.

In the graded response model (Ostini & Nering, 2006; Samejima, 1969), the probability of a person responding positively at a category boundary, given all previous categories, is modeled with a 2PLM model. If b_{ij} represents a category boundary parameter, then the probability for person n with trait level θ_n responding with j' in or above category j is given by P^* in equation 2.3:

$$P^*_{ij}(j' \geq j|\theta_n) = \frac{e^{a_i(\theta_n - b_{ij})}}{1 + e^{a_i(\theta_n - b_{ij})}} \tag{2.3}$$

The b_{ij} represents a category "difficulty level" and indicates the trait level needed to respond in or above threshold j (i.e. beyond category $j-1$) with probability 0.5. There are equal category slope parameters, a_i, within an item (Embretson & Reise, 2000). This views the item as a series of $M = J-1$ dichotomies (0 vs. $1, 2, \cdots, M$; 0, 1 vs. $2, 3, \cdots, M$, etc.).

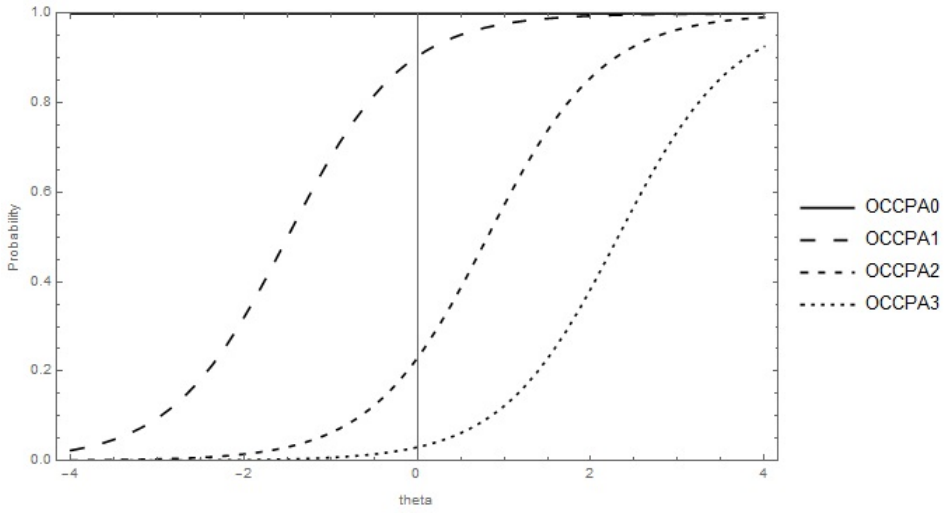
The $P^*_{ij}(\theta)$ is known as a cumulative category response function. The graph of the probability functions are known as Operating Characteristic Curves (OCC). Note that $P^*_{i0} = 1$ and that $P^*_{iM} = 0$ for any item since a person is assumed to choose any one of the categories.

The probability of endorsing specific category, $j' = j$, depends upon the probabilities of endorsing the previous categories. This probability is denoted by $P_{ij}(\theta_n)$ and is known as the category response function. Its graph is a category response curve (CRC). The probability for endorsing a specific category, $j' = j$, is given by $P_{ij}(j' = j|\theta_n)$ as in equation 2.4:

$$P_{ij}(j' = j|\theta_n) = P^*_{ij}(j' \geq j) - P^*_{i(j+1)}(j' \geq j + 1) \tag{2.4}$$

for a person with trait level θ_n.

Figure 3: Operating Characteristic Curves for GRM

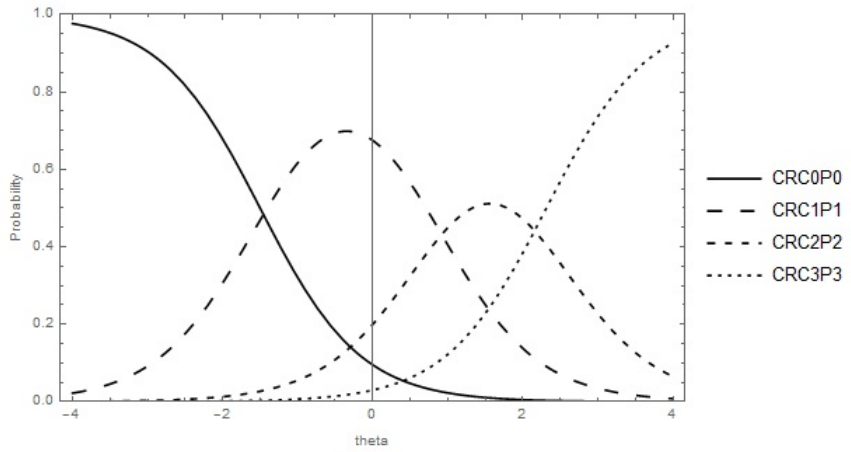


Note: OCCPA0 = P^*_{i0} , OCCPA1 = P^*_{i1}, OCCPA2 = P^*_{i2}, OCCPA3 = P^*_{i3}.
See text for definitions of curves and parameters.

For example, consider an item with $J = 4$ categories (0, 1, 2, 3). There are $M = 3$ category threshold or step difficulty parameters. Suppose that they have values $b_{i1} = -2.25$, $b_{i2} = 1.21$, and $b_{i3} = 3.47$, and that the item has a common slope parameter of $a_i = 1.5$. The four nonzero OCCs for this item are shown in Figure 3.

A graph of the category response curves (CRCs) of the GRM for this example item is shown in Figure 4. The two monotone CRCs are for the first and last categories, 0 and 3, while the nonextreme categories have non-monotone CRCs, CRC1P1 and CRC2P2. To have the highest likelihood of obtaining the highest category score for this item, a person needs a trait level θ_n greater than $b_{i3} = 3.47$. This can be seen at the intersection of the the CRC2P2 curve with the monotone increasing curve CRC3P3 at the right of the figure.

Figure 4: Category Response Curves for GRM



Note: CRC0P0 = P_{i0} , CRC1P1 = P_{i1}, CRC2P2 = P_{i2}, CRC3P3 = P_{i3}.
See text for definitions of curves and parameters.

2.1.3 Partial Credit Model

The Partial Credit Model (PCM) models the probability of responding in a particular response category differently than the GRM. In the PCM, the probability of endorsing specific category, $j' = j$ is modeled using all categories as in equation 2.5:

$$P_i(j' = j|\theta_n) = \frac{e^{\sum_{x=0}^{j}(\theta_n - \tau_{ix})}}{\sum_{s=0}^{M} e^{\sum_{x=0}^{s}(\theta_n - \tau_{ix})}}. \tag{2.5}$$

Note the summation expression in the denominator involves all of the categories and that $\sum_{x=0}^{0}(\theta_n - \tau_{ix}) = 0$.

The parameter τ_{ik} in item i is called the step difficulty which is related to category score k and indicates where the item threshold k is located on the trait continuum (Eid & Raubner, 2000; Embretson & Reise, 2000). Note that these step difficulty parameters are different from the step difficulty parameters of the GRM. In the PCM, each τ_{ik} parameter indicates the relative difficulty of each step and the point on the trait scale where the person has a probability of .5 of responding in the adjacent category, $k-1$. For the PCM, the step difficulty parameters of an item are the only item characteristics which help to explain the response behavior of persons since the slopes for all items are equal.

2.1.4 Nominal Response Model

A model constructed similarly to the PCM is the Nominal Response Model (NRM,Bock 1972). This model was designed to describe the probability that an examinee n with trait level of θ_n selects one of J categories for a nominally scored item on a multiple choice test. This model has also been used to test the assumption that items expected to yield ordered category responses have actually done so (González-Romá & Espejo, 2003; Thissen, Cai, & Bock, 2010). Thus,the model can be used to check if categories have any order and if the categories fall in the order expected (Ostini & Nering, 2006). If the items do not yield ordered responses, then the typically used integer scoring system is not tenable (González-Romá & Espejo, 2003).

Suppose that $j = 1, 2, ..., J$ are the score categories. Let a_{ij} represent the category slope parameters for item i and let c_{ij} represent the category intercepts. The model can be identified by $\sum_{j=1}^{J} a_{ij} = 0$ and $\sum_{j=1}^{J} c_{ij} = 0$ for each item i. For item i, the NRM gives the probability of endorsing specific category, $j' = j$, as in equation 2.6:

$$P_i(j' = j) = \frac{exp(a_{ij}\theta_n + c_{ij})}{\sum_{m=1}^{J} exp(a_{im}\theta_n + c_{im})}. \tag{2.6}$$

Item parameters in the original NRM can be difficult to interpret since a large slope parameter for a category in the NRM does not mean that an item will discriminate well as it does for the GRM (Wollack, Bolt, Cohen, & Lee, 2002). In the reparameterized NRM,

there is an overall single discrimination parameter that eases explanation of item analysis with the model (Thissen et al., 2010). This overall single discrimination parameter can be compared to those in other IRT models such as the GRM.

For the reparameterized NRM for item i, the probability of endorsing specific category, $j' = j$, is given in equation 2.7:

$$P_i(j' = j) = \frac{exp(a_i^* a_{ij}^s \theta_n + c_{ij})}{\sum_{m=1}^{J} exp(a_i^* a_{im}^s \theta_n + c_{im})}, \tag{2.7}$$

where a_i^* is the overall item slope parameter, a_{ij}^s is the scoring function (category slope) for response j, and c_{ij} is the intercept parameter in the original model. It is necessary to have identification restrictions such as $a_1^s = 0$, $a_{im}^s = m - 1$ and $c_1 = 0$ which are implemented by reparameterizing and estimating parameter vectors α (scoring function contrasts) and γ (intercept contrasts) in equation 2.8:

$$a^s = \mathbf{T}\alpha, \mathbf{c} = \mathbf{T}\gamma. \tag{2.8}$$

Note that when the NRM has been used in previous work, the contrast matrix $\mathbf{T}$ has included"deviation constrasts" from analysis of variance or a set of polynomial terms (Thissen et al., 2010). With the reparameterized NRM, the $\mathbf{T}$ matrix includes a column with linear terms and columns with Fourier function terms which provide a more numerically stable, symmetric orthogonal basis than using polynomial terms. Parameter estimation has been improved and the model has become more flexible in its use. Constrained versions of the NRM allow researchers to estimate the PCM or General PCM.

The product of the item slope a_i^* with a^s (the vector of category slopes) or $a_i^* \mathbf{T}\alpha$ gives the vector of original NRM category slope parameters in model described by Bock (1972). The $\mathbf{c} = \mathbf{T}\gamma$ gives the vector of original NRM intercept parameters.

2.1.5 Item Response Theory and Factor Analysis Models

While the discussion thus far has focused on IRT models, it is important to briefly summarize the equivalence of these models with Factor Analysis (FA) models since some of the literature

discusses the mixture IRT models in a CFA context. Additionally, some software programs (e.g., *Mplus*, L. K. Muthén & Muthén, 1998-2012) used to estimate the IRT models do so in a factor analytic framework. This equivalent parameterization is used to interpret the FA parameter estimates from software output.

Both IRT and FA models are used to describe an unobserved continuous variable of interest. This latent variable is referred to as a "trait" in the IRT setting and as a "factor" in the FA setting. The IRT model is a factor analysis model with categorical outcomes.

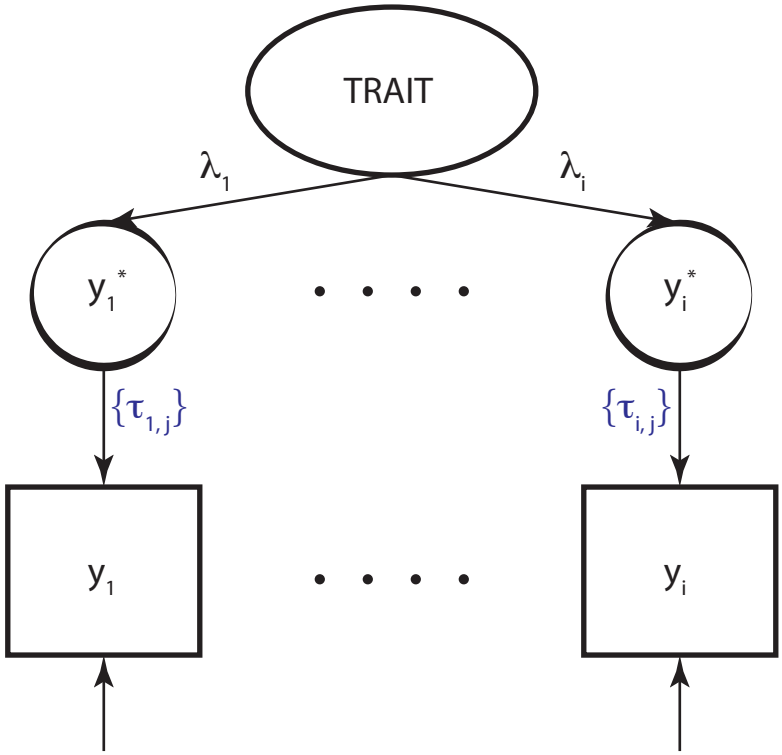
In typical factor analysis, both the observed responses and latent factor are continuous variables and there is a linear relationship between an observed response and the factor score for the person. In categorical confirmatory factor analysis (CCFA), a threshold structure is used to relate the discrete observed responses to continuous underlying latent response variables which are linearly related to the factor scores (Kim & Yoon, 2011; Wirth & Edwards, 2007).

Consider a one factor model to represent the relationship between the common factor scores (θ_n) and continuous latent response score variables, y^*_{ni} , that underlie the observed discrete scores, y_{ni} , for person n to item i. This model can be represented by

$$y^*_{ni} = \mu_i + \lambda_i \theta_n + \epsilon_{ni} \tag{2.9}$$

where ϵ_{ni} is the unique residual. In this model, the λ_i is the factor loading for the item. Typically the item intercept, μ_i , is set to 0 to impose a scale on the y^*_{ni} continuous response tendencies (McDonald, 1999).

Figure 5: One Factor Model with Latent Response Score Variables and Discrete Scores



Note: y_i^* = continuous latent response to item i , y_i = observed response to item i, λ_i = Factor loading for item i, $\{\tau_{ij}\}$ = Set of Thresholds for item i. Adapted from Kim and Yoon (2011).

Figure 5 shows how the factor influences the continuous response variables, y_i^*, that underlie the observed discrete scores, y_i. The latent response variables are shown from the discrete response options with a set of threshold parameters, τ_{ij} (Kim & Yoon, 2011). With respect to the continuous, latent response distribution for item i, the threshold parameters τ_{ij} define the J ordered-categorical responses (Wirth & Edwards, 2007) as seen in equation 2.10:

$$y_{ni} = j, \text{ if } \tau_{ij} < y_{ni}^* \leq \tau_{i(j+1)}, \tag{2.10}$$

where $j = 0, 1, ...$ M , and $\tau_{i0} = -\infty$ and $\tau_{iJ} = \infty$. There are $M = J-1$ finite thresholds for the J categories. The latent response variables, y_{ni}^* , are assumed to have a multivariate normal distribution. Correlations between these variables are estimated using the proportions of observed responses in the categories (Kim & Yoon, 2011; Wirth & Edwards, 2007).

The output from the CCFA includes estimates for the factor loadings (λ_i) and thresholds (τ_{ij}). The output for the IRT analysis includes estimates for the difficulty and slope parameters previously discussed. Assuming standardized FA and IRT models (a zero mean and unit variance for the latent factor) and a variance of one for the ϵ_{ni}, the two models are equivalent (Kamata & Bauer, 2008; Sawatzky et al., 2012). Furthermore, the parameters for the GRM in IRT can be determined using equation 2.11:

$$a_i = \lambda_i \tag{2.11}$$

and equation 2.12:

$$b_{ij} = \tau_{ij}/\lambda_i. \tag{2.12}$$

If these equivalent FA parameters for the slope and difficulty parameters are substituted into equation 2.3, then the equivalent factor analysis parameterization for the probability in equation 2.3 is given by equation 2.13:

$$P_{ij}^*(j' \geq j|\theta_n) = \frac{e^{\lambda_i\theta_n - \tau_{ij}}}{1 + e^{\lambda_i\theta_n - \tau_{ij}}} \tag{2.13}$$

This equivalent parameterization is used to determine the parameters in the IRT models estimated in this study. The models are extensions of the PCM, GRM, and 2PLM and describe the multidimensionality in the data. A multidimensional model is used to explain performance or item responses arising from more than one primary dimension and this type of model is discussed next.

2.1.6 Multidimensional Models

Researchers and practitioners use a multidimensional model to estimate a set of latent trait scores for each person since more than one trait affects item responses. Reckase (2009) describes how a multidimensional IRT (MIRT) model is either compensatory or noncompensatory (partial compensatory). For a compensatory model, the components of the trait vector are combined additively with item parameters in a linear combination. With such a model, a high value on one trait compensates for a low value on a different trait so that the same sum could result for different combinations of trait levels. The probability of a particular response is then calculated from this linear combination using an IRT model(Reckase, 2009). An example of a compensatory MIRT model can be seen in the multidimensional extension of the one dimensional 2PL model given in equation 2.2 to a model with t traits. In such a model, the exponent is a linear combination of the components of the t-dimensional trait vector ($\theta_{\mathbf{n}}$) for each person and there is an associated t-dimensional vector of slope parameters for an item ($\mathbf{a_i}$) and an intercept term for the item (d_i). The probability of response $j = 1$ for the values of the traits and item parameters is given by equation

$$P_i(j = 1|\theta_n, \mathbf{a_i}) = \frac{e^{\mathbf{a_i}\theta_{\mathbf{n}}{}' + d_i}}{1 + e^{\mathbf{a_i}\theta_{\mathbf{n}}{}' + d_i}}. \tag{2.14}$$

For a noncompensatory model, a unidimensional model is used for each separate trait needed to complete a questionnaire item. The product of the probabilities from the separate models gives the probability of a particular response. This model is different from the compensatory type since the probability of a particular response will not be greater than the highest probability for a given trait. The compensation of a high trait value for a low trait value is reduced. Reckase uses the term "partial compensatory" for these models since a high value on one trait means that the probability associated with this trait will be higher than it would if there was a low value on the trait. Some compensation does occur. The multi-process model (M-PM, discussed below) is a partial compensatory model.

In this study, five different multidimensional models are compared: the M-PM, the Multidimensional Partial Credit Model (MPCM), the Multi-dimensional Nominal Response Model (MNRM), and the mixture graded response (mixGRM) and partial credit (mixPCM) models.

These models are useful for determining personality trait estimates that have been adjusted for response style bias. Accounting for response style bias is important since this bias can affect statistics based on survey research.

2.1.7 Multidimensional Partial Credit Model

The Multidimensional Partial Credit Model (MPCM) is a compensatory model and was discussed in Kelderman (1996). Suppose there are t traits to be estimated by the model and for person n, the trait parameters are represented by θ_{nq} (q = 1,...,t). An indicator variable, ω_{qij}, is used to designate how items are assigned to the dimensions. This variable is 1 when an item i response measures a dimension q and 0 otherwise. The j indicates the item category for categories $j = 0, 1, ..., M$. As in the PCM, τ_{ix} represents the threshold parameter between two categories $x-1$ and x. The MPCM models the probability of endorsing specific category, $j' = j$, as in equation 2.15:

$$P_i(j' = j) = \frac{exp[\sum_{x=1}^{j}\sum_{q=1}^{t}(\omega_{qij}\theta_{nq} - \tau_{ix})]}{1 + \sum_{s=1}^{M} exp[\sum_{x=1}^{s}(\sum_{q=1}^{t}(\omega_{qij}\theta_{nq} - \tau_{ix})]}. \tag{2.15}$$

2.1.8 Multidimensional Nominal Response Model

The multidimensional nominal response model (MNRM, Bolt & Johnson, 2009; Bolt & Newton, 2011; Falk & Cai, 2015; Takane & De Leeuw, 1987) is also a compensatory model, like the MPCM, and assumes that the measured traits combine together additively to produce the item response. Suppose there are $j = 1, 2, ..., J$ score categories and Q traits to be estimated by the model and for person n, the trait vector is represented by $\theta_{\mathbf{n}}$. Each component θ_{nq} is a score on the trait q, $(q = 1, ..., Q)$. A $Q \times 1$ slope vector $\mathbf{a_j}$ represents the loadings of category j on the Q latent variables. Each c_j is the intercept for category j.

The $\mathbf{a_j}$ and $\mathbf{c}$, respectively, contain the slopes and intercepts. The model can be identified by $\sum_{j=1}^{J} a_{ijq} = 0$ and $\sum_{j=1}^{J} c_{ij} = 0$ for each item i and trait q. The MNRM models the probability of endorsing specific category, $j' = j$, as in equation 2.16:

$$P_i(j' = j) = \frac{exp(\mathbf{a_j'}\theta_\mathbf{n} + c_j)}{\sum_{m=1}^{J} exp(\mathbf{a_m'}\theta_\mathbf{n} + c_m)}. \tag{2.16}$$

Falk and Cai (2015) illustrated a reparameterization of the MNRM due to Thissen and Cai (2016). As in the preceding discussion, suppose that there are Q traits to be estimated by the model and for person n, the trait vector is represented by $\theta_\mathbf{n}$. The Q slope parameters are given in vector $\mathbf{a}$ and J intercept parameters are given in vector $\mathbf{c}$. $\mathbf{S}$ is a $Q \times J$ matrix with scoring function values for the item categories and modeled dimensions. Each column $\mathbf{s_j}$ is for category j and each row corresponds to a trait dimension.

With this reparameterization, the overall item slopes $\mathbf{a}$ are separated from the scoring functions (order of the categories) $\mathbf{S}$. The order of the categories can be fixed to hypothesized values for an interesting dimension, such as one measuring Midpoint (or Extreme) response style. Let the $\circ$ represent the entrywise (Schur) product. The MNRM models the probability of endorsing specific category, $j' = j$, as in equation 2.17:

$$P_i(j' = j) = \frac{exp([\mathbf{a} \circ \mathbf{s_j}]'\theta_\mathbf{n} + c_j)}{\sum_{m=1}^{J} exp([\mathbf{a} \circ \mathbf{s_m}]'\theta_\mathbf{n} + c_m)}. \tag{2.17}$$

The reparameterization enables the estimation of overall item slopes for response style dimensions that can differ for each item. User-defined response style scoring functions that differ across items can also be estimated. The scoring functions can differ across latent dimensions so that two known response styles (e.g., ERS and MRS) could be modeled simultaneously.

2.2 SURVEY RESEARCH AND RESPONSE BIAS

Surveys are one of the most commonly used social science research methods to test behavior theories (Groves et al., 2004), to measure attitudes (Fitzpatrick, Sanders, & Worthen, 2004), and to assist program evaluators. Surveys are conducted in healthcare, education, marketing, and many other fields. The paper or online questionnaire is an efficient way to collect survey data and the questionnaire often includes self-report scales of items with a Likert response format. This format is particularly popular in psychological assessments of attitudes or personality traits (Khorramdel & von Davier, 2014; McCoach, Gable, & Madura, 2013).

To use Likert's method, experts design questionnaire items with the goal of placing respondents on a continuum that represents the latent trait, θ, to be measured. The experts also choose a response scale format (i.e., number of response options, use of a midpoint or not, and anchor labels for the options). The extent of agreement is expressed by selecting one of the ordered categories which are typically labeled.

Unfortunately, using the common Likert method can potentially introduce undesired method variance or measurement error (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). To complete survey items, respondents go through a complex five-step process and in the last step mark a category on the response scale (Tourangeau, Rips, & Rasinski, 2000). The steps of this process, activities accompanying each step, and potential method biases are seen in Table 3. Respondents interpret the item, retrieve general and specific memories related to the item, use the information to make a judgment, select a response option, and report a response. The last step can involve editing the response since some respondents may satisfice, provide a socially desired response or one that is consistent with previous responses (Krosnick, 1991, 1999; Podsakoff et al., 2003; Tourangeau et al., 2000). Due to possible method biases, the response options might not capture what the researcher intends. Thus, using Likert response formats can affect the response process and create the potential for response bias.

Table 3: How Common Method Biases Can Affect the Response Process

Stage	Activities involved with each stage	Potential Method Biases
Comprehension	Attend to items and directions, represent logical form of item, identify information sought, and link key terms to relevant ideas	Item ambiguity
Retrieval	Generate retrieval strategy and cues, retrieve general and specific memories, and fill in missing details	Measurement context, item context, item embeddedness, item intermixing, priming effects, item social desirability, scale size, transient mood states
Judgment	Assess completeness and accuracy of memories, draw inferences based on accessibility, inferences that fill in gaps of what is recalled, integrate retrieved material, and make estimate for partial retrieval	Consistency motif (if it is an attempt to increase accuracy in face of uncertainty), implicit theories, priming effects, item demand characteristics, and item context-induced mood states
Select Response	Map judgment onto response category	Item characteristic effects, common scale anchors and formats and item-induced anchoring effects
Report Response	Possible editing of response for consistency, acceptability or other criteria	Consistency motif (if it is an attempt to appear rational), leniency bias, acquiescence bias, social desirability, and demand characteristics

Note: Adapted from Podsakoff et al. (2003).

There are two broad types of response bias: response set and response style (D. Paulhus, 1991; Peer & Gamliel, 2011). A **response set** occurs when the person unconsciously or consciously gives item responses to present a certain self-image. This refers to a specific situation or temporary reaction to items (Cronbach, 1950; Van Herk et al., 2004). The responses given depend upon the item content, such as during an application for employment. In such situations, employers using the items would be concerned about Social Desirability Responding. This type of response bias is often controlled for by using a separate scale of items such as the Marlowe-Crowne Social Desirability Scale (Fischer & Fick, 1993).

When a person tends to respond to a set of items independent of the item content across situations and time, a **response style** occurs (Jackson & Messick, 1958; Van Herk et al., 2004). A person's response style is a trait indicated by overuse of certain response options and is independent of the latent trait being measured. The presence of substantial response style traits contaminates measurement of the desired latent trait. Controlling for this type of response bias has initiated a multitude of studies due to the different types of response styles that exist and to the different methods that are available to account for response style use.

2.3 RESPONSE STYLES AND MODELS TO ACCOUNT FOR RESPONSE STYLES

Three of the most well-known response biases in the psychological literature are acquiescence, extreme response, and socially desirability responding (SDR)(D. Paulhus, 1991; D. L. Paulhus & Vazire, 2007; Van Herk et al., 2004). Although SDR is especially important in self-reports of sensitive behavior and personality scales (D. Paulhus, 1991; Van Herk et al., 2004), this study takes the position that SDR is a response set, not a response style, as Schimmack, Böckenholt, and Reisenzein (2002) and Van Herk et al. (2004) have done. Thus, methods to account for SDR are not discussed here; however, discussion and review of some methods can be found in Nederhof (1985) and Helmes, Holden, Carstensen, and Ziegler (2014).

Table 1 presented several different types of response styles and their definitions. While the most common response styles are Acquiescence (ARS), Disacquiescence (DRS), Extreme (ERS), and Midpoint response style (MRS), the two most commonly researched response styles are ARS and ERS (A. W. Harzing, Brown, Köster, & Zhao, 2012; Weijters et al., 2010a). Due to the number and variety of response styles, researchers and practitioners have developed, studied, and used many different approaches depending on what goal they want to address. First, researchers decide whether they will address any response styles at all. Plieninger (2016) outlines the three potential ways that researchers and practitioners can view response styles: (1) They ignore response styles since they may not understand enough or cannot implement a control for them; (2) They see variance due to response styles as small error variance or as negligible compared to content and thus, do not address them (e. g. Schimmack, Böckenholt, & Reisenzein, 2002); or (3) They see response styles as serious threats to data quality which, in turn, affect inferences from scores or trait estimates.

Researchers and practitioners with the third viewpoint must make decisions about how many and which response styles will be addressed (i.e., ERS, ARS, MRS, etc.), how they will be measured (using substantive trait items or a large set of uncorrelated items), and how they will be examined (e.g., traditional methods, CFA method, IRT method). It is beyond the scope of this chapter to discuss fully the more than 60 years of research concerning response styles; however, some key methods available are briefly summarized to illustrate some of the different methods that have been used and types of models available. The methods are classified by type of items (heterogeneous or homogeneous) in the questionnaire.

2.3.1 Methods where Heterogeneous Content is Available

For surveys which contain many different uncorrelated items in addition to the items related to the substantive trait, a simple method involves using the count procedure to measure how many extreme, midpoint or acquiescent responses have been used. These response style measures can be used in ANCOVA models to obtain scores which have been "purged" of response style variance due to the added covariates. For example, Reynolds and Smith (2010)

used ANCOVA to control for ERS, MRS, dispersion, and ARS. The researchers found there were less cultural differences on the substantive constructs related to Service Quality (e.g. assurance, responsiveness) when taking response styles into account.

Slightly more complicated methods using Representative Indicators for Response Styles (RIRS) involve single or multilevel regression models along with the count procedure. Greenleaf (1992b) used a logistic transformation of the ERS scores as the dependent variable in a regression model. He found that gender did not explain a significant difference in ERS; however, increasing age, decreasing education, and decreasing household income were significantly associated with increasing use of ERS. Greenleaf (1992a) used regression to remove the response bias due to standard deviation (response range), a response style related to ERS. As age increased, standard deviation increased. Greenleaf found that with the adjustment for this response style, the mean age increased and mean education levels decreased in certain marketing segments.

Baumgartner and Steenkamp (2001), using RIRS for five different response styles in a multilevel regression model, found that noncontingent (random, careless) responding did not bias scale scores systematically, but ERS and MRS did affect variation in scale scores, particularly when the scale mean (on the response scale) differed the most from the scale midpoint. The researchers also found that using scales with positively and negatively worded items helped to effectively account for 60-62% of the variance due to ARS and DRS. This supports the common suggestion of using scales with reversed items to control for ARS and DRS .

Using a more complex, factor analytic approach with the RIRS, Weijters et al. (2010a) found that persons with higher levels of ARS had higher score levels and persons with higher levels of ERS tended to use extreme options much more than persons with lower levels of ARS and ERS. Both ARS and ERS led to bias in the same questionnaire and were consistent across several different scales. For questionnaires with reversed items and carefully balanced content of the positively and negatively worded items, the response style bias from ARS becomes small and can be controlled with a measurement model for these items. Billiet and McClendon (2000) illustrate this model.

Using a means and covariance structure model, Weijters et al. (2008) illustrated the use of extra heterogeneous content items to account for four different response styles (ARS, DRS, ERS, MRS). The RIRS measured the latent response style factors and the variance due to these response style factors was removed from the indicators for the substantive trait. Differences in the latent trait from respondents in the three survey administration modes (online, telephone, and pencil paper) that appeared due to the presence of response styles were removed. The response style measurement model had improved the latent trait estimates for the three groups by correcting for the bias that occurred in the factor model which did not address response styles.

From these studies, Weijters and colleagues recommend using a dedicated set of heterogeneous items to account for response styles when possible. This can be difficult when additional items are not available since many surveys measure several constructs. In these cases, another method is needed and can be found among the various methods for using homogeneous content items.

2.3.2 Methods where only Homogeneous Content is Available

Some surveys have one or more substantive scales and do not contain a large number of heterogeneous items. In such cases, there are many different types of methods which typically address one or two response styles and one or two substantive traits. The following sections illustrate applications of some of the different models available. This review focuses on studies using models with attributes similar to the models in this study; i. e., the studies reviewed involve methods using a latent classes or multidimensional modeling to address ERS or ERS and MRS. Studies addressing ARS are only briefly described since the models in the current study are not being used to address ARS or DRS. More thorough discussions of handling ARS can be found in (Billiet & McClendon, 2000; Cheung & Rensvold, 2000; Ferrando, Morales-Vives, & Lorenzo-Seva, 2016; Morren, Gelissen, & Vermunt, 2011; Savalei & Falk, 2014).

2.3.3 Methods related to Latent Class Analyses

Some approaches to address ERS and other response styles use a latent class model to show existence of distinct classes which differ in response style use. For example, Van Rosmalen, Van Herk, and Groenen (2010) used latent class regression to investigate what different classes with respect to item content and response style factors would emerge. The Bayesian Information Criterion (BIC) was used to determine the number of dimensions and classes. The researchers imposed a bilinear parameter structure on a multinomial logit model and illustrated with graphs how respondent characteristics affect response behavior and content ratings. The researchers used their mixture model to account for response heterogeneity and dependencies between observations in a study of the nine item "List of Values" scale given in five European countries. Each value was rated with a nine-point scale ranging from 1 ="very important" to 9 = "not important at all" The model chose two different dimensions. One set of latent classes was formed based upon item content ratings and another based upon response tendencies. The response tendencies dimension (set) revealed 11 different classes: strong acquiescence, moderate acquiescence, weak acquiescence, nuanced positive, moderate, wide response range, extreme scoring, midpoint responding, weak disacquiescence, strong disacquiescence, and incomplete response. One notable finding was that the midpoint responding class also had a high degree of extreme responses.

The item content dimension (set) was composed of five classes. Three of the classes met the content expectations. The first class consisted of "hedonist" persons for whom "fun and enjoyment" and "self-respect" items were very important. Many persons in this group tended to be relatively young, highly educated, and Spanish or French. The second class consisted of "group-oriented" persons for whom "belonging" was the most important value. Members tended to be Germans over the age of 40 and with lower education levels. The third class was "self-oriented" for whom "self-respect" was most important; its members tended to be relatively older Italians. The other two classes were described as "indifferent" and "mixed opinions". The "indifference" class had many young Spaniards using moderate

(nonextreme), midpoint, and extreme responses; while the "mixed opinions" class members were young British persons who showed strong acquiescence or disacquiescence. The latter two classes were less outspoken than the first three described.

Morren et al. (2011) used latent class factor analysis (LCFA) to investigate two substantive factors and an ERS factor in a study of four ethnic groups. On the "Attitudes toward Dutch society" factor, the four groups were basically similar. For the other factor "Autonomy of the children", Moroccans and Turks had higher scores on average than the Antilleans and Surinamese did. Including a factor to account for ERS improved model fit. Three latent classes were identified, which for substantive factors, designated negative, neutral, and positive attitudes. For the style factor, the three classes were labeled Low ERS, Middle ERS, and High ERS. Persons with a tendency for using the Agree/Disagree categories (and low use of extreme options) were classified in the Low ERS class. The Middle ERS class had a high tendency to use midpoints and the High ERS class had high use of extreme options. Moroccans tended to use extreme options slightly more than Turks while Surinamese did slightly less.

The group differences could only be partially explained by including ERS in the model. There were large individual differences in response style use in the groups. The presence of ERS suppressed the ethnic group differences and the bias due to ERS was removed using an LCFA model with item ratings specified to be ordinal with respect to substantive factors and nominal with respect to the ERS factor. Then the differences among ethnic groups were validated with the Antilleans and Surinamese scoring much higher than the Turks and Moroccans on the substantive factors.

2.3.4 Using Multidimensional Item Response Theory Models to account for Response Styles

Much of the work with IRT models has been in developing models to account for ERS since this is one of the most common response styles. To address ERS, researchers use existing IRT models to build models that involve latent individual difference variables for the measured target construct and for the tendency toward extreme response (Thissen-

Roe & Thissen, 2013). The researchers make decisions regarding the parametric model to measure the substantive construct(s), the number of latent substantive constructs, and the parametric model form to describe the individual ERS effect on the response (Thissen-Roe & Thissen, 2013). Researchers then decide if parameters for items and response categories are constrained to be equal or not. For multidimensional construct models, determination of constraints to identify the common factor model must be made also.

In the multilevel model developed by De Jong et al. (2008), the items which measure ERS do so differentially which is important since items do not invoke ERS in persons in the same way. In this model, a dichotomous IRT model is used to estimate ERS_{ij} the latent ERS trait level for person i in county j. For an item k from scale r, the model includes parameters for item discrimination (a_{kj}), item difficulty parameters (b_{kj}) and a testlet parameter for a scale effect (Ψ_{ijr_k}; c.f. Wainer, Bradlow, & Wang, 2007). A testlet is a set of items that relate to one content area and is developed as a unit. The testlet effect parameter was needed as there were 14 different scales and items within the same scale are correlated. For the standard normal cumulative distribution function Φ, the model is shown in equation 2.18:

$$P(EXT_{ijk} = 1|ERS_{ij}, \Psi_{ijr_k}, a_{kj}, b_{kj}) = \Phi[a_{kj}(ERS_{ij} - \Psi_{ijr_k} - b_{kj})] \tag{2.18}$$

In this model the substantive traits and ERS trait were not assumed to be correlated. The summed scores were adjusted by regression of the ERS scores on the sum scores to create 'purified' scores which could be used in for other analyses.

The researchers investigated the framework of cultural dimensions by Hofstede (2001). One dimension is cultural individualism which measures whether individual attitudes are largely regulated by individual preferences (individualistic society) or society's preferences (collectivist society). Cultural femininity/masculinity refers to the degree a society is characterized by modesty, gentleness, and nurturance versus assertiveness, achievement, and ambition (Hofstede, 2001). A third cultural attribute "uncertainty avoidance" refers to the extent a society feels nervous or threatened by undefined, risky, or ambiguous situations and chooses rigid rules and attitudes to avoid such situations. The fourth cultural dimension is "power distance". Countries with high levels of this variable stress conformity and societies

tend to be more authoritarian than not (Hofstede, 2001; Johnson, Kulesa, Cho, & Shavitt, 2005) and persons tend to be acquiescent. Cultures with low power distance emphasize equality in status and modesty.

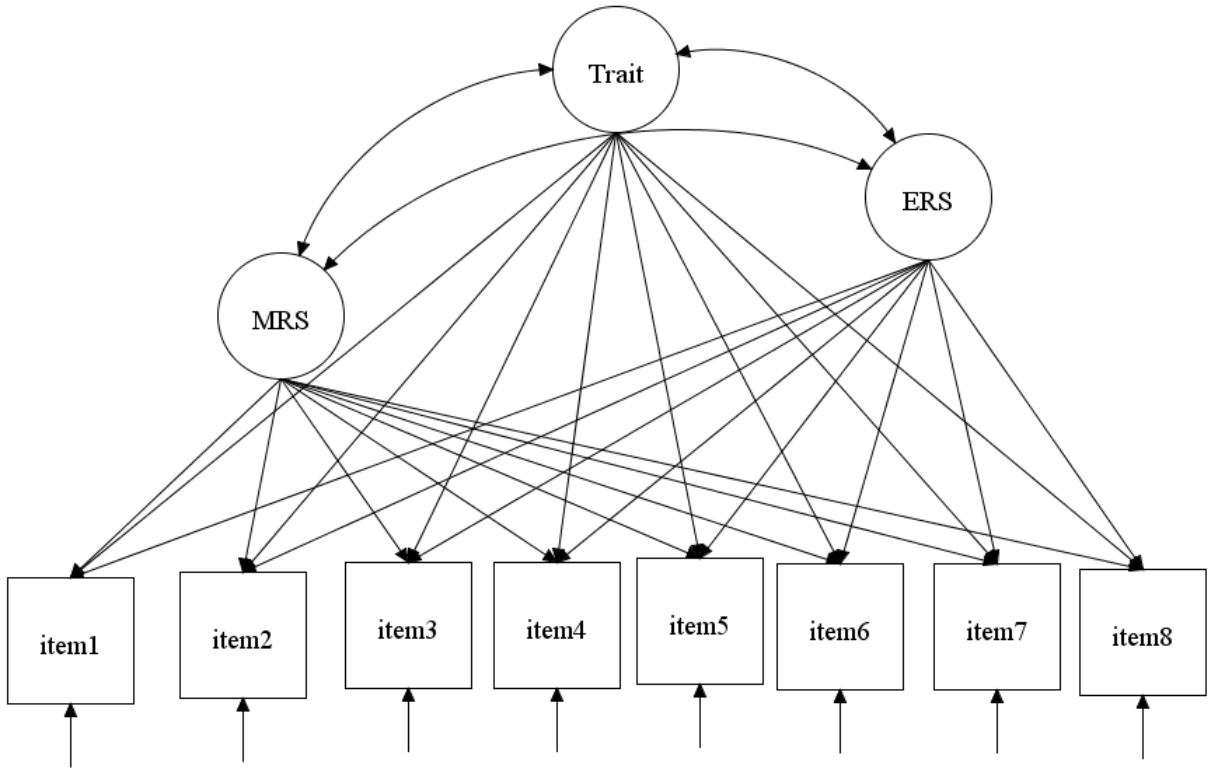
The researchers found no relationship between ERS and power distance; however, ERS was positively related to national cultural individualism, uncertainty avoidance, and cultural masculinity. De Jong et al. (2008) also found that women tended to score higher on ERS than men. The positive effect for both the country level masculinity and sociodemographic variable gender illustrated that femininity and masculinity are not merely simple personality traits and contemporary gender role patterns should not be equated with historical gender role patterns. Thus, there can be differences in relationships between variables at the cultural and individual levels since the underlying mechanisms differ (De Jong et al., 2008; Hofstede, 2001). The positive effect of cultural individualism on use of extreme options is illustrated by countries such as the U.S., Italy, and the Netherlands having high use of ERS and countries such as Taiwan, China, and Thailand which have much lower use.

A different approach is the multidimensional nominal response model for ERS (MNRM, Bolt & Johnson, 2009; Bolt & Newton, 2011). Bolt and Johnson (2009) examined the substantive trait "tobacco dependence level" using a 68-item scale and an ERS trait to account for differential use of the response scale. In the study, the latent trait and ERS trait were assumed to be uncorrelated. A two dimensional model with equal category slopes for the substantive trait and equal slopes for the ERS trait fit better than one-, two-, or three-dimensional NRMs with no slope constraints using BIC and CAIC fit measures. The model accounted for the effects of the substantive trait and the ERS trait on the item responses. Bolt and Newton (2011) used the MNRM to address for ERS in short attitude scales with responses to the five-item Science Enjoyment scale and the ten-item Science Value scales from the 2006 Programme for International Student Assessment (PISA). The model used assumed that the substantive traits and ERS traits were correlated. The MNRM allowed accurate estimation of and accounting for ERS, even though the scales were correlated on the substantive traits. Stronger links between the science enjoyment, value and achievement variables were found when the effects of ERS were removed.

Falk and Cai (2015) extended research with the MNRM using a reparameterization which allows response styles to be defined differently across items and enables items slopes to vary for response style and substantive trait dimensions across items. An advantage of the reparameterization is its use to address both ERS and MRS in addition to the substantive traits. The model assumes that the latent traits and response style traits are correlated. Using empirical data with six correlated "cigarette smoking" constructs, Falk and Cai illustrated one use of the model. A model accounting for both ERS and MRS traits with freely estimated ERS slopes and fixed MRS slopes across items fit better than a model allowing MRS slopes to vary across items. The best fitting model also fit better than a model with the latent traits and ERS. Including the response style factors results in adjusted scores for the substantive factor means. High ERS scores resulted in adjustments toward the substantive trait means while low ERS scores were adjusted away from the means. Other response styles such as ARS or SDR can be addressed using different scoring functions for indicators for these response styles in the model. The model is quite flexible since it can address other response styles and it includes the models described by Bolt and Johnson (2009) and Johnson and Bolt (2010).

In another study, Wetzel and Carstensen (2015) used different two- and three-dimensional partial credit models to examine the relationships between personality traits and one or two response styles (ERS, MRS, ARS, or DRS). These models assumed that the personality trait and response style traits were correlated. A picture of a three dimensional model is given in Figure 6. The researchers determined best fitting models by using minimum Akaike, Bayesian, and Consistent Akaike information criteria. Wetzel and Carstensen (2015) examined models which used the same items (i.e., homogeneous scale items) since these models provided a scale-specific response style trait estimate and a desired trait estimate that has been corrected for the "scale-specific response style" effects. The analyses revealed that ERS was the most important response style (over MRS, ARS, and DRS) for explaining variance in item responses. Three-dimensional models involving ERS, MRS, and the trait of interest fit best for 26 of 30 personality facets. Three-dimensional models that did not incorporate an ERS trait never fit as well as models that did. Adding MRS to a model with trait and ERS dimensions explained more variance than adding ARS or DRS.

Figure 6: Three Dimensional Partial Credit Model



Note: Three traits affect responses to a five-point item in a compensatory (additive) way. ERS = Extreme Response Style, MRS = Midpoint Response Style. Scoring for trait dimension(0,1,2,3,4); Scoring for ERS dimension (1,0,0,0,1); Scoring for MRS dimension (0,0,1,0,0). Adapted from Wetzel and Carstensen (2015).

In models which used additional items items to measure response style effects, the additional items captured more general response style trait estimates. In such models, the desired trait estimate is different due to the correction for a "general response style tendency". The models using the same items to measure response style traits as the scale trait adjusted trait estimates for response style use to a higher degree than other models. Models using additional heterogeneous items to measure response style traits did not provide adjusted trait estimates unless the items loaded on the scale content items also. This changed the adjustment of the desired trait estimate for response style use. The adjustment was less than that from models using the same items since "general response style tendencies" were measured.

2.4 THE MULTI-PROCESS MODEL

The models reviewed above do not attempt to explain how respondents make decisions in choosing response options. In contrast, however, Thissen-Roe and Thissen (2013) proposed a Two-Decision Model (TDM) for ERS. The TDM combines the first and second steps of the Multi-process Model and like the M-PM is a noncompensatory model. With this model, items elicit ERS differentially and the substantive trait and ERS trait are correlated. The TDM assumes that to give a response, the respondent asks two questions "Do I agree with the statement?" and "How strongly do I want to express my position?" The midpoint could be chosen to express indifference or to de-emphasize agreement or disagreement with the item content.

Like the TDM, the multi-process model is discussed as a way to study response styles as continuous variables. Imagine Figure 5 with the latent trait measured by responses to eight items. This figure indicates that only a single trait or one process determines the item responses.

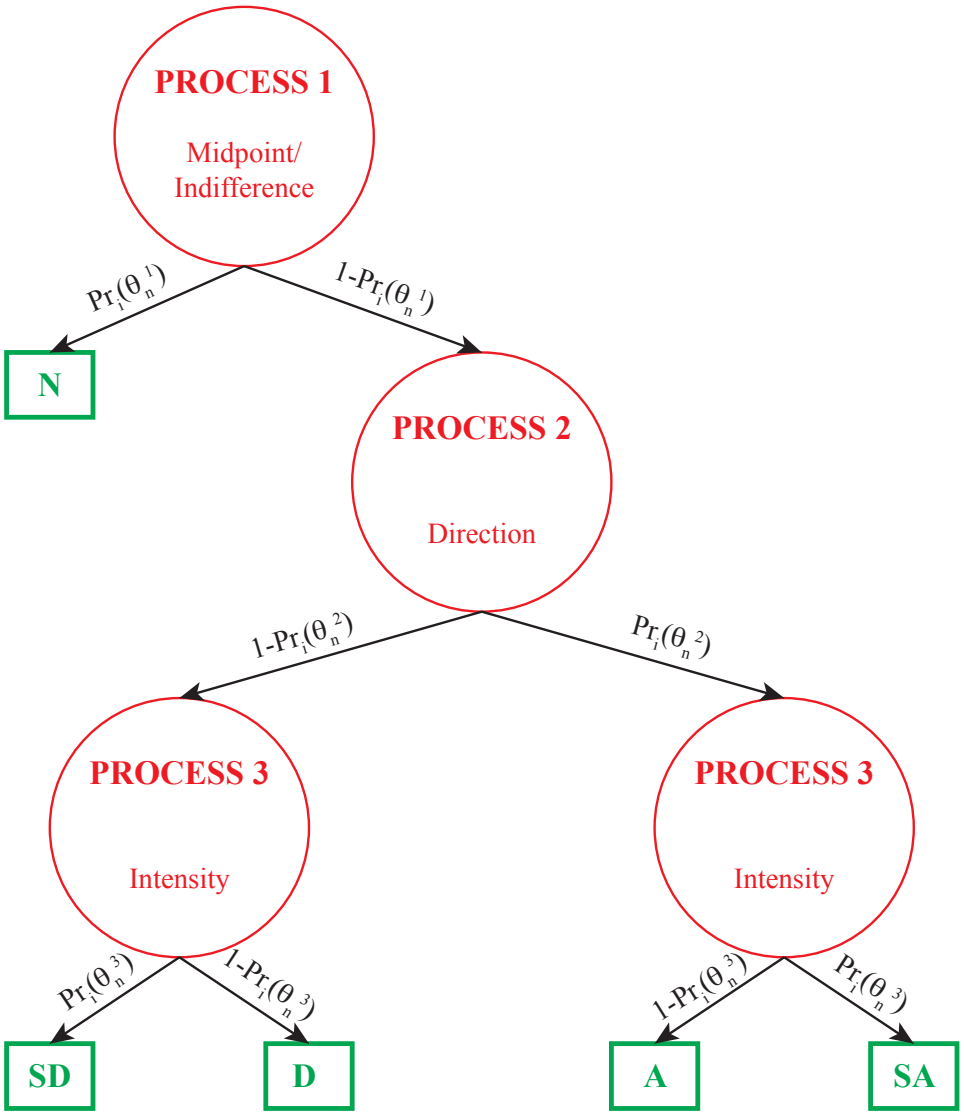
2.4.1 Presentation of the Multi-Process Model

The multi-process model (M-PM) acknowledges that two or more traits may be needed to account for the observed response. With the M-PM, a profile of scores, not a single trait score, is obtained for each person. The M-PM breaks down the response process into a series of steps or stages. In this study, a three-process model is used to analyze five-point response format data. The tree-like structure in Figure 7 shows the three successive processes of indifference (1, use of *m*idpoint or not), *d*irection (2, agree or disagree), and intensity (3, *e*xtreme or not). If a person does not have a distinct opinion about a given item's content, the person selects the middle category and the response process ends. The indifference process is the only one used for the given item. On the other hand, if the person has a well-defined opinion about the item content, the person uses the direction process to express agreement or disagreement with the item. Lastly, the intensity process is used to express how strongly the opinion is held; the person chooses an extreme option or not.

The data needed for the M-PM come from creating binary pseudo-items (BPIs) to model the outcomes of the processes by recoding the original five-point response data into binary format. Khorramdel and von Davier (2014) label the three BPIs as *m*-items, *d*-items, and *e*-items to reflect that choices have been made regarding use of *m*idpoint (*m*-items), *d*irection from the midpoint (*e*-items), and *e*xtreme option (*e*-items). Each *m*-item (BPI 1) is coded 1 if the *m*idpoint is chosen for the original item and 0 otherwise. Each *d*-item (BPI 2) is coded 1 for the agreement *d*irection, 0 for the disagreement direction, and missing if the midpoint was chosen. Each *e*-item (BPI 3) is coded 1 for an *e*xtreme response, 0 for a non-extreme response, and missing if the midpoint was chosen.

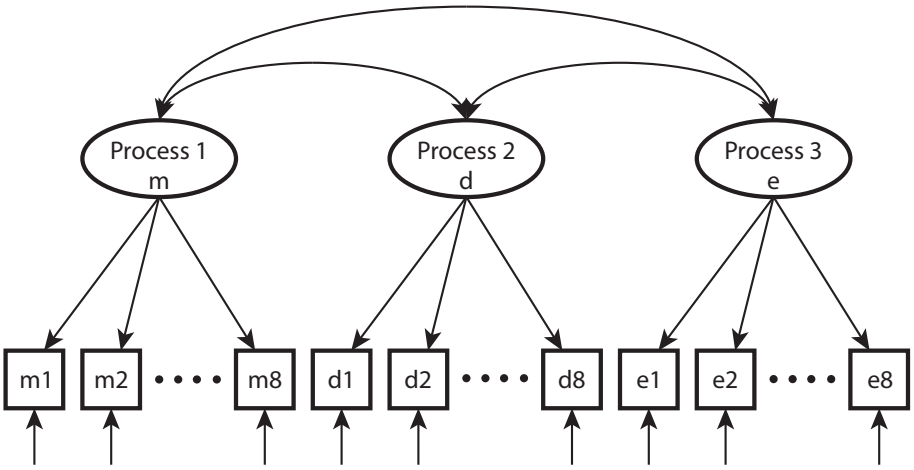
Thus, for each item and process, a BPI is created, so that all of the information from the observed data is maintained, yet expressed with three dichotomous items. These BPIs indicate how the person used the three distinct response processes (1, 2, 3) for each item. To estimate the M-PM, the recoded data are used in a correlated three-factor model with the eight BPIs for each process as shown in Figure 8. Analysis of the model yields item and person parameters for each of the processes. The model also estimates correlations between the substantive (direction) trait and the ERS (intensity) and MRS (indifference) traits.

Figure 7: Tree structure of Three Successive Processes



Note: Unobserved processes used to respond to a five-point item. SD = Strongly Disagree, D = Disagree, N = Neither Disagree or Agree, A = Agree, SA = Strongly Agree, $Pr_i(\theta_n^h)$ = Probability person n with trait level θ_n^h uses process h to response to item i. Adapted from Böckenholt (2012).

Figure 8: Multi-Process Model



Note: A three dimensional multi-process model for response styles where the three response processes combine in a noncompensatory way to produce the item response. Process 1 m = indifference (midpoint) process, m1 = Binary Pseudoitem 1 for process 1 m; Process 2 d = direction (trait) process, d1 = Binary Pseudoitem 1 for process 2 d; Process 3 e = Intensity (extremeness) Process, e1 = Binary Pseudoitem 1 for process 3 e.

2.4.2 Studies Using a Multi-Process Model to Account for Response Styles

In empirical studies with very large samples (greater than 63,500), Thissen-Roe and Thissen (2013) found the TDM outperformed the MNRM and proportional Thresholds model (PTM) when estimating the substantive latent trait and the ERS trait. All three models (MNRM, PTM, and TDM) outperformed models which did not have an ERS dimension (e. g., GRM, NRM) and indicates the importance of accounting for ERS. From three disjoint representative samples in one of the studies, the mean correlation between the extreme response and main dimensions was larger for the TDM (.331) than the mean correlations between the same dimensions for the PTM (.091) and MNRM (-.052). This indicated a difference in the meaning of the ERS traits estimated by the three models. The MNRM did not include a way to address midpoint responding since only trait and extreme response dimensions were included. On the other hand, the PTM views midpoint responding as a negative counterpart of extreme responding and has no way to treat midpoint responses. The Two-Decision Model

permits midpoint responses by those with High ERS and consistent MRS. Unfortunately, the TDM does not provide an estimate for the midpoint response style as the multi-process model does.

Leventhal (2017) compared three different MIRT models accounting for ERS: an IRTree model similar to the M-PM, a Multidimensional Nominal Response Model (MNRM), and a Modified Generalized Partial Credit Model (MGPCM). These models assumed that the substantive and ERS traits were independent. The empirical study examined a unidimensional scale with items measured with a four-point response format. The correlations between substantive trait estimates were quite strong (all greater than 0.974).

The correlations between ERS trait estimates revealed differences among the models even though the correlations were all strong. There were weaker correlations when the IRTree model was paired with the MGPCM and MNRM compared to when the MNRM and MGPCM are paired. Leventhal (2017) suggested that the difference may be due to the trait definition. The IRTree model models an ERS response process. This differs from ERS treated as an extreme response tendency as in the other models. The MNRM and MGPCM were favored over the IRTree model in terms of fit.

Some researchers have used the M-PM to examine 5-point Likert data. Böckenholt (2012) examined two items measuring whether ethical (or unethical) behavior of a firm is important in consumer purchasing decisions. Model fit was examined by using the log-likelihood ratio statistics which compared the predicted and observed category frequencies. He found that a one process GRM fit the data poorly. A two-process GRM was then tested. In this model, the first process modeled if the midpoint was chosen by using a BPI. The second process used an *ordinal four-point* pseudo-item to model both direction and intensity of the original response. Absolute fit was improved, but still poor. Böckenholt then used a three-process GRM with m, d, and e BPIs to model outcomes of the indifference, direction, and intensity processes. This three-process M-PM represented and fit data better than the two or one process models.

Khorramdel and von Davier (2014) analyzed multi-scale data consisting of 50 items measuring the Five Factor Model of Personality dimensions. They found that the different BPIs measured the factors (response style or personality) well. A strong general ERS factor

existed behind the e BPIs and indicated that the persons have a strong ERS tendency across personality dimensions. The MRS factor could not be separated as clearly from the trait measure; however, Khorramdel and von Davier (2014) concluded that extreme response style (ERS) and midpoint response style (MRS) could be measured by the e and m BPIs. The estimated latent intercorrelations between the five traits in a model using the d BPIs were lower than the score inter-correlations based on trait estimates from a model using the original 5-point data. This indicated that the M-PM had accounted for ERS and MRS effects, while the other model did not. The three unique BPIs differentiated scoring on the Big Five scales better than the original 5-point items and improved trait measurement for all personality dimensions.

In their study, Plieninger and Meiser (2014) validated use of the M-PM with responses to a self-confidence scale. These researchers decomposed nine seven- point items into four binary pseudoitems (BPIs). The four BPIs respectively represented midpoint preference, direction preference, extreme response option preference, and a central tendency preference (i.e., choosing an option close to the midpoint). See Figure 1 for a diagram of these four processes. A one parameter logistic model was used to model each process. The four latent traits were assumed to follow a multivariate normal distribution. The Bayesian Information criterion (BIC) was used to assess model fit.

Since the trait estimates for the first (midpoint preference) and fourth (central tendency preference) processes were highly correlated, a model constraining the fourth process BPIs to load on the same latent trait as the first process BPIs, was tested. This model fit better than the four process model. The three process model was validated with a model for ERS and MRS traits measured by external heterogeneous items. The M-PM extreme (midpoint) process correlated with the ERS (MRS) trait. Thus, the M-PM could be used to estimate the levels of extreme and midpoint response styles. Additionally, the direction preference items retained the necessary information related to the validity criteria for the measured trait. The researchers concluded that transforming the 7-point responses into simple BPIs retained the necessary construct-information.

Zettler et al. (2015) used the M-PM to study the HEXACO personality inventory which has six domains (Honesty-Humility, Emotionality, eXtraversion, Agreeableness, Conscien-

tiousness, and Openness to Experience Feelings) each consisting of four lower order facet scales. Model fit was assessed with the BIC. The researchers examined correlation patterns of the indifference, direction, and intensity processes across traits. The mean for the bivariate correlations between process estimates from each trait pair was determined within self- and observer reports. The mean correlations for the intensity values showed substantial correlations within observer (.46) and self (.56) which indicated that a person's use of Extreme categories was consistent across traits. Although the mean correlations for indifference were slightly smaller, use of midpoints was also fairly consistent across traits.

Mean correlations concerning the direction process were very similar to mean correlations for raw scores which indicated how the direction process traits could be used to measure the scale trait. The authors concluded that intensity and indifference were mainly person-specific response processes which could carry some variance related to content. The direction process mimics a content-specific response process. The authors concluded that responding to Likert scales involves judgment related to the indifference and intensity processes and the targeted construct.

While these researchers viewed response style traits as continuous variables, other researchers have viewed response styles as a difference in item parameters across response style groups which model response heterogeneity as a function of differences in item characteristics. With this categorical view of response styles, respondents are seen as having a certain response style trait or not. Researchers then use a mixture IRT model to sort the respondents into classes whose members use the response scale format in qualitatively different ways (Wetzel, 2013; Wetzel, Lüdtke, Zettler, & Böhnke, 2015).

2.5 THE MIXTURE IRT MODEL

A mixture IRT model acknowledges that in addition to the latent trait, the observed responses are due to unobserved heterogeneous classes which make up the population. The mixture IRT model combines a latent class model (LCM) and an IRT model. The latent class model assumes that there is a heterogeneous population and that item response pat-

terns determine the composition of the classes, but that there are no residual covariances between items after persons are put into classes (Clark et al., 2013). The mixture IRT model assumes that covariation of observed data within class is explained by one or more continuous latent factors and that unknown population heterogeneity is explained by a categorical latent variable (e.g. Leite & Cooper, 2010).

A mixture IRT model is a type of factor mixture model (FMM). The model assumes that the items measure the trait they were designed to measure; however, the factor structure can be class specific (Cho, 2013; Clark et al., 2013; G. Lubke & Neale, 2008; Sawatzky et al., 2012). Thus, in models which account for response styles, the factor loadings are fixed across classes while the category thresholds can vary. The differences between thresholds indicate how persons in the respective classes use the response categories. Persons in a class with thresholds close together tend to use extreme options while persons in a class with thresholds farther apart tend to use non-extreme options.

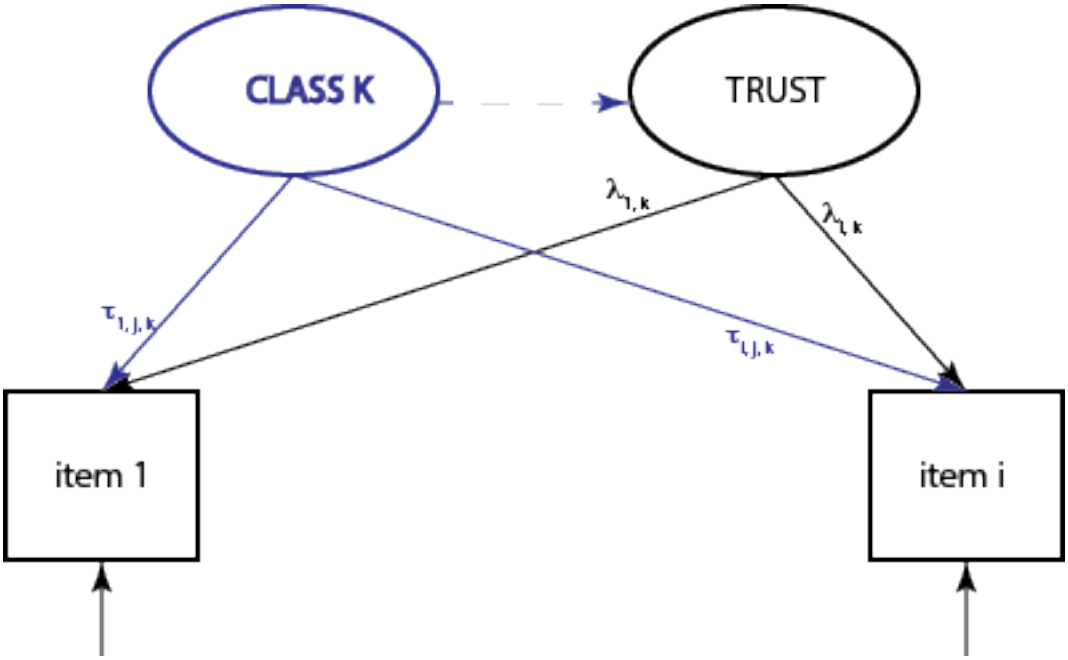
2.5.1 Presentation of the Mixture IRT Model

A mixture model combines a between-class model defining a subject's probability to be in a certain class k (of K classes), π_k, and a within-class probability model defining a data-generating mechanism for subjects in class k (Sterba, 2013). The combination gives a marginal (across class) probability function to explain the observed distribution. Class sizes (mixing proportions) satisfy two constraints: $0 < \pi_k < 1$ and $\sum_{k=1}^{K} \pi_k = 1$.

For the polytomous items in this study, the Graded response model, (Samejima, 1969) described earlier is chosen for the within class probability model. Recall that each item i has J response categories and $J - 1$ category thresholds.

The one factor model shown in Figure 5 is extended to include a latent class variable. Figure 9 shows how this can be done by adding a latent categorical variable (class k or c_k) to the common factor model to show that a finite number of different classes ($k = 1, 2, ..., K$) exist and that different item parameters exist for each class in the population. Now the item factor loadings, λ_{ik}, and category thresholds, τ_{ijk}, have an added subscript k to indicate that the thresholds and loadings are conditioned on unobserved class k membership.

Figure 9: Factor Mixture Model



Note: In this model, the trait (Trust) is impacted by each class K. Each class uses the response options differently. The factor loadings and threshholds can vary by class. τ_{1jk} = Thresholds j for item 1 in class k, τ_{ijk} = Thresholds j for item i in class k, λ_{ik} = factor loading for item i in class k. Adapted from Sawatzky et al. (2012).

For the mixture graded response model (mixGRM), Equation 2.13 is modified to include the unobserved categorical variable c_k with K classes to get equation 2.19:

$$P^*_{ijk}(y_{ni} \geq j|\theta_n, c_k = k) = \frac{e^{\lambda_{ik}\theta_n - \tau_{ijk}}}{1 + e^{\lambda_{ik}\theta_n - \tau_{ijk}}}. \tag{2.19}$$

Equation 2.19 gives the probability of person n in class k responding at or above category j. The probability of a specific response category j is found by the difference $p(y_{ni} = j) = P^*_{ijk}(y_{ni} \geq j) - P^*_{i(j+1)k}(y_{ni} \geq j + 1)$.

For the mixture partial credit model (mixPCM), equation 2.5 is modified to get equation 2.20:

$$P_i(j' = j|\theta_n) = \frac{e^{\sum_{x=0}^{j}(\theta_n - \tau_{ixk})}}{\sum_{s=0}^{M} e^{\sum_{x=0}^{s}(\theta_n - \tau_{ixk})}}. \tag{2.20}$$

In this equation the k is added to the threshold parameter to reflect that the parameter is class specific as in equation 2.19.

Then the within class probability model for a pattern of responses $\mathbf{y_n} = \{y_{ni}\}$ for person n in class k is given by equation 2.21:

$$p(\mathbf{y_n}|c_k = k) = \int (\prod_{i=1}^{I} p(y_{ni}|c_k = k, \theta) f_k(\theta)) d\theta, \tag{2.21}$$

where local independence in the LCM is relaxed and $f_k(\theta)$ is the within class trait distribution (Sterba, 2013).

The between class model is given by the model-implied probability, π_k, that person n belongs to class k. This class probability is found using a multinomial logistic parameterization as in equation 2.22:

$$\pi_k = Pr(c_k = k) = \frac{e^{w(k)}}{\sum_{k=1}^{K} e^{w(k)}}, \tag{2.22}$$

where $w(k)$ is a multinomial intercept and $w(K) = 0$ for identification purposes (Sterba, 2013). The probability gives the class size or proportion of the sample assigned to class k and is also referred to as the class probability.

The unconditional probability for pattern of responses $\mathbf{y_n}$ in the population can be expressed as:

$$p(\mathbf{y_n}) = \sum_{k=1}^{K} \pi_k p(\mathbf{y_n}|c_k = k) = \sum_{k=1}^{K} \pi_k \int (\prod_{i=1}^{I} p(y_{ni}|c_k = k, \theta) f_k(\theta)) d\theta, \tag{2.23}$$

and the items in class k have different parameters (cf., Cho, 2013; Sterba, 2013). The unconditional probabilities are multiplied to form the likelihood function for all of the data.

The class probabilities are used with Bayes' Theorem to determine the posterior probabilities, p_{nk}. The posterior probabilities are used to determine each person's class membership. A person is assigned to the class for which his or her posterior probability is highest. For each person n with response pattern $\mathbf{y_n}$, the posterior probabilities, p_{nk}, of being in a particular class k are given by equation 2.24:

$$p_{nk} = p(c_k = k|\mathbf{y_n}) = \frac{\pi_k p(\mathbf{y_n}|c_k = k)}{p(\mathbf{y_n})}, \tag{2.24}$$

where $p(\mathbf{y_n}) = \sum_{k=1}^{K} \pi_k p(\mathbf{y_n}|c_k = k)$ by the law of total probability (Sterba, 2013). The posterior probabilities are also used in different criteria that help to select the number of classes in the mixture model. Use of a mixture IRT model to account for response styles is described in the following section.

2.5.2 Studies Accounting For Response Styles With Mixture IRT Models

Several researchers have applied mixture Rasch models (Eid & Raubner, 2000; Wetzel, 2013) and this work is reviewed first. Studies that use the mixGRM to account for response styles are then discussed.

The mixPCM is an extension of the partial credit model (PCM, Masters, 1982). In the mixPCM, the item parameters may differ in each of the latent classes (Wetzel et al., 2013). Rost (1997) illustrated use of the mixPCM to analyze 12 items from the Conscientiousness scale of a personality measure. While the AIC identified too many non-interpretable classes, the BIC selected two classes. The larger class (about 65% of the sample) used the response scale in a typical way. The distances between thresholds were reasonably spaced

for each item. The smaller class did not use the response scale in an ordinary way. The thresholds were much closer together than those for the first class. This second class was considered the extreme class and the first class the non-extreme class.

Eid and Raubner (2000) also used the mixPCM in their analysis of a six-item scale assessing employee satisfaction with a work superior. The analysis revealed that the BIC selected a two-class solution over a one or three class solution. The larger group (71% of the company's employees) used the response scale as it was intended. The thresholds between categories were ordered and the fifth category was preferred over the other categories for all items. In the second group, the differences between thresholds were always smaller than one. There were unordered thresholds and the second category was always avoided. The second group tended to use the two extreme categories as indicated by a larger first threshold and smaller last threshold than the corresponding thresholds in the first group.

In another study, Wetzel et al. (2013) used an unconstrained mixPCM and a constrained mixPCM to study responses from 11,724 participants who completed the German *Revised NEO Personality Inventory (NEO-PI-R)* (Ostendorf & Angleitner, 2004) which measures the Big Five Personality traits. In the constrained mixed PCM, the mean of the thresholds for each item (item location) is constrained to be equal across classes.

In the study, the sample was randomly divided into two parts for separate analyses and the results compared. The second part was used to validate the results from the first part. Five personality facets (Positive Emotions, Trust, Altruism, Modesty, and Competence) were removed due to estimation problems in 7 unconstrained or constrained mixture PCM. For 16 of the remaining 25 facets, the constrained model fit better than the unconstrained mixPCM. With the constrained mixPCM, the classes can be compared since the same construct is measured in each class, but the differences in thresholds imply that different response styles are used by each class (Wetzel et al., 2013). With the unconstrained mixPCM, the classes cannot be compared since differences in item location parameters implies that different constructs are being assessed in each class.

For most of the 16 personality facets, a two class solution described the data best. The two classes consisted of persons who preferred extreme options (ERS) and persons who preferred use of more moderate options (NERS). A three class solution fit the data better

for the (Conscientiousness) *Deliberation* and Openness to Experience Feelings *Actions* facets. There was an ERS class while the other two classes differed in how they used the Likert scale midpoint. One class rarely used the midpoint while the other class never did. Model fit was assessed using CAIC.

Recently, Böckenholt and Meiser (2017) used a two dimensional scale to compare the mixture PCM and the Multi-Process Model (M-PM). The Personal Need for Structure (PNS, Neuberg & Newsom, 1993) has two dimensions (response to lack of structure and desire for structure) measured with a six-point response format. The two dimensional PCM was embedded in a mixture Rasch model to examine heterogeneity of threshold parameters (Mixed 2dimPCM). To examine a model with less parameters than the Mixed 2dimPCM, a constrained mixedPCM imposed constraints on the threshold distances across subpopulations. That is, item specific threshold distances in the second class were assumed to be a linear function of the threshold distances in the first class. In both mixture models, the item difficulties (locations) were constrained to be equal across classes.

Along with the M-PM, a constrained M-PM was tested. In the constrained M-PM, the discrimination and difficulty parameters used to measure the two dimensions were constrained to be proportional for the strong and weak attitudinal positions. The constrained mixedPCM fit better than the Mixed 2dimPCM and constrained and unconstrained versions of the M-PM. In the two multi-process models and the mixture models, the correlation between the two dimensions of the PNS was lower than the correlation between the two dimensions in the PCM. Thus, both M-PM and mixture models were useful in accounting for response styles.

Three studies illustrate use of the mixGRM to analyze questionnaire data. In the first study, Egberink et al. (2010) used the mixGRM to study a Conscientiousness scale. This 30-item scale had five subscales (Perfectionism, Organization, Drive, Concentration, and Methodicalness) each with six items. The Drive subscale and two other subscale items were not used due to low reliability and effects of social desirability responding. Although the scale items had five response options, persons rarely endorsed the first two categories and these categories were collapsed.

The researchers selected a four class mixture model since this had better fit criteria and better meaningful interpretation than any of the corresponding criteria and interpretations for other numbers of classes tested. Of the four subscales used, the persons in the four classes differed in their responses to the Perfectionism and Concentration subscales. The persons in the first and third classes were most consistent in their use of the response categories. Class one persons preferred the third category most of the time while persons in class three preferred the fourth category (extreme option) most of the time. The persons in class two preferred category one most often for Perfectionism items and persons in class four often chose category one for the Concentration. For the Organization and Methodicalness scales, all four classes selected the third and fourth categories more frequently than the first two.

In a different study, Sawatzky et al. (2012) used the mixGRM to determine how three different classes could explain the heterogeneity of responses in a 10-item physical functioning subscale. The three classes differed in use of the response scale with the items. The first class tended to use the lower part of the response scale. The second class tended to use the middle to upper parts of the scale. The third class also tended to use the middle to upper parts, but had many trait estimates higher than those of the other two classes. The first class had trait estimates lower than the other two classes.

As part of a study concerning differential item functioning analyses of an adolescent self-regulation questionnaire, Gnambs and Hanfstingl (2014) studied three subscales using the mixGRM. The model identified two classes of students: those using an extreme response style (ERS) and those using a non-extreme response style. A derived ERS score for each person was found using the log odds ratio of being in the ERS class instead of the NERS class. The researchers concluded the response styles showed consistency across the three subscales. This indicates that persons were consistent in their use of response styles across the three different scales.

2.6 SUMMARY OF THE LITERATURE REVIEW

In survey research, use of the common Likert response option format can create conditions under which some respondents may use extreme and midpoint response styles when completing the items. This is problematic since the use of response styles produces a biased summed score and therefore an inaccurate estimate of the measured trait. To account for response styles and provide an improved trait estimate, researchers have developed different methods using extensions to standard IRT models (e.g., PCM, GRM).

One line of research methods involves use of multidimensional IRT (MIRT) models. With MIRT models, the substantive and response style traits are measured by different dimensions. For example, some work has been done with the multidimensional nominal response model (MNRM). Bolt and Newton (2011) extended the work of Bolt and Johnson (2009) so that data from two scales could be used to improve the estimates of the ERS and two substantive latent traits. Falk and Cai (2015) extended the methodology further by using a parameterization of the MNRM which could address six substantive traits and two response styles, in particular ERS and MRS. Their model is perhaps the most flexible of all of the models since it enables modeling of other response styles such as acquiescent and social desirability response styles.

To study the relationship between ERS and MRS, Wetzel and Carstensen (2015) used the multidimensional PCM (MPCM) which is a constrained version of the MNRM. They found that only five of thirty personality traits had small to moderate latent correlations with either MRS or ERS. These studies have not compared the MPCM and MNRM. The MNRM incorporates different item-discrimination parameters for each category in each item which is an advantage over the MPCM since response styles can affect items and category use differentially. Thus, a study comparing the MPCM and MNRM can contribute to research related to using the MPCM and MNRM to address ERS and MRS. Since the MPCM and MNRM are compensatory models, the response style traits and substantive traits combine additively to produce the item response. High values of one trait can compensate for low values on a different trait.

On the other hand, the multi-process model (M-PM) is a partial compensatory MIRT model since the traits do not combine additively. In this model, the substantive, ERS,and MRS trait levels are estimated for each person. The probabilities of using each of the substantive, ERS, and MRS traits are determined. The probabilities are then multiplied to get the probability of a particular item response. The advantage of the M-PM over the MNRM and MPCM is that the M-PM is proposed to explain how persons differ in the use of three sequential decision-making processes: indifference, direction, and intensity. These estimates are proposed to be different from the substantive, ERS, and MRS trait estimates from the MNRM and MPCM because the M-PM and MPCM/MNRM have different assumptions about how the traits combine to produce the response.

The use of the M-PM trait estimates can possibly fill the void of conceptual ideas in explaining person differences in the response process (Zettler et al., 2015). Another key point that Zettler and colleagues emphasize is that the M-PM assumes that intensity and indifference traits may be related to item content. Thus, the intensity and indifference traits may differ among persons yet should be consistent across different scales. The direction traits however should differ across different scales. Khorramdel and von Davier (2014) and Zettler et al. (2015) illustrated use of the M-PM to estimate the three different traits across different personality scales. The M-PM provided trait estimates which had been adjusted for the effects of MRS and ERS. These researchers did not compare use of the M-PM to other MIRT models. Thus, a study that compares the M-PM with other MIRT models would add to the research concerning the M-PM, MPCM, and MNRM.

A different approach for addressing response style use is to use a latent class method. With this approach, the analysis involves a model which uncovers classes that are not directly observed in the population such as groups based on variables such as gender or age. Latent class factor analysis (LCFA), for example, involves substantive and response style dimensions and a categorical variable. Morren et al. (2011) found that the three classes uncovered by an LCFA model differed in both attitudes and level of ERS. The probabilities conditional on amount of extreme options used indicated the likelihood of persons belonging to a particular class. The ERS dimension identified the tendency to choose and avoid use of extreme options.

In a mixture IRT model, the latent continuous response style dimension variable is replaced by a latent categorical variable. As with the LCFA, persons are also assigned probabilities to belong to each of the latent classes. Wetzel (2013) compared MIRT PCM and mixture PCM models for the same set of scales. Her use of mixture Rasch models highlights the importance of using a constrained mixture model to ensure that the trait is measured in the same way across classes. Wetzel et al. (2013) used a random sample of half of the data for the first part of their study. The results showed that for five scales, the occurrence of null categories prevented convergence of either an unconstrained or a constrained mixture PCM. However, more studies using mixture IRT models are needed since few studies exist.

Wetzel (2013) questions if use of the mixture PCM (categorical approach) is sufficient to address those respondents who used the non-extreme (moderate) categories since a two class PCM fit best for most of the facets examined. As she indicates, use of a MIRT model (dimensional approach) models each person's preference for extreme or non-extreme responses. While the categorical approach can distinguish between extreme response style and non-extreme response style (preference for moderate categories), she found it did not identify other response styles such as ARS or MRS. A study comparing the mixture PCM, mixture GRM, and MIRT models such as the M-PM, MPCM, and MNRM would be useful to personality measurement practitioners and researchers.

Wetzel (2013) compared use of the MPCM and mixPCM for personality data; however, she did not compare these models with the MNRM, mixGRM, and M-PM. This study examines how the M-PM, mixGRM, mixPCM, and multidimensional PCM and multidimensional NRM perform in addressing scales with different hypothesized levels of ERS and MRS use.

This study contributes to the research on use of the M-PM in comparison to other MIRT models, specifically the MPCM and MNRM. This study also adds to the research that involves the use of mixture models since it provides a comparison between the mixPCM and mixGRM and to the MIRT models previously described. Thus, this study also contributes to the research involving mixture and multidimensional IRT modes that was begun by Wetzel (2013).

3.0 METHODS

This research examines performance of the multi-process model (M-PM), multidimensional partial credit (MPCM) and nominal response (MNRM) models, and mixture partial credit (mixPCM) and graded response (mixGRM) models. Specifically, the goal of this study is to compare how these five models account for extreme (ERS) and midpoint response style (MRS) use in personality trait measurement. Examining responses to personality questionnaires for detection of possible response style effects is worthwhile to study for three reasons: (1) Organizations are increasingly using personality measures as part of making personnel hiring decisions; (2) Personality measures are contributing to job performance prediction(Rothstein & Goffin, 2006); and (3) Practitioners and researchers are expressing renewed interest in the Five Factor Model (FFM) of personality and relationships between personality and job performance in organizations (Rothstein & Goffin, 2006), in educational settings (Peeters & Lievens, 2005; Pozzebon, Ashton, & Visser, 2014), and in the military (Stark, Chernyshenko, Drasgow, & White, 2012; Zickar & Drasgow, 1996).

In this study, the response data to three personality facet subscales from a Big Five questionnaire are used in the analyses. After the list of research questions is presented, the instrument, sample, and preliminary analyses are described. This is followed by a discussion of the analyses using each model and the statistics used to answer the research questions. Finally, a discussion of the limitations of the study is given.

Table 4 lists the three general research questions pursued in this study. To support the answer for general question A, there are also three specific questions to be addressed. The questions are answered by analyzing data from facet subscales from personality domains measured by the instrument described next.

Table 4: Research Questions Pursued in this Study

A	Does modeling response styles with mixture, multidimensional, and multi-process models improve model-data fit for scales exhibiting Extreme (ERS) or Midpoint Response style (MRS) over the standard IRT models (Partial Credit (PCM) and Graded Response (GRM)?
1	For scales showing ERS and MRS, how do the mixture PCM and mixture GRM compare with each other and with the PCM and GRM in terms of fit ?
2	Does the multi-process model (M-PM) improve the fit of the model to the data over the PCM and GRM for scales exhibiting MRS and ERS?
3	Do the multidimensional PCM (MPCM) and multidimensional nominal response (MNRM) models improve the model-data fit over the PCM, GRM, for scales exhibiting MRS and ERS?
4	How do mixture and multidimensional models compare in explaining variability in item responses?
B1	For the multi-dimensional models, how do the estimated latent correlations between the substantive trait and the response style traits compare with each other?
B2	How do correlations between latent trait estimates based on the different IRT models compare with each other?
C	Which model, the mixture model (mixPCM or mixGRM), a multi-dimensional PCM, multidimensional nominal response model (MNRM), or the multi-process model (M-PM), is best for addressing extreme and midpoint response styles?

3.1 INSTRUMENT

Costa and McCrae (1992) developed the *Revised NEO Personality Inventory (NEO-PI-R)* and this study used subscales of the German version of the inventory (Ostendorf & Angleitner, 2004). The instrument measures the Big Five domains of personality: Extraversion (E), Openness to Experience Feelings (O), Neuroticism (N), Agreeableness (A), and Conscientiousness (C). The five domains each consist of six lower-order facets. The eight items in each facet subscale have five ordered response options (*strongly disagree*, *disagree*, *neutral*, *agree*, and *strongly agree*). The reliability and validity of the NEO-PI-R have been examined in several studies (e.g. Costa & McCrae, 2014).

The Cronbach alpha reliability coefficients for total scores on the 30 facet subscales ranged from .53 to .85 (Wetzel et al., 2013). The Cronbach coefficient indicates the proportion of variance in scale scores that is attributed to the true score (DeVillis, 1991) and is a common measure of reliability in the behavioral sciences.

3.2 SAMPLE

The subjects completing the German *NEO-PI-R* were the nonclinical standardization sample of 11,407 participants. The subjects' ages were between 16 and 60 years ($M = 28.88$, $SD = 10.46$) and the sample was 64% female. The sample represents a subset of a larger sample of 11,724 cases. Persons older than 60 years (317 subjects) were excluded. The data collection for the complete dataset occurred in over 50 separate studies in different places in Austria, Germany, and Switzerland (Wetzel & Carstensen, 2015) from 1992 to 2001.

3.3 SELECTION OF FACET SCALES

This study uses responses to three selected personality facet subscales of the *NEO-PI-R*. The goal was to select different scales with varying degrees of MRS and ERS for comparison of

the mixture, multidimensional and multi-process IRT models. At the same time, facet scales that did not reflect multidimensionality in terms of content were desired. While existence of multidimensionality is expected given the presence of response styles, models used to account for response styles still assume there is no multidimensionality in terms of content.

Moderate to high reliability on the selected facet scales was also desired. The higher the reliability, the less the evaluation of the methods to be compared is affected by the amount of measurement error. The following sections describe how the above goals for choosing facets were met.

3.3.1 Reliability and Exploratory Factor Analysis for Potential Scales

Sixteen facets: Anxiety (N1), Self-consciousness (N4), Assertiveness (E3), Positive Emotions (E6), Openness to Experience Feelings Fantasy (O1), Openness To Experience Feelings (O3), Trust (A1), Altruism (A3) , Compliance (A4), Modesty (A5), Competence (C1), Order (C2), Dutifulness (C3), Achievement striving (C4), Self-Discipline (C5), and Deliberation (C6) were first chosen for potential analyses. The E6, A1, A3, A5, and C1 scales had not been completely analyzed with both constrained and uncontrained mixture models in the Wetzel et al. (2013) study due to model estimation problems. All six facets from the Conscientiousness domain were chosen since this trait is the most vital personality predictor of work performance (Chamorro-Premuzic & Furnham, 2010).

Using *SPSS* (IBM Inc., 2015), the Cronbach α reliablity coefficient for these 16 facets was examined to aid in selecting the facets. Reliabilities from .80 to .90 are considered very good and from .70 to .80 are respectable (DeVillis, 1991). In Table 5, it can be seen that the reliability for all of the scales is .70 or greater, except for A4, C1, C3, and C4.

One and two dimensional exploratory factor analyses were also examined for the scales to examine content-based multidimensionality. Scales were retained if there were significant loadings for all items; if there was a high correlation (approximately .7 or greater) between the two factors in a two factor solution, and if there was a small number of significant loadings (e.g., one or two) for items which loaded on a secondary factor. The scales that were rejected

violated one or more of these criteria. The idea was to select scales where there could be a potential secondary dimension due to person differences and not due to content differences.

Relevant statistics from the EFA for each scale are presented in Table 5. The table has seven scales in bold font for discussion about scale rejection and acceptance. The Anxiety (N1), Assertiveness (E3), Openness to Experience Feelings (O3), Modesty (A5), and Self-Discipline(C5) scales have significant loadings for all scale items. The Compliance (A4) and Deliberation (C6) scales have significant loadings on seven of the eight items. This can be seen by examining the numbers of significant loadings on the first and second factor which are presented in the fifth and sixth columns. For most scales, the first factor has the higher number of significant loadings. For the Openness to Experience Feelings, Modesty, and Self-Discipline scales, there are four significant loadings on each factor. Thus, it was important to also examine the correlation between the two factors.

Most of the scales have a high correlation between the two factors. The scales for which the correlation is not as strong are Modesty (.47) and Deliberation (.63).

The ratios of the first to second and second to third eigenvalues are shown in the third and fourth columns of Table 5. When the first eigenvalue is large relative to the second and the second is not very big compared to the others, the set of items is approximately unidimensional (Lord, 1980). As can be seen from the table, the first of the two ratios is nearly four times the second for facets N1, E3, O3, and C5. The two selected facets from the Agreeableness Domain, A4 (Compliance) and A5 (Modesty), and C6 (Deliberation) from the Conscientiousness domain show a ratio of first to second eigenvalue that is not as large as it is for the other four selected facets.

3.3.2 Response Category Use for Potential Scales

Table 6 illustrates another rationale for subscale selection and involved the use of midpoint and extreme response categories. In this table, the final selected scales are presented in boldface font and reflect a subset of the seven scales identified in Table 5. The scales are the Anxiety (N1), Openness to Experience Feelings (O3), and Compliance (A4) facets. The scale Cronbach alpha values were 0.76 for Openness to Experience Feelings, 0.63 for Compliance,

Table 5: Facet Subscale Exploratory Factor Analysis Summary

		Eigenvalue Ratio		Sig. Loadings		
NEO Facet(Domain)	Rel.	1st/2nd	2nd/3rd	1st Fac.	2nd Fac.	2 Factor r
N1 Anxiety(N)	**.82**	**4.68**	**1.15**	**5**	**3**	**.75**
N4 Self-consciousness(N)	.72	3.06	1.23	6	2	.63
E3 Assertiveness(E)	**.81**	**4.14**	**1.07**	**7**	**1**	**.68**
E6 Positive Emotions(E)	.80	2.82	1.88	4	4	.51
O1 Fantasy(O)	.81	3.74	1.24	5	2	.59
O3 Feelings(O)	**.76**	**4.16**	**1.07**	**4**	**4**	**.76**
A1 Trust(A)	.76	2.93	1.43	4	3	.58
A3 Altruism(A)	.70	2.71	1.08	6	2	.45
A4 Compliance(A)	**.63**	**2.53**	**1.12**	**5**	**2**	**.71**
A5 Modesty(A)	**.74**	**2.55**	**1.50**	**4**	**4**	**.47**
C1 Competence(C)	.65	2.77	1.09	4	3	.57
C2 Order(C)	.70	2.92	1.06	4	4	.57
C3 Dutifulness(C)	.67	3.10	1.08	3	3	.69
C4 Achievement Striving(C)	.69	2.81	1.11	4	2	.60
C5 Self-discipline(C)	**.81**	**4.27**	**1.20**	**4**	**4**	**.71**
C6 Deliberation (C)	**.75**	**3.54**	**1.06**	**5**	**2**	**.63**

Note: N = Neuroticism, E = Extraversion, O = Openness to Experience Feelings, A = Agreeableness, C = Conscientiousness. Rel. = Cronbach Alpha Reliability. 1st Fac. = Number of significant factor loadings on primary factor, 2nd Fac. = Number of significant loadings on secondary factor, 2 Factor r = correlation between two factors in the two factor EFA solution.

and .82 for Anxiety. The subscale reliability was good for the Anxiety scale and respectable for the Openness to Experience Feelings scale, but not as high for the Compliance scale. The Compliance scale was included, despite the low reliability, because of the low response style use that was found.

The final three facets were also chosen since they reflected different uses of the response scale and are hypothesized to show possibly different levels of response style use. As the analyses presented below indicate, the Compliance scale tended to show potentially low use of extreme and midpoint response styles. The Openness to Experience Feelings scale showed low potential use of MRS and marked potential use of ERS. The Anxiety scale tended to show medium potential use of MRS and medium potential use of ERS.

To explore the differential use of response options across the scales, the proportions of extreme and midpoint responses used by each person for each facet scale were determined to get a measure of the scale-specific response styles for each facet. The MRS proportion (TMRS) was found by determining the number of midpoint responses in a set of items for each person and dividing by the number of items. The ERS proportion (TERS) was calculated similarly. The number of highest and lowest categories endorsed were combined when calculating measures of extreme response for subjects for two reasons: (1) it is unlikely that persons ignore scale content altogether; and (2) Persons who tend to give extreme negative responses also tend to give extreme positive ones (Baumgartner & Steenkamp, 2001).

For the 16 facets, the percentages of midpoints and extreme options used were also calculated. First the total number of responses in each of the five categories was found. The five totals were summed to get the total number of responses used for the scale. Then the total number of midpoints used was divided by the total number of responses used and converted to a percentage. To get the percentages of extreme options used, the total number of high extreme and low extreme options used were combined to get the total number of extreme options used. The total number of extremes used was divided by the total number of responses used and converted to a percentage.

These percentages are shown in the last two columns of Table 6. The percentages of midpoints and extreme options used were all greater than 10% and less than 31%. To facilitate the discussion, percentages less than 20% designated low use of the categories; percentages between 20% and 30% designated medium use of the categories. Percentages greater than 30% indicated relatively high use of the categories.

As can be seen in the table, the Anxiety (N1) scale has the highest percentage of midpoints used of the other three selected scales. The Openness to Experience Feelings (O3) Scale has the highest percentage of extreme options used. The use of the midpoint in the Openness to Experience Feelings scale is less than its use in the other four selected scales. Due to the differences in use of midpoints and extreme options in the scales, persons may have used different response styles.

With the percentages though, there are two limitations. One does not know whether the extreme or midpoint responses occur due to stylistic or substantive reasons. Also, there is no standard to determine if a given percentage of extreme or midpoint responses is low or high. Therefore, correlation analyses were also done to argue for the potential use of response styles with the scales.

With correlations, there are common decision rules suggested by many scholars (De Beuckelaer, Weijters, & Rutten, 2010; Franzblau, 1958). Correlations with magnitude between 0 and .2 are considered negligible. Then, the next three designations are, respectively, .2-.4, low; .4-.6, moderate; .6-.8, marked. Lastly, those correlations with magnitude between .8 and 1.0 are considered strong or high.

The Pearson correlations between the summed score and proportions of midpoint and extreme responses are presented in Table 6. As can be seen from this table, the Compliance score has negligible correlation with the proportion of midpoints (r = .02) and extreme options (r = -.10) used. There were slightly higher but still negligible correlations for the Anxiety score (r = -.11 for Sum and Midpoint proportion and r = .15 for sum score and Extreme proportion).

The Openness to Experience Feelings sum score has a moderate, negative correlation with proportion of midpoints used (r = -.59) and a marked, positive correlation with proportion of extreme options used (r = .74). Thus, the Compliance and Anxiety facets show

negligible correlation with simple measures of MRS and ERS, while the Openness to Experience Feelings has a higher correlation with at least one of the response styles examined. These statistics help to justify use of the chosen facet scales to compare the performance of the models in this study.

3.3.3 Demographic Variables and Potential Use of Response Styles

Since previous research has found relationships between use of response styles and demographic variables, the potential uses of response styles with respect to age and gender were each examined. The Pearson correlations between age and proportions of Midpoints and Extreme Categories used are presented in Table 7. As can be seen in this table, most of the correlations between age and the response style measure are negligible. None of the correlations is greater than .16 in size for the chosen scales. This indicates that age does not explain much variability in the use of midpoints and extreme categories in the proposed dataset.

The use of different categories with respect to gender was also examined. The percentages of midpoints used by each group are given in Table 8. As can be seen in this table, males and females tend to use about the same percentage of midpoints for the Assertiveness (25%) scale. For the Deliberation scale, males used 0.23% more midpoints than females (20.45%) and males used 1.05% more midpoints than females for the Self-Discipline scale. Males used 1.39% more midpoints than females used (17.40%) for the Compliance (A4) scale. Males used 2.56% more midpoints than females used (20.45%) for the Anxiety scale. While these differences may not reflect practical differences, the differences between the two groups for the Openness to Experience Feelings scale is larger. Males used 6.94% more midpoints than females (11.58%) for the Openness to Experience Feelings scale.

The percentages of extreme options used by each group are given in Table 9. For the Self-Discipline scale, males used 1.36% more extremes than females used (12.13%) and used 0.74% more extremes than females for the Deliberation scale. Males used 0.60% more extremes than females (17.42%) for the Compliance scale. On the other hand, females used only slightly more extremes (13.33%) than males (12.88%) for the Anxiety scale. Females used 8.27%

Table 6: Subscale Rationale Summary Based on Category Use

Facet(Domain)	Rel.	S-TMRS r	S-TERS r	Mid Use	Ext Use
N1 Anxiety(N)	**.82**	**-.11**	**.15**	**21.37% M**	**13.17% L**
N4 Self-consciousness(N)	.72	-.05	.15	22.43% M	15.42% L
E3 Assertiveness(E)	.81	.04	-.09	25.16% M	12.74% L
E6 Positive Emotions(E)*	.80	.42	.57	16.11% L	24.58% M
O1 Open to Fantasy(O)	.81	-.44	.58	18.91% L	21.93% M
O3 Openness to Feelings(O)	**.76**	**-.59**	**.74**	**14.08% L**	**25.24% M**
A1 Trust(A)*	.76	-.21	.15	25.36% M	10.73% L
A3 Altruism(A)*	.70	-.54	.61	21.16% M	16.49% L
A4 Compliance(A)	**.63**	**.02**	**-.10**	**17.90% L**	**17.64% L**
A5 Modesty(A)*	.74	-.12	.24	30.40% H	12.04% L
C1 Competence(C)*	.65	-.52	.48	27.25% M	11.71% L
C2 Order(C)	.70	-.23	.31	16.86% L	18.61% L
C3 Dutifulness(C)	.67	-.36	.64	15.52% L	24.59% M
C4 Achievement Striving(C)	.69	-.26	.38	20.88% M	19.27% L
C5 Self-discipline(C)	.81	-.19	.22	20.51% M	13.07% L
C6 Deliberation(C)	.75	.07	.06	20.53% M	12.62% L

Note: N = Neuroticism, E = Extraversion, O = Openness to Experience, A = Agreeableness, C = Conscientiousness. Rel. = Cronbach Alpha Reliability. S-TMRS r = Pearson Correlation between Sum score and Proportion of Midpoints Used; S-TERS r = Pearson Correlation between Sum score and Proportion of Midpoints Used. Mid Use = percentages of scale midpoints used; Ext Use = percentages of extreme options used on scale. L = Low Use, M = Medium Use, H = High Use. *Due to the existence of null categories which led to problems with mixture model estimation, these facets had not been analyzed by Wetzel et al. (2013).

Table 7: Correlations between Age and Midpoint and Extreme Proportions

Facet	Age-TMRS r	Age-TERS r	Facet	Age-TMRS r	Age-TERS r
N1	**-.01**	**-.04**	N4	-.01	-.06
E3	-.10	.02	E6	.08	-.18
O1	.10	-.17	**O3**	**.10**	**-.16**
A1	-.08	-.04	A3	-.04	-.02
A4	**-.02**	**-.10**	A5	-.04	-.01
C1	-.12	.03	C2	-.07	.01
C3	-.13	.14	C4	-.01	-.09
C5	-.09	.01	C6	-.06	-.04

Note: N1 = Anxiety, N4=Self-consciousness, E3 = Assertiveness, E6 = Positive Emotions, O1 = Openness to Experience Feelings Fantasy, O3 = Openness to Experience Feelings , A1 = Trust, A3 = Altruism, A4 =Compliance, A5 = Modesty, C1 = Competence, C2 = Order, C3 = Dutifulness, C4 = Achievement Striving, C5 = Self-discipline, C6 = Deliberation. Age-TMRS r = Correlation between Age and Proportion of Midpoints Used; Age-TERS r = Correlation between Age score and Proportion of Midpoints Used. N = 11,407 for all facets.

more extremes than males (19.95%) for the Openness to Experience Feelings scale. The more practical difference with respect to gender occurred in the Openness to Experience Feelings scale.

3.3.4 Preliminary Data Analyses Identifying Possible Use of Response Styles

To illustrate that there are different groups of respondents who tended to use some categories more than others (i.e., use the response scale differently) for different scales, K-means clustering was used (Dolnicar & Grün, 2007; Sarstedt & Mooi, 2014). Since the K-means algorithm can be sensitive to the starting values used (Steinley, 2003), ten thousand different initial random seeds were used to examine two and three group solutions using MATLAB (MathWorks, 2016). Use of the software's parallel processing enables the choice of the seeds (initial randomizations) and algorithm execution to be processed quickly.

A "good" cluster analysis depends upon the clustering variables , the number of clusters, a distance measure, and cluster validation (Steinley & Brusco, 2011). The cluster membership for the groups was formed based on the proportion of midpoints used and proportion of extreme options used. These proportions were determined for each person for each scale separately.

The software uses a two step process (batch updates and online updates) for the K-means algorithm. The first phase estimates a solution that is used in the second phase. The second phase is used to find the solution with a global minimum. The solution with the minimum total distance (between and within clusters) was selected by the software.

The minimum total distance is used to find the potentially best clustering solution. The best total distance for two and three cluster solutions for the chosen scales are presented in Table 10. The total distance is based on the between cluster distance between cluster centroids and the within cluster distance from cases in cluster to their cluster centroid. As the table illustrates, the three cluster K-means solutions have a smaller total distance than the two cluster solution for each scale. For example, the Anxiety scale has a minimum total distance of 473.8 between clusters in the two group solution and a minimum total distance 248.0 for the three group solution. The table also indicates that the two cluster solutions

Table 8: Group Differences in Midpoint Use based on Gender

Differences in Percentages of Midpoints Used					
Scale	Males	N	Females	N	z-score
N1 Anxiety	**23.01%**	32,816	**20.45%**	58,290	9.06**
N4 Self-consciousness	23.45%	32,819	21.86%	58,326	5.52**
E3 Assertiveness	**25.11%**	32,817	**25.19%**	58,303	**-0.28**
E6 Positive Emotions	18.11%	32,815	14.98%	58,290	12.34**
O1 Open to Fantasy	20.57%	32,801	17.98%	58,259	9.58**
O3 Open to Feelings	**18.51%**	32,820	**11.58%**	58,306	28.88***
A1 Trust	26.15%	32,809	24.91%	58,292	4.13**
A3 Altruism	23.23%	32,819	20.00%	58,290	11.46**
A4 Compliance	**18.79%**	32,819	**17.40%**	58,314	5.25**
A5 Modesty	30.69%	32,801	30.24%	58,296	1.42
C1 Competence	26.03%	32,810	27.93%	58,275	−6.18**
C2 Order	17.80%	32,816	16.33%	58,339	5.69**
C3 Dutifulness	15.94%	32,824	15.28%	58,280	2.64**
C4 Achievement Striving	21.70%	32,821	20.41%	58,326	4.60**
C5 Self-Discipline	21.18%	32,807	20.13%	58,302	3.77**
C6 Deliberation	**20.68%**	32,824	**20.45%**	58,319	**0.83**

Note: The number of males was 4,108 and number of females was 7,299. N = Total Number of Responses.
*p<.05 ; **p<.01 ; ***p<0.00001.

Table 9: Group Differences in Extreme Options Use based on Gender

Differences in Percentages of Extreme Options Used					
Scale	Males	N	Females	N	z-score
N1 Anxiety	**12.88%**	32,816	**13.33%**	58,290	**-1.93**
N4 Self-consciousness	14.61%	32,819	15.72%	58,326	-4.47^{**}
E3 Assertiveness	**13.05%**	32,817	**12.57%**	58,303	2.09^{*}
E6 Positive Emotions	22.72%	32,815	25.648%	58,290	-9.83^{**}
O1 Open to Fantasy	19.96%	32,801	17.98%	58,259	7.36^{**}
O3 Open to Feelings	**19.95%**	32,820	**28.22%**	58,306	-27.58^{***}
A1 Trust	10.59%	32,809	10.81%	58,292	-1.03
A3 Altruism	14.90%	32,819	17.39%	58,290	9.72^{**}
A4 Compliance	**18.02%**	32,819	**17.42%**	58,314	2.28^{*}
A5 Modesty	11.91%	32,801	12.11%	58,296	-0.87
C1 Competence	13.94%	32,810	10.46%	58,275	15.68^{**}
C2 Order	18.95%	32,816	18.40%	58,339	2.05^{*}
C3 Dutifulness	24.04%	32,824	24.91%	58,280	-2.93^{**}
C4 Achievement Striving	19.41%	32,821	19.19%	58,326	0.81
C5 Self-Discipline	13.54%	32,807	12.80%	58,302	3.18^{**}
C6 Deliberation	**13.49%**	32,824	**12.13%**	58,319	5.94^{**}

Note: The number of males was 4,108 and number of females was 7,299. N = Total Number of Responses.
$^{*}p<.05$; $^{**}p<.01$; $^{***}p<0.00001$.

Table 10: Best Total Distance for One to Three K-means Cluster Solutions

Scale	K = 1	K = 2	K = 3
N1 Anxiety	812.3	473.8	248.0
O3 Open to Feelings	1150.1	493.8	303.3
A4 Compliance	645.1	374.2	224.3

Note. The best total distance shown for each scale and number of groups K is based on 10,000 replications (random initial starting values).

are better than one cluster solutions. For the one cluster solutions, the best total distance is always greater than the distances for the two and three cluster solutions. The anxiety (N1) scale has a total distance 812.3 for the one group solution.

For the one group solution, the Openness to Experience Feelings (O3) scale has a total distance of 1,150.1. For Compliance (A4), the total distance is 645.1. For the three cluster solutions, the total distances for the O3 and the A4 scales are 303.3 and 224.3, respectively. These are lower than the corresponding one and two total distance solutions for these scales.

The sizes of the groups formed by K-means clustering were different depending on the scale content and respondent characteristics (i.e., response scale use) as in previous research (e.g., Dolnicar & Grün, 2007). The different sizes for the analyses with two group solutions can be seen in appendix G. Category use for the scales for the two K-means solutions can be seen in appendix H.

The different sizes and respondent characteristics for the K means analyses with three groups for the chosen scales are presented in Table 11. The three groups are designated as Midpoint (in which the Midpoint is preferred over Extremes), Extreme (the Extremes are preferred over a Midpoint), and General (the Agree or Disagree options are preferred over the Midpoint and Extremes). The General groups were the largest for all scales. The sizes of the Midpoint and Extreme groups were different from the General groups and differed

across scales. For the sake of discussion, less than 25% of the sample represents a small group; between 25% to 50% of the sample is a medium group; and greater than 50% is a large group.

The Anxiety (N1) scale had a medium sized Midpoint group that was almost twice as its small Extreme group (26.7% versus 14.2%). The Anxiety general group was larger than the other two and large in size (59.1%). The Openness to Experience Feelings (O3) scale medium sized Midpoint group was slightly larger (25.7%) than its small Extreme group (23.1%). The Openness to Experience Feelings general group was larger than the other two and large in size (51.2%). The medium-sized Compliance (A4) Midpoint Group was more than twice as large (38.6%) as its small Extreme Group (18.6%). The A4 general group was medium sized (42.9%).

The Midpoint and Extreme groups are further differentiated by the mean proportions of Midpoints and Extremes used across scales. For the sake of discussion, the mean proportions are described with language similar to correlation description (e.g., .2-.4, low; .4-.6, moderate; .6-.8, marked). Table 11 also shows the mean proportion of midpoints used per person (M TMRS) and the mean proportion of extremes used per person (M TERS) in each of the response style groups (Midpoint, Extreme, and General). The M TMRS is always negligible (less than .20) for the Extreme Group for a scale, while the M TERS is always negligible for the Midpoint Group as would be expected due to each group's category preference. The M TMRS and M TERS are always negligible in the General group due to this group's preference for the Agree and Disagree categories.

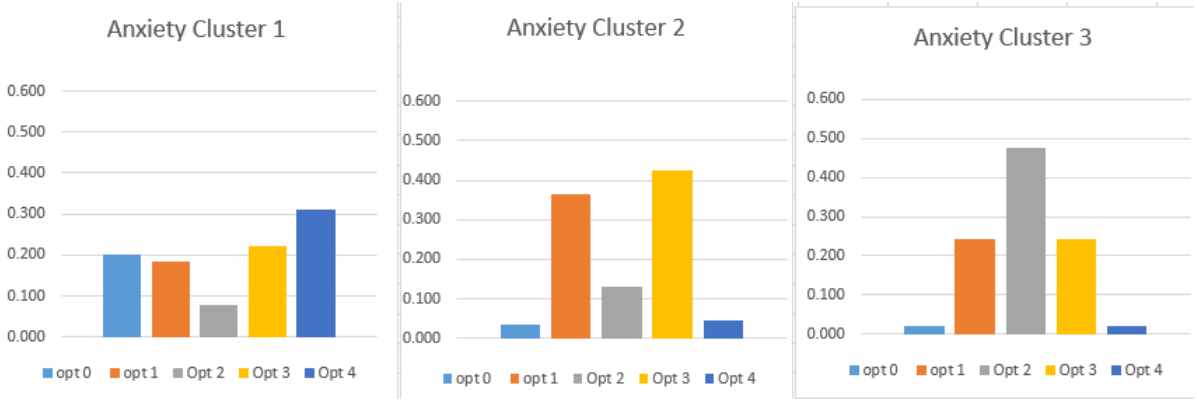
Comparing the mean proportions across the three different groups within scale illustrates the characteristics of the persons in the groups. For the Anxiety (N1) scale, the M TMRS is .47 for the Midpoint group which is believed to indicate moderate use of MRS. This is higher than the M TMRS for the Openness to Experience Feelings (O3) Scale (.38) and Compliance (A4) Scale (.34) which are presupposed to indicate low use of MRS. The persons tend to use a midpoint less for these scales than the persons in the midpoint group for the Anxiety scale. For Compliance The midpoint class is larger (38.6%) than it is for the Anxiety (26.%) and Openness to Experience Feelings (25.7%) scales.

Table 11: K-means Cluster Results for Three Different Response Style Groups

Scale	Midpoint Size TMRS, TERS M(SEM)	Extreme Size TMRS, TERS M(SEM)	General Size TMRS, TERS M(SEM)
N1 Anxiety	26.7% **.47(.002)**, .04(.001)	14.2% .08(.003), **.51(.004)**	59.1% .13(.001), .08(.001)
O3 Open to Feelings	25.7% **.38(.003)**, .05(.002)	23.1% .04(.001), **.67(.003)**	51.2% .07(.001), .17(.002)
A4 Compliance	38.6% **.34(.002)**, .10(.002)	18.6% .09(.002), **.47(.003)**	42.9% .07(.001), .12(.001)

Note. The percentage of the sample (N = 11,407) assigned to the group designated as Midpoint (Midpoint preferred over Extremes, Extreme (Extremes preferred over Midpoint), or General (the Agree and Disagree options were preferred over the Midpoint and Extremes). TMRS = proportion of midpoints used by person within scale, TERS = proportion of extremes used by person within scale, M = Mean of the proportion of midpoints(Extremes) used, SEM = Standard of the mean.

Figure 10: Anxiety (N1) Item Category Use by Three Different Response Style Groups



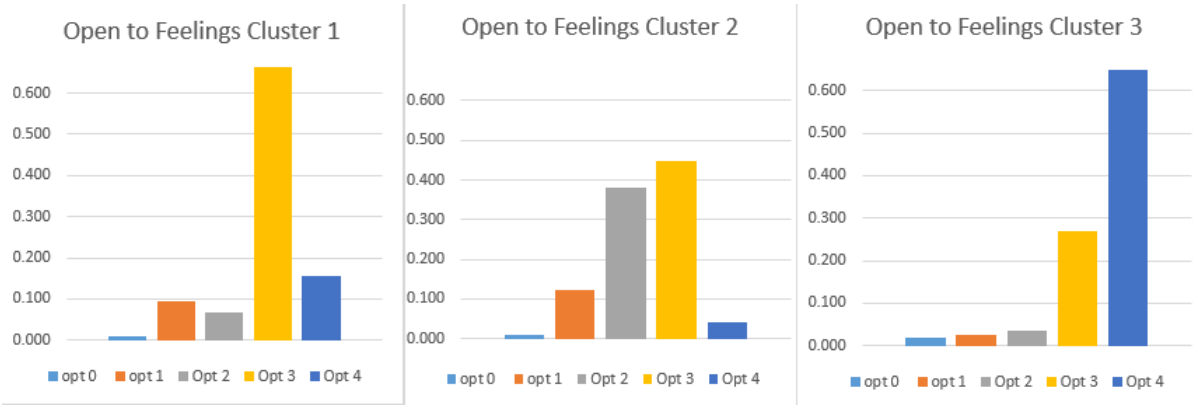
Note: Sample sizes for the groups are $N_{Ext} = 1620$, $N_{Gen} = 6745$, and $N_{Mid} = 3042$. These, respectively, represent 14.2 %, 59.1 %, and 26.7% of the total sample.

Regarding ERS, the Anxiety and Compliance scales are thought to have medium use of ERS as indicated by the moderate M TERS (.51 and .47, respectively). The Openness to Experience Feelings scale is presupposed to have marked use of ERS since its M TERS is .67. The Openness to Experience Feelings scale has a larger Extreme group (23.1%) than the Anxiety (14.2%) and Compliance (18.6%) scales do.

The standard errors for the mean proportions are given as well to illustrate the precision of the estimates for response style use. For the M TMRS and M TERS, the standard errors ranged from .001 to .004.

Therefore, for each scale, three different groups of respondents were assumed to generate the data. Each group used the response scale differently with one group tending to use the midpoint over the extreme options. A second group preferring extreme options over the midpoint. The general (third) group tended to prefer the agree or disagree categories over the others as indicated by the low M TMRS and M TERS for each scale (.08-.14).

Figure 11: Openness to Experience Feelings (O3) Item Category Use by Three Different Groups



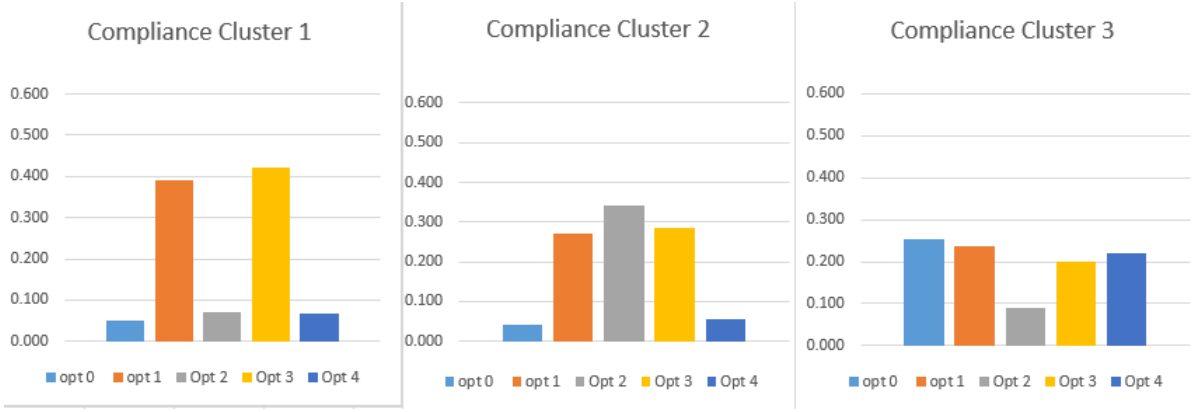
Note: Sample sizes for the groups are $N_{GenAgree} = 5842$, $N_{MidAgree} = 2932$, and $N_3 = 2633$. These, respectively, represent 51.2%, 25.7%, and 23.1% of the total sample.

Figure 10 shows how respondents in three different groups used the categories for items in the Anxiety scale. The first group uses the extreme options the most and the midpoint the least of the five possible responses. The second group tends to use the Agree and Disagree categories (options 1 and 3) more than the other categories. The third group tends to use the midpoint more than the extreme options.

Category use by three groups for the Openness to Experience Feelings (O3) items can be seen in Figure 11. The distribution of responses is negatively skewed for the first and third groups. As can be seen in the figure, the first group tended to use the extreme options more than the other groups. The second group tends to use the Agree option (option 3) more than the other groups. The third group tends to use the midpoint (option two) more than the other groups.

Category use by the three groups for the Compliance (A4) items can be seen in Figure 12. The second group tends to use the midpoint more than the other two groups. The first group tends to use the Agree and Disagree options more than the other categories. The third group tends to use the midpoint less than the other four categories. The third group also tends to use the extremes more than the other groups.

Figure 12: Compliance (A4) Item Category Use by Different Response Style Groups



Note: Sample sizes for the groups are $N_{Gen} = 4890$, $N_{Mid} = 4401$, and $N_{Ext} = 2116$. These represent, respectively, 42.9%, 38.6%, and 18.6% of the total sample. Gen = General, Mid = Midpoint, Ext = Extreme.

Although K-means clustering can be used to show different groups of respondents, it is limited since it does not have a rigorous statistical method for supporting the number of classes while latent class analysis (LCA) does (Magidson & Vermunt, 2002). With LCA, the "choice of clustering criterion is less arbitrary and includes rigorous statistical tests" according to Magidson and Vermunt (2002). Similar to LCA, the mixture model analyses in this study also involved using statistical tests to help support the selection of the number of groups.

Finally using the correlation analyses and the K-means cluster results for the three groups, some possible response style effects are described after examining the size of the groups and the mean TMRS and TERS for each response style group. This is done assuming that the size, M TMRS and M TERS will determine the impact on the models. The potential midpoint and extreme response style effects for the scales are given in Table 12.

The purpose of the descriptions is to describe response style use qualitatively and to make connections with results from the models. Note also that these statements are only tentative since they are based on the mean proportions of response options used (M TMRS, M TERS) which are possibly tainted by response style effects. Persons may select an extreme

Table 12: Possible Effects due to Use of Response Styles in Scales

Scale	Potential Response Style Effects	
N1	Medium sized Group w/ Moderate MRS	Small Group w/ Moderate ERS
O3	Medium sized Group w/ Low MRS	Small Group w/ Marked ERS
A4	Medium sized Group w/ Low MRS	Small Group w/ Moderate ERS
N1	Negligible r between Score and MRS	Negligible r between Score and ERS
O3	Moderate, negative r between Score & MRS	Marked, positive r between Score & ERS
A4	Negligible r between Score and MRS	Negligible r between Score & ERS

Note: The descriptions were made based upon examining results of correlation and K means cluster analyses using mean proportions of extreme and midpoint response options used. N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance, MRS = Midpoint Response Style, ERS = Extreme Response Style. r = correlation, Score = Sum Score.

option (or midpoint) due to the substantive trait, the response style trait, or a combination of both. The effects of the substantive and response style traits cannot be separated with any simple model; an IRT model is required.

Mixture models are used to address statements concerning response style groups. There is no estimated latent correlation between substantive and response style traits as there is for multidimensional models. The MIRT models are used to obtain the estimated latent correlations between substantive and response style traits. The MIRT model response style estimates were also used with K-means clustering to demonstrate the existence of three different response style groups for each scale. Scales are examined with the following analyses.

3.4 MIXTURE MODEL ANALYSES

Although much work has been done with the mixture PCM, the PCM has a discrimination parameter that is common to all items. Since the item discrimination parameter in IRT is analogous to the item-test biserial correlation (Yen & Fitzpatrick, 2006), the item to test biserial correlations for the above facets were examined for amount of variation. The range of these correlations for the items for each scale follow: Anxiety (.49 to .65), Compliance

(.18 to .44), and Openness to Experience Feelings (.34 to .60). The variation in these biserial correlations implies using a two parameter IRT model such as the graded response model could provide better fit to the data.

For the chosen scales, a one class GRM and PCM were estimated. Then constrained and unconstrained mixture models were examined to test if the scales measured the trait in the same way for all participants in each class. In a mixture model analysis, better fit of a constrained model over an unconstrained model implies that the trait is measured in the same way for all persons across classes. When the unconstrained model fits better, the trait is measured differently across the classes. In an unconstrained mixture model, differences between latent classes can be interpreted as content-related differences (such as different traits being measured) and/or content-unrelated differences (e.g. differences in response scale usage, Wetzel, Carstensen, & Böhnke, 2013). When the item discrimination and difficulty parameters and factor (trait) covariance matrix are free to vary across classes, the trait might not be measured in the same way or possibly could be different within each class (Clark et al., 2013). The parameter constraints used in this study are discussed further in the section below. Only trait estimates from the constrained mixture model should be used to compare trait levels of persons in different latent classes (Wetzel et al., 2013). Thus, it was important to check that a constrained model fit better than an unconstrained model for the scales in this study to compare the two different mixture IRT models.

The responses to the chosen personality facet scales were examined with the mixture partial credit model (mixPCM, Rost, 1991; Wetzel, Carstensen, & Böhnke, 2013) and the mixture graded response model (mixGRM, Samejima, 1969, Sawatzky, Ratner, Kopec, & Zumbo, 2012). Estimating the mixture models required numerical integration since the likelihood function for the responses involves an integral over the trait distribution in each class. To maximize the log likelihood of responses, the Maximum likelihood robust (MLR) option in *Mplus* (L. K. Muthén & Muthén, 2010) was used for estimation. The MLR estimator has different algorithms such as Newton-Raphson, Expectation-Maximization (EM), quasi-Newton, and Fisher scoring, any of which could be selected by default when the software performs iterations (Han & Paek, 2014). Twelve processors was used to speed up the estimation. To be sure that the best log-likelihood was replicated, the analyses were performed

by generating 500 different random starting values in the initial stage and 100 optimizations carried out in the final stage. With the software, the standard rectangular integration algorithm with 15 quadrature points was used.

3.4.1 Estimation and Model Selection Criteria

First the PCM and GRM were fit to the data and measures of fit were calculated for the one class models. Then the number of classes was increased by one each time and the model was estimated. In the unconstrained mixture PCM (mixPCM), the factor loadings (item discriminations) were constrained to be equal and the common factor loading and class thresholds were estimated. In the constrained mixPCM, the factor loadings were constrained to be equal and the means of the item thresholds were constrained to be equal across classes. The common factor loading and thresholds were estimated. For the constrained mixture GRM (mixGRM), the factor loadings were forced to be equal across classes and the thresholds were allowed to vary in each class (Egberink et al., 2010; Grove, Baillie, Allison, Baron-Cohen, & Hoekstra, 2015). This was done to maximize classification of the persons into classes (Grove et al., 2015; G. H. Lubke & Muthén, 2005) and to ensure that the latent factor measured by the items in each class was the same (Gnambs & Hanfstingl, 2014). In the unconstrained mixGRM, the loadings and thresholds were allowed to vary across classes. For the estimated models, the factor mean was set to zero and the variance was set to 1 in each class.

The unconstrained and constrained MixPCM and mixGRM for two to three classes were fit to the data. Although previous analyses with the mixPCM used six classes, as the number of classes increases, the classes became more difficult to interpret. Three classes were expected to be sufficient to detect differences in response style use in this study.

In the constrained mixPCM, the item locations were constrained to be equal across classes. The item thresholds could vary within each class, but the item locations (means, or equivalently, the sums of item thresholds) were forced to be equal between classes so that the measured trait was the same across classes. That is, the differences in classes were due

Table 13: Model Selection Criteria to Determine Number of Classes in Mixture Model

Criterion Type	Criteria	Model Selection Rule
Information	AIC, BIC, ssBIC	Least
Statistical	VLMR, aVLMR, BLRT	$p \leq 0.05$ for alternate model
Classification Quality	sEn	Greatest
Interpretability of the solution	Mean class assignment probabilities	Best if the diagonal probabilities are greater than .8 for high quality classification

Note: AIC = Akaike Information criterion (Akaike, 1974), BIC = Bayesian Information criterion (Schwarz, 1978), ssBIC = sample size adjusted BIC (Sclove, 1987), VLMR = Vuong Lo Mendell Rubin test (Lo, Mendell, & Rubin, 2001), aVLMR = adjusted VLMR (Lo et al., 2001), BLRT = Bootstrap Likelihood Ratio Test (McLachlan & Peel, 2000; Nylund, Asparouhov, & Muthén, 2007), sEn = entropy (B. Muthén et al., 2002; Ramaswamy, DeSarbo, Reibstein, & Robinson, 1993).

to response style effects only and not some other trait being measured (Wetzel et al., 2013). In the mixGRM, the factor loadings (item discriminations) were constrained to be equal across classes for the same reason.

Choosing the number of latent classes was based on the model selection criteria in Table 13. Three information criteria, used in many studies of model fit, were based on lowering the log-likelihood function by adding a term related to number of model parameters or sample size. These were the Akaike Information Criterion (AIC, Akaike, 1974; Burnham & Anderson, 2002),the Bayesian Information Criterion (BIC, Schwarz, 1978), and the sample-size adjusted BIC (ssBIC, Sclove, 1987).

Since information criteria can often conflict in suggested number of classes, it was important to consider other criteria. Three other approaches involve a statistical method used to estimate the likelihood ratio test statistic. The Vuong-Lo-Mendell-Rubin (VLMR; Lo, Mendell, & Rubin, 2001) test is an extension of a generalized likelihood ratio test to compare two competing models (Vuong, 1989). An adjusted VLMR test (aVLMR; Lo, Mendell, & Rubin, 2001) uses a corrected test statistic based on sample size and number of estimated

parameters (Henson, Reise, & Kim, 2007). These two tests have been criticized for selecting too many classes (Jeffries, 2003). The Bootstrap Likelihood Ratio Test (BLRT; McLachlan & Peel, 2000; Nylund, Asparouhov, & Muthén, 2007) assesses the p value of the LRT statistic.

Efficient use of the adjusted Vuong Lo Mendell Rubin Likelihood Ratio Test (VLMR LRT) and the Bootstrap Likelihood Ratio Test (BLRT) was implemented with the OPTSEED option (Asparouhov & Muthén, 2012; L. K. Muthén & Muthén, 1998-2012). When warnings were indicated with the BLRT to do so, the number of LRT random starts (for the K class model) was increased .

3.4.2 Checking Statistics based on Interpretations of Classes

Other model selection approaches involved the set of posterior probabilities for each person being in each latent class (equation 2.24) which were part of the output for a model with $K > 1$ classes. The posterior probabilities were used to determine two statistics that are based on clear interpretation of the solution and classification uncertainty of class members. The entropy measure is used to assess a proposed model's ability to separate classes (Celeux & Soromenho, 1996). It is a measure of the classification quality of the solution from the model. While various entropy criteria exist (Henson et al., 2007; Peugh & Fan, 2013), this study used a scaled entropy measure (sEn) to compare models with $K > 1$ (B. Muthén et al., 2002; Ramaswamy et al., 1993). sEn was given by the equation:

$$sEn = 1 - \frac{-\sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}\ln(p_{nk})}{N\ln(K)},$$

where p_{nk} is the posterior probability of membership of person n in class $k > 1$ for a mixture model with $K > 1$ classes. Higher values of sEn, on the 0 to 1 scale, indicate better classification quality.

Another part of the mixture model *Mplus* output (L. K. Muthén & Muthén, 2010), related to model fit evaluation and solution interpretation, is the set of average latent class probabilities for most likely class membership or mean class assignment probabilities (MCAPs). These probabilities are the means of the probabilities for persons to be in each latent class.

Thus, for a K-class model, a KxK table was formed. The MCAPs on the main diagonal are associated with classification quality or reliability. For high classification quality, these probabilities should have been greater than or equal to 0.8 (Geiser, 2013).

The model selection criteria summarized in Table 13 were examined to help select the number of classes. The proportions of midpoints and extreme options used in each class were determined to help interpret the classes. After consideration of both model selection criteria and class interpretation, the number of classes in the mixture model was chosen.

Once the number of classes for the mixture model was selected, the OPTSEED feature was used to run an analysis with the starting values to produce the best loglikelihood to get the trait estimates (L. K. Muthén & Muthén, 1998-2012). The software produced expected a posteriori (EAP) estimates for the person trait (factor score).

3.4.3 Comparing fit of the one class models (PCM, GRM) and mixture models (mixPCM, mixGRM) to the data

To answer research question 1, the fit of the PCM, GRM, mixPCM, and mixGRM to the data was first assessed using the Information criteria from the PCM, GRM, and selected mixture models. For the models, univariate model fit at the item level was assessed by comparing observed and expected item response proportions. The item residuals and standardized residuals were examined. The overall univariate Pearson χ^2 statistic was checked for significance.

Bivariate model fit (local independence) was assessed using the joint distributions of the observed and predicted item responses for each pair of items. The bivariate Pearson χ^2 statistic for each item pair was given in the software output. The total bivariate Pearson χ^2 statistic from the item pairs was examined, yet could not be compared to a known distribution (McCrea, 2013). For two models, the model with the lower bivariate Pearson χ^2 value was interpreted as the one with better fit.

For the cells of the bivariate cross table of scores, *Mplus* (L. K. Muthén & Muthén, 1998-2012) also calculated the Bivariate standardized Pearson residuals (Agresti, 1996). The percentage of absolute values of the standardized residuals greater than 3 was calculated to

determine how many were significantly large since values greater than 3 were considered large (Agresti, 2010; McCrea, 2013). The model with the lowest percentage of of Absolute standardized bivariate residuals greater than 3 was considered to fit best. The results from the mixPCM and mixGRM were compared to determine if a two parameter model was better than a one parameter model.

3.5 MULTI-PROCESS MODEL ANALYSES

While the mixture model takes a categorical approach and classifies persons based on the response heterogeneity and presence of a response style or not, the multi-process model (M-PM) assumes that persons respond to an item using the three processes (indifference, direction, and intensity) described in the previous chapter. A person decides if he/she is indifferent or not (ie., chooses a midpoint or not). If a person has an opinion, then he/she expresses the agreement or disagreement (ie., makes a direction choice). Lastly, if the person expressed the (dis-)agreement, then he/she decides how strong the opinion is held (ie., indicates an intense/extremeness choice or not).

For analysis with the three process M-PM model, the original response data was recoded into binary pseudoitems (BPIs). The recoding was done as described in the left side of Table 14. A two parameter logistic model (2PLM, Khorramdel & von Davier, 2014) was used due to the variation in item biserial correlations described earlier in section 3.4.

To check whether a two process model fit better than the standard and three process models, a two process model which involved one response process (either indifference (MRS) or intensity (ERS), exclusively) was estimated. While one process modeled use of the response style (e.g., MRS), the other modeled use of direction and the other response style. This involved using a binary pseudoitem for the response process and an ordinal three-point or four-point pseudoitem which captured the direction and other response process (cf., Böckenholt, 2012). The recoding for the two process models is presented in Table 15.

Table 14: Recoding Five-point Likert data into Binary Pseudo-items for Three Process Model

Opt	BPI 1 m	BPI 2 d	BPI 3 e	Category option Probabilities
0	0	0	1	$Q_i(a_i^1(\theta_n^1 - b_i^1))Q_i(a_i^2(\theta_n^2 - b_i^2))\Pr_i(a_i^3(\theta_n^3 - b_i^3))$
1	0	0	0	$Q_i(a_i^1(\theta_n^1 - b_i^1))Q_i(a_i^2(\theta_n^2 - b_i^2))Q_i(a_i^3(\theta_n^3 - b_i^3))$
2	1	-	-	$\Pr_i(a_i^1(\theta_n^1 - b_i^1))$
3	0	1	0	$Q_i(a_i^1(\theta_n^1 - b_i^1))\Pr_i(a_i^2(\theta_n^2 - b_i^2))Q_i(a_i^3(\theta_n^3 - b_i^3))$
4	0	1	1	$Q_i(a_i^1(\theta_n^1 - b_i^1)\Pr_i(a_i^2(\theta_n^2 - b_i^2))\Pr_i(a_i^3(\theta_n^3 - b_i^3))$

Note: Opt = Category Option. Dashes show that data are missing by design.
BPI = Binary Pseudo Item
$\Pr_i(a_i^h(\theta_n^h - b_i^h))$ is the probability of using process h.
$\mathrm{Q}_i(a_i^h(\theta_n^h - b_i^h))$=1- $\Pr_i(a_i^h(\theta_n^h - b_i^h))$ is probability process h is not used.
Adapted from Böckenholt (2012, p. 668).

The latent trait distributions for indifference (MRS), direction (substantive trait), and intensity (ERS) processes in the two and three process models were standardized with means set to 0 and variances set to 1 to identify the models. The correlations between these latent traits were computed as part of the output.

The AIC, BIC, and ssBIC for the different models were compared to address research question 2. The best fitting model was determined by choosing the model with the lowest information criteria.

The Maximum likelihood robust (MLR) option in *Mplus* (L. K. Muthén & Muthén, 1998-2012) was used to estimate the models. Four processors were used to speed up the estimation. The best fitting two or three process model was used to get the expected a posteriori (EAP) estimates for the person traits. *flexMIRT* was used to get the M_2 statistic since this statistic is not provided in *Mplus*. It is a useful statistic for assessing absolute fit of a model. See appendices C and D for example code.

Table 15: Recoding Five-point Likert data into Pseudo-items for Two Process Models

Indifference Process (Mid-PM)			
Opt	MBPIm	FOPIod	Category option Probabilities
0	0	0	$Q_i(a_i^1(\theta_n^1 - b_i^1))P_i(a_i^2(\theta_n^2 - b_i^2))$
1	0	1	$Q_i(a_i^1(\theta_n^1 - b_i^1))P_i(a_i^2(\theta_n^2 - b_i^2))$
2	1	-	$\Pr_i(a_i^1(\theta_n^1 - b_i^1))$
3	0	2	$Q_i(a_i^1(\theta_n^1 - b_i^1))\Pr_i(a_i^2(\theta_n^2 - b_i^2))$
4	0	3	$Q_i(a_i^1(\theta_n^1 - b_i^1)\Pr_i(a_i^2(\theta_n^2 - b_i^2))$
Intensity Process (Ext-PM)			
Opt	EBPIe	TOPIod	Category option Probabilities
0	1	0	$P_i(a_i^1(\theta_n^1 - b_i^1))P_i(a_i^2(\theta_n^2 - b_i^2))$
1	0	0	$Q_i(a_i^1(\theta_n^1 - b_i^1))P_i(a_i^2(\theta_n^2 - b_i^2))$
2	-	1	$\Pr_i(a_i^2(\theta_n^1 - b_i^1))$
3	0	2	$Q_i(a_i^1(\theta_n^1 - b_i^1))\Pr_i(a_i^2(\theta_n^2 - b_i^2))$
4	1	2	$P_i(a_i^1(\theta_n^1 - b_i^1)\Pr_i(a_i^2(\theta_n^2 - b_i^2))$

Note: Opt = Category Option. Dashes show that data are missing by design.
MBPI = Binary Pseudo Item for Indifference Process
FOPI = Four-point Ordinal Item for Direction
EBPI = Binary Pseudo Item for Intensity Process
TOPI = Three-point Ordinal Item for Direction
$\Pr_i(a_i^h(\theta_n^h - b_i^h))$ is the probability of using process h.
$\mathrm{Q}_i(a_i^h(\theta_n^h - b_i^h))$=1- $\Pr_i(a_i^h(\theta_n^h - b_i^h))$ is probability process h is **not** used.
Adapted from Böckenholt (2012, p. 668).

3.6 OTHER MULTI-DIMENSIONAL MODEL ANALYSES

To compare the Multi-process model (M-PM) with other multidimensional IRT models and to replicate the results of Wetzel and Carstensen (2015), the data were analyzed with the Multi-dimensional Partial Credit Model (MPCM) using *flexMIRT* (Houts & Cai, 2015). Two and three dimensional partial credit models which involved the substantive trait, and at least one or both of the midpoint and the extreme response style traits were estimated to check if the three dimensional model fit better than two dimensional and standard IRT models and the mixPCM.

Although including a multi-dimensional graded response model (MGRM, De Ayala, 1994; Muraki & Carlson, 1995) was also initially considered for comparison with the mixture GRM, the MGRM does not have scoring functions to handle response style effects. The MPCM and the reparameterized Multi-dimensional Nominal Response Model (MNRM, Falk & Cai, 2015; Thissen & Cai, 2016) do, however, have the necessary scoring functions that allow the researcher to fix the order of the item categories and thereby, model known response styles. Therefore, the data were also analyzed using the MNRM so that the results of this model could be compared with the M-PM and other models. *flexMIRT* allows use of scoring functions for category slopes that are separate from overall item slopes. See appendices E and F for example code.

The MPCM constrains item slopes to be equal across items for each of the substantive and response style dimensions. Since the item slopes are constrained to be equal within a dimension, the trait is assumed to affect each item in the same way. The scoring functions (order of the categories) are fixed to identify and interpret known response styles and the substantive trait responses.

The MNRM is different from the MPCM since its freely estimated item (and category) slopes enable a researcher to test whether the trait affects items differentially or not for the substantive and response style traits. Although freely estimated scoring function constraints for the response style dimensions are possible, this was not done. Instead, the scoring function constraints were fixed as they were with the MPCM so that known response style traits can be identified and interpreted.

The three dimensional models involved the substantive trait and both midpoint and extreme response style traits. The overall item slopes for the substantive dimension were freely estimated in each of the models; however, the overall item slopes for the MRS and ERS dimensions were constrained to be equal or freely estimated for the three dimensional models. There were four different versions of the three dimensional model with substantive, midpoint, and extreme response style traits that were tested. The models differed regarding constraints on the overall item category slopes: freely estimated for each dimension (F F F), constrained equal for the second (MRS) dimension and freely estimated for the others (F E F), constrained equal for the third (ERS) dimension and freely estimated for the others (F F E), and freely estimated for the substantive (first) dimension and constrained equal for the MRS and ERS dimensions (F E E).

Additionally, two dimensional models which involved the substantive trait and one response style trait (MRS or ERS) were tested to see if either one fit better than the three dimensional model which involved both response style traits. For the two dimensional models involving one response style only, the category slopes were freely estimated. The information criteria were compared to find the best fitting MNRM of all tested models.

The information criteria for the MPCM and MNRM were compared with those for the M-PM, PCM, and GRM to assess model fit. The EAP trait estimates were found for each trait using the best fitting three dimensional model of each type (M-PM, MPCM, and MNRM).

3.7 MODEL FIT ANALYSES

The information criteria were examined for the standard PCM and GRM models and the five different IRT models (mixture and multidimensional) with response style dimensions. The information criteria assess relative fit, that is, to see how one model compares with another for each facet. It is also important to examine how well the model can reproduce the response data, that is, to assess absolute fit.

For the standard and multi-dimensional models, the absolute fit was assessed using the M_2 statistic (Cai & Hansen, 2013; Maydeu-Olivares, 2013; Maydeu-Olivares & Joe, 2006). This statistic is a limited information statistic based on the first and second moments. This statistic is recommended over the G^2 and Pearson's χ^2 statistics for situations when there are large and/or sparse contingency tables, as in this study. The M_2 statistic is approximately distributed as a χ^2 random variable. The M_2, the Root Mean Square Error of Approximation (RMSEA), and the 95% confidence intervals for the RMSEA from the PCM and GRM, M-PM, MPCM, and MNRM analyses with *flexMIRT*. The RMSEA values and confidence intervals were compared. Models where the confidence intervals limits were below .05 were believed to fit better than models which did not.

Absolute fit indices such as the M_2 statistic and the RMSEA are not available directly for the mixture models in *Mplus*. Global model fit can only be obtained indirectly using a simulation study to obtain a parametric bootstrap (Geiser, 2013). Instead, the amount of absolute standardized bivariate residuals greater than three for the standard IRT and mixture models were compared. The model with the least amount was understood to fit better than the other models.

3.8 EXAMINING MODEL BASED RESPONSE STYLE USE

To answer **research question B1**, the estimated latent correlations between substantive and response style traits for each model were examined. The size and signs of the correlations were compared to each other within facet. These correlations are expected to be different from the correlations between sum score and mean proportion response style measures since the IRT model has separated the response style effects from the substantive trait effects. The results from the multi-dimensional models were compared with the findings of Wetzel and Carstensen (2015). The estimated latent correlations are not available from mixture models for the unidimensional scales. The response style groups from the mixtures can be examined

3.9 MULTI-DIMENSIONAL MODEL AND MIXTURE MODEL COMPARISONS

To examine the statements made about potential effects of response styles, the mixture models were examined for the size of the classes and the mean proportions of midpoints and extremes used in each class. The response style trait estimates from multi-dimensional models were also used to form groups using K-means clustering. The sizes and mean traditional response style measures of each group were compared with the size and mean TMRS and TERS of the classes formed by the mixture models. This was done to show that response style groups differ across traits and possibly due to the models. One reason why MIRT response style estimates are useful is to control for the effects of substantive trait level on the selection of midpoint or extreme options which have been removed. There are some limitations to using these estimates and these are described in the limitations section.
To address **research question B2**, correlations between model substantive trait estimates and between model response style trait estimates were examined. The correlations between model substantive trait estimates were expected to correlated highly since the models accounted for response style use, although in different ways.

For the mixture models, it was expected that the probability to belong to the extreme class would exist for each person due to the findings of Wetzel et al. (2013). This probability was used as an estimate for the ERS trait. The probability of being in an extreme class was correlated with the intensity trait estimates from the M-PM and with the ERS estimates from the multidimensional models to see if the correlations were high. A high correlation was interpreted that the respective models could provide comparable estimates for extreme response style.

If a class could be interpreted as a group of persons who overused the midpoint, the probability for being in this class was correlated with the indifference or MRS trait estimates from the M-PM or multi-dimensional IRT model. The correlations between the probability to be in a non-extreme class and the M-PM indifference or multi-dimensional MRS trait estimates were examined to see if the correlations were large. If any of these correlations were large, then the respective models could provide comparable estimates for midpoint response style.

The correlations between response style (RS) trait estimates were compared to see if any differences could be determined from the models. The multiprocess model (M-PM) RS estimates are expected to be different from the other MIRT model RS estimates since the M-PM is a noncompensatory model which estimates a response process. The other MIRT models are compensatory models in which estimate ERS and MRS tendencies for each person.

To address general **research question C** of which model is best for addressing extreme and midpoint response styles, the following comparisons were made for each scale. First, the information criteria were compared for all of the models to see which was the best fitting model for each substantive trait.

The amount of explained variability in the responses due to the models was examined when the standard models were nested in the mixture or MIRT models. This was possible for the mixture models, Multidimensional PCM, and Multidimensional NRM, but not for the Multi-Process Models. Explained variability in this study is the general coefficient of determination R^2 (Nagelkerke, 1991).

Next, the correlations from **research question B2** were compared for differences in size to see what possible conclusions could be drawn. Then, the different features of the models were compared. That is, the models were summarized by the approaches each takes for addressing response styles and by the available output from the models.

3.10 SUMMARY OF SUBSCALE SELECTION AND PURPOSE OF STUDY

To summarize: the purpose of this study is to examine how five IRT models (mixPCM, mixGRM, M-PM, MPCM, and MNRM) account for ERS and MRS in three personality subscales. The three subscales were chosen using Exploratory Factor Analysis, Correlational analyses, and K-means clustering analyses. The anxiety (N1) scale was chosen since it seemed to exhibit moderate use of MRS and low use of ERS. The Openness to Experience Feelings (O3) scale was chosen since it appeared to exhibit low use of MRS and marked use of ERS. The Compliance (A4) scale was chosen since it appeared to exhibit low use of MRS and low use of ERS. The complex IRT models in this study are expected to provide a better picture of potential response style use and substantive trait estimates than standard IRT models.

4.0 RESULTS

In this chapter, the results for the analyses in this study are presented. This study compared the trait estimates from five different IRT models which account for midpoint (MRS) and extreme (ERS) response styles. The mixture Partial Credit and mixture Graded Response models, the Multi-Process Model, the Multidimensional Partial Credit Model, and the Multidimensional Nominal Response Model were examined.

The estimates from fitting mixture and multidimensional models are proposed over using subscale sum scores to estimate persons' trait levels since the models account for response style effects. The subscales (Anxiety, Openness to Experience Feelings, and Compliance) were chosen from the NEO revised Personality Inventory (Costa & McCrae, 1992) since they were believed to illustrate different possible effects due to MRS and ERS. Table 16 summarizes the possible response style effects across the analyzed subscales.

First the fit criteria for the mixture and multi-dimensional models are presented and compared to the fit criteria for the unidimensional models. The amount of large standardized bivariate Pearson residuals from the mixture models are also compared to the amount of large standardized bivariate residuals from the standard IRT models. Also, within each scale the correlations between model substantive trait estimates and correlations between model response style trait estimates are examined to compare the models.

This is followed by an examination of the model based output concerning the use of response styles. The classes of the mixture models are examined for size and mean proportions of midpoints and extreme options used. For the MIRT models, the estimated latent correlations between substantive and response style traits are presented and compared. For the mixture and MIRT models, statements regarding response style effects are made since

Table 16: Possible Effects due to Use of Response Styles in Scales

Scale	Potential Response Style Effects	
N1	Medium sized group w/ Moderate MRS	Small group w/ Moderate ERS
O3	Medium sized group w/ Low MRS	Small group w/ Marked ERS
A4	Medium sized group w/ Low MRS	Small group w/ Moderate ERS
N1	Negligible r between Score and MRS	Negligible r between Score and ERS
O3	Moderate, negative r between Score & MRS	Marked, positive r between Score & ERS
A4	Negligible r between Score and MRS	Negligible r between Score & ERS

Note: The statements were made based upon examining results of correlation and K means cluster analyses using mean proportions of extreme and midpoint response options used. N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance, MRS = Midpoint Response Style, ERS = Extreme Response Style. r = correlation, Score = Sum Score.

the estimates are based on models which account for response styles. This differs from possible effects based on mean proportions of response options used in the previous chapter and presented in Table 16.

4.1 COMPARISONS OF MODELS ACROSS SCALES

Research questions and results first address subsets of models. Then at the end of the section all models are compared against each other. This is done since some fit statistics across the models can differ. For example, the mixture models involve statistics which are used to help determine the number of classes while the multidimensional models do not. *Mplus* was needed to estimate the mixture models, while *flexMIRT* was needed for the multidimensional IRT models. Only the standard IRT and multi-process models could be estimated in both software packages.

There are also differences in what statistics are readily available from the software output. For example, the *Mplus* output for the mixture model does not provide any absolute fit statistics such as the RMSEA as *flexMIRT* does. The *flexMIRT* output does not directly provide the bivariate standardized Pearson residuals which are available in *Mplus* output

for the standard and mixture IRT models. Furthermore, examining the amount of absolute bivariate standardized Pearson residuals (ABSPR) greater than three for the multi-process model (M-PM) does not make sense. For model comparisons, the residuals should look at the same features of the data. The M-PM models in this study used binary, three-point, and four-point pseudoitems which differ from the original five-point item responses. Since the pseudoitems capture different functions of the data, examining the amount of ABSPR > 3 is not useful for the M-PM.

The model selection criteria are presented for the standard IRT and constrained mixture IRT models. Using the constrained mixture model trait estimates ensures that the trait is measured the same way across the different classes which is needed for comparing trait estimates (Wetzel et al., 2013). The same trait is measured across different classes which differ in use of response styles. With unconstrained mixture models, the effects due to measuring substantive and response traits are confounded. The measured trait may be different across classes or persons in the classes may view the items differently. Thus, model comparison criteria, class assignment probabilities, and distribution of response options by class assignments for the constrained mixPCM and mixGRM models are compared. Since the adjusted Vuong-Lo-Mendell-Rubin Test (aVLMR) and Bootstrap Likelihood Ratio Test (BLRT) were not helpful in selecting the number of classes, these are not presented. Finally, the information criteria indicated similar results, thus the BIC is provided for all models in this study since this is a commonly used measure.

4.1.1 Mixture Model Results

4.1.1.1 Anxiety subscale(N1) For the Anxiety subscale, traditional analyses expected the presence of the following response styles: a Medium sized group using moderate MRS and a small group with moderate ERS use. A negligible correlation between Anxiety and MRS traits was expected. A negligible correlation between Anxiety and ERS traits was also expected.

Table 17: Mixture Model Selection Criteria for Anxiety Facet

	Anxiety Facet			
K (Model)	BIC	sEn	MMCAP	ABSPR > 3
1(GRM)	230,152	—	1	38.9%
1(PCM)	232,563	—	1	51.4%
2(mixGRM)	227,645	.51	.82	14.1%
2(mixPCM)	229,001	**.61**	**.85**	25.3%
3(mixGRM)	**227,003**	.44	.69	**5.7%**
3(mixPCM)	228,065	.50	.74	14.9%

Note: K = Number of classes, BIC = Bayesian Information Criterion, sEn = scaled entropy, MMCAP = Minimum Diagonal value of Mean Class Assignment Probabilities table. ABSPR = Percent of Absolute Bivariate Standardized Pearson residuals that are large (i.e., above 3). The values in boldfont indicate that the respective number of classes, K, be considered. 2(mixGRM) = two class constrained graded response model.3(mixPCM) = three class constrained partial credit model.

The model selection criteria for the Anxiety facet are presented in Table 17. As can be seen, the BIC for the constrained three class GRM (227, 003) is lower than the BIC for the constrained two class GRM (227, 645).

The second and third columns of Table 17 provide the scaled entropy (sEn) and the Minimum Diagonal value of the Mean Class Assignment Probabilities table (MMCAP, See also Table 18). The mean class assignment probabilities (MCAPs) indicate the classification quality and are used to determine the entropy statistic. The values of the entropy and MMCAPs for the three class mixPCM and mixGRM solutions were lower than those for the two class solutions. Lower values of entropy indicate that the overall classification quality for the three class solution is not as good. Thus, while the information criteria suggested a three class solution fit the data better, the classification quality statistics indicated that a two class solution fit better. Therefore, the use of categories and classification quality for the two and three class solutions were explored further.

Table 18: Mean Class Assignment Probabilities tables for the Anxiety scale

N1 2mixGRM			N1 2mixPCM			N1 3mixGRM				N1 3mixPCM			
	E	N		N	E		E	G	M		E	M	G
E	**.82**	.18	N	**.90**	.10	E	**.78**	.14	.08	E	**.82**	.04	.15
N	.13	**.87**	E	.15	**.85**	G	.11	**.69**	.20	M	.03	**.78**	.19
						M	.06	.20	**.74**	G	.08	.18	**.74**

Note: 2mixGRM = two class constrained graded response model. 3mixPCM = three class constrained partial credit model. E = Extreme class, N = Non-extreme class, M = Midpoint class, G = General class. Probabilities in bold indicate persons are classified with high probabilities in the respective class.

Table 18 contains the Mean Class Assignment Probability tables for two and three class solutions for the Anxiety scale. Results for the two class solutions (2mixGRM and 2mixPCM) are shown in the left side of the table. The values on the main diagonal for the tables in the two class solutions are .82 or higher and this indicates good classification (Geiser, 2013). There is higher classification quality for the 2mixPCM solution than for the 2mixGRM. Thus, there is somewhat less overlap between the classes for the 2mixPCM than for the 2mixGRM. The MCAP tables for the Anxiety scale under three class mixture GRM (3mixGRM) and three class PCM (3mixPCM) can be seen in the right side of Table 18. The diagonal MCAP values for the E, M, and G classes (.82, .78, and .74) for the 3mixPCM are generally greater than those for the mixGRM for the **respective** classes (.78,.74, and .69). These indicate higher classification quality for the mixPCM solution than for the mixGRM. With lower diagonal MCAPs for the three class solutions, the classification quality is not as good for the three class solutions as it is for the two class solutions. This might be expected since as the number of classes increases, the classes are not separated as clearly (Geiser, 2013). The classification uncertainty can be seen in the off-diagonal elements of the tables. As might be expected, the persons in the Extreme and Midpoint classes have negligible MCAPs to be assigned to the Midpoint and Extreme classes respectively. Similar results occur for the 3mixGRM.

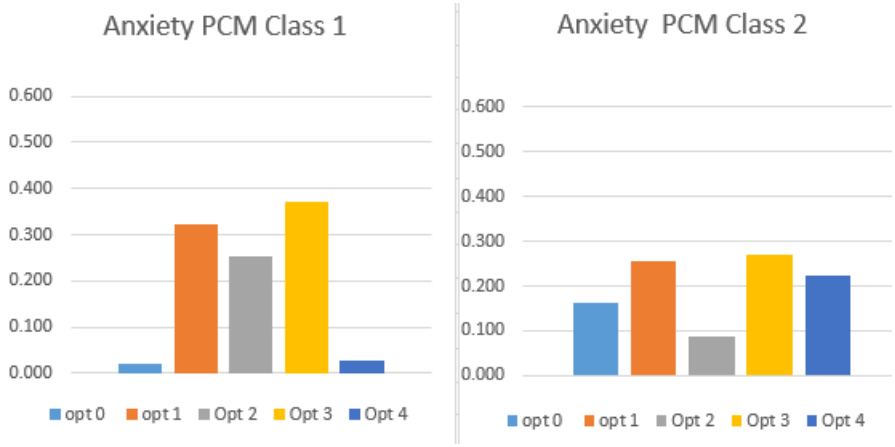
The fourth column of Table 17 shows the amount of large absolute bivariate standardized Pearson residuals (ABSPR) for each model. From this column, it can be seen that the improvement in fit of the GRM to the data over the PCM and improvement in fit of the mixture models over the standard IRT models. As can be seen, the PCM and GRM did not predict the joint distributions of items as well. The amount of large standardized bivariate residuals is less for the GRM (38.4%) than the PCM (51.4%). Note that because these standard models did not predict the joint distribution of items well, a violation of local independence of the items was indicated (Sawatzky et al., 2012).

Since the Anxiety and other scales were assumed to be unidimensional in terms of item content due to the analyses in the previous chapter, the idea that measurement model parameters could be noninvariant across latent classes was proposed to explain the violation of local independence. The amount of absolute standardized residuals > 3 for the joint distributions of items was examined for the two and three class mixture models. The three class mixture model had a smaller amount of ABSPR greater than three than the corresponding two class mixture model (e.g., 25.3% vs. 14.9% for mixPCM). Additionally, the mixGRM fit better than the mixPCM for the same number of classes. For example, the amount of large ABSPR was 5.7% for the 3mixGRM compared to 14.9% for the 3mixPCM.

The distribution of categories used in the 2mixPCM solution for the Anxiety scale is presented in Figure 13. In the figure, the first class uses the midpoint more than the second class; however, the first class prefers the Agree and Disagree options. The second class prefers extremes more than the first class, yet also tends to prefer Agree or Disagree options over all others.

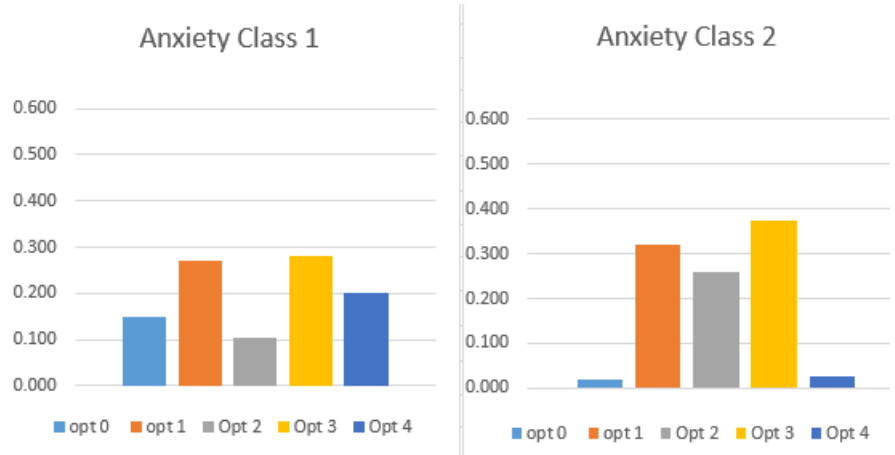
The distribution of categories used in two class mixGRM solution for the Anxiety scale is presented in Figure 14. In the figure, the first class uses more extreme options than the second class yet prefers the Agree and Disagree options. The second class uses the midpoint more than the first class; however, the second class prefers the Agree and Disagree options. The two class GRM mixture for the Anxiety scale indicated category use similar to the two class PCM mixture.

Figure 13: Anxiety (N1) Item Category Use for Two Class PCM mixture



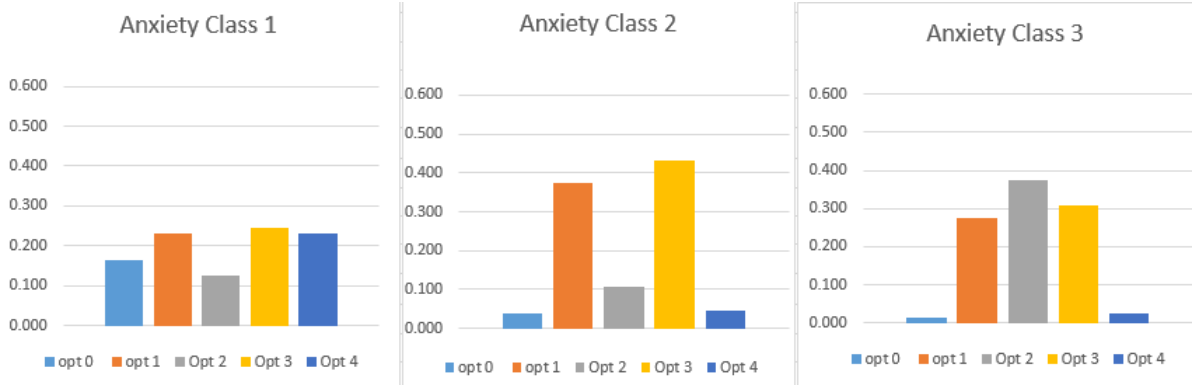
Note: Class sizes are $N_1 = 8612$ and $N_2 = 2795$. These represent 75.5% and 24.5% of the sample.

Figure 14: Anxiety (N1) Item Category Use for Two Class GRM mixture



Note: Class sizes are $N_1 = 3240$ and $N_2 = 8167$. These represent 28.4% and 71.6% of the sample.

Figure 15: Anxiety (N1) Item Category Use for Three Class GRM mixture

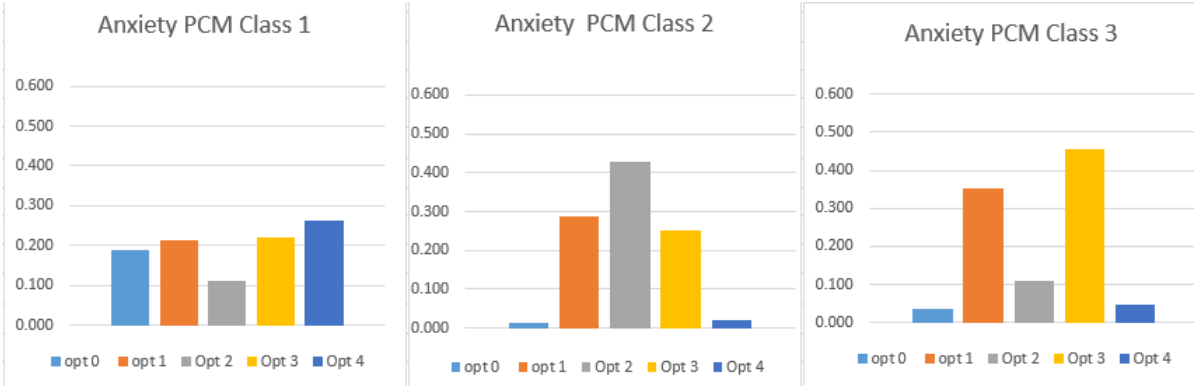


Note: Class sizes are $N_1 = 2314$, $N_2 = 4729$, and $N_3 = 4364$. These represent 20.3%, 41.5%, and 38.3% of the sample.

Figure 15 shows how respondents in the three different classes used the categories for the Anxiety scale under the 3mixGRM. These classes are interpreted as Extreme (E), General (G), and Midpoint (M) classes, respectively, due to the preferred use of categories in each class compared to the other classes. The first (E) class used the extreme categories more often than the other two classes. The second (G) class used the Agree and Disagree categories (options 1 and 3) more than the other classes. The third class used the midpoint more than the other five categories and used the extreme options the least. This third (M) class also used the midpoint more often than the first two classes.

The category use for the three class mixture PCM can be seen in Figure 16. The respective class sizes are similar to those for the mixture GRM. The class preferring the extreme options over the midpoint is the smallest. This class also prefers the Strongly Agree option over all others, although not by a large amount. The class preferring the midpoint is larger, but not as large as the class preferring the Agree and Disagree options.

Figure 16: Anxiety (N1) Item Category Use for Three Class PCM mixture



Note: Class sizes are $N_1 = 2041$, $N_2 = 3711$, and $N_3 = 5655$. This represents 17.9%, 32.5%, and 49.6% of the sample.

4.1.1.2 Openness to Experience Feelings subscale(O3)

For the Openness to Experience Feelings scale, traditional analyses expected the following response styles: a medium sized group with low MRS and a small group with marked ERS. It was also expected that the correlation between the Openness to Experience Feelings and MRS traits is moderate and negative. The correlation between Openness to Experience Feelings and ERS was expected to be marked and positive.

Table 19 shows the mixture model selection criteria for the Openness to Experience Feelings (O3) facet. For this facet, a three class mixture graded model was suggested by the information criteria. As with the first scale, the BIC is lowest for the three class mixGRM solution $(196,041)$ when compared with BIC for the other models.

The entropy and MMCAP values were also comparable to those for the Anxiety scale. The values were higher for the two class solution than the three class solution and were higher for the mixPCM than the mixGRM. Thus, two and three class solutions were examined for the Openness to Experience Feelings scale.

The classification quality tables can be seen in Table 20. As the left side of the table indicates, for the two class solutions, there is good classification for both the extreme and non-extreme classes. The diagonal MCAP values are .84 or higher. For the two class mixPCM, yet there is higher classification quality than the two class mixGRM. The diagonal MCAPs

Table 19: Mixture Model Selection Criteria for Openness to Experience Feelings Facet

	Openness to Experience Feelings Facet			
K(Model)	BIC	sEn	MMCAP	ABSPR > 3
1(GRM)	198,870	—	1	36.0%
1(PCM)	203,391	—	1	48.1%
2(mixGRM)	196,413	.53	.84	10.1%
2(mixPCM)	198,035	**.67**	**.88**	23.0%
3(mixGRM)	**196,041**	.43	.68	**4.9%**
3(mixPCM)	197,002	.60	.79	12.0%

Note: K = Number of classes, BIC = Bayesian Information Criterion, sEn = scaled entropy, MMCAP = Minimum Diagonal value of Mean Class Assignment Probabilities table. ASBPR = Percent of Absolute Bivariate Standardized Pearson residuals greater than three. 2(mixGRM) = two class constrained graded response model. 3(mixPCM) = three class constrained partial credit model.

are lower for the three class solutions and this indicates that the classification quality is not as good as for the two class solutions. The values range from .68 to .73 for the 3(mixGRM) and from .79 to .87 for the 3(mixPCM). For the 3(mixGRM), there is more class overlap (higher classification uncertainty) between the General and Midpoint classes than for the Extreme and General classes. For the 3(mixPCM), the classification uncertainty is higher between the Extreme and General classes than the Midpoint and General classes.

In the fourth column of Table 19, the amount of absolute bivariate standardized Pearson residuals (ABSPR) > 3 for the models is presented. As can be seen, the PCM and GRM did not predict the joint distributions of items as well. The amount of large standardized bivariate residuals is less for the GRM than the PCM. In the cross-tabulation table, 36.0% of the cells had an absolute standardized residual ≥ 3 under the GRM which is lower than 48.1%, the amount under PCM. As with the Anxiety scale models, the Openness to Experience Feelings scale results show that the two class mixture PCM has a larger amount (23.0%) than the three class mixPCM (12.0%). Under the three class mixGRM, 4.9% of the cells had an absolute bivariate standardized residual > 3.

Table 20: Mean Class Assignment Probabilities Tables for the Openness to Experience Feelings scale

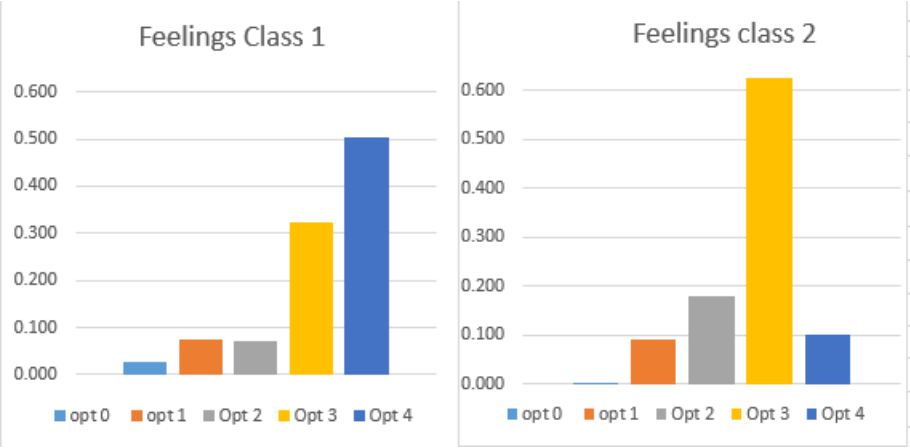
O3 2(mixGRM)			O3 2(mixPCM)			O3 3(mixGRM)				O3 3(mixPCM)			
	E	N		E	N		G	E	M		G	M	E
E	**.84**	.16	E	**.88**	.12	G	**.68**	.10	.22	G	**.79**	.09	.13
N	.13	**.87**	N	.08	**.92**	E	.13	**.79**	.08	M	.11	**.87**	.02
						M	.19	.08	**.73**	E	.17	.03	**.80**

Note: 2(mixGRM) = constrained two class graded response model. 3(mixPCM) = constrained three class partial credit model. E = Extreme class, N = Non-extreme class, M = Midpoint class, G = General class. Probabilities in bold indicate persons are classified with high probabilities in the respective class.

The distribution of categories used in 2mixGRM solution for the Openness to Experience Feelings scale is presented in Figure 17. In the figure, the first class uses more extreme options than the second class. The second class uses the midpoint more than the first class; however, the second class prefers the Agree option. Although not shown, the two class PCM mixture for the Openness to Experience Feelings scale indicated category use similar to the two class GRM mixture. The results can be seen in appendix I.

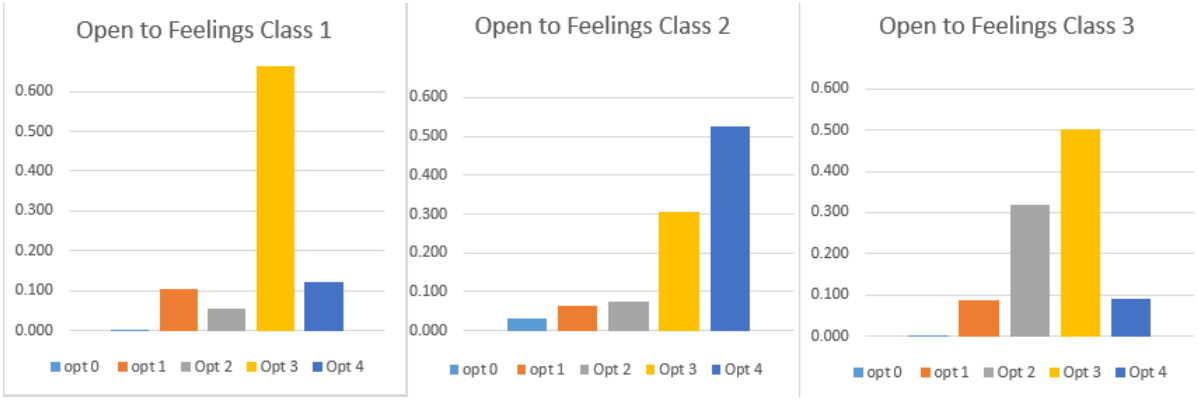
For the three class solutions, the Openness to Experience Feelings scale was different from the Anxiety scale in the use of categories and number of persons put in the three respective classes. Category use for the three class mixture GRM for the Openness to Experience Feelings (O3) items can be seen in Figure 18. As can be seen in the figure, the first class preferred the Agree option (option three) more than the other two classes. The second group tends to use the Strongly Agree option (option 4) more than the other classes. The third group used the midpoint (option two) more than the other classes; however, the Agree option was more frequently used than the midpoint option. The third class contains less persons ($N_3 = 3462$) than the first two classes ($N_1 = 4321$ and $N_2 = 3624$). The Openness to Experience Feelings scale indicated higher use of the strongly agree extreme category than the midpoint in the larger two classes.

Figure 17: Open to Experience Feelings (O3) Item Category Use for Two Class GRM mixture



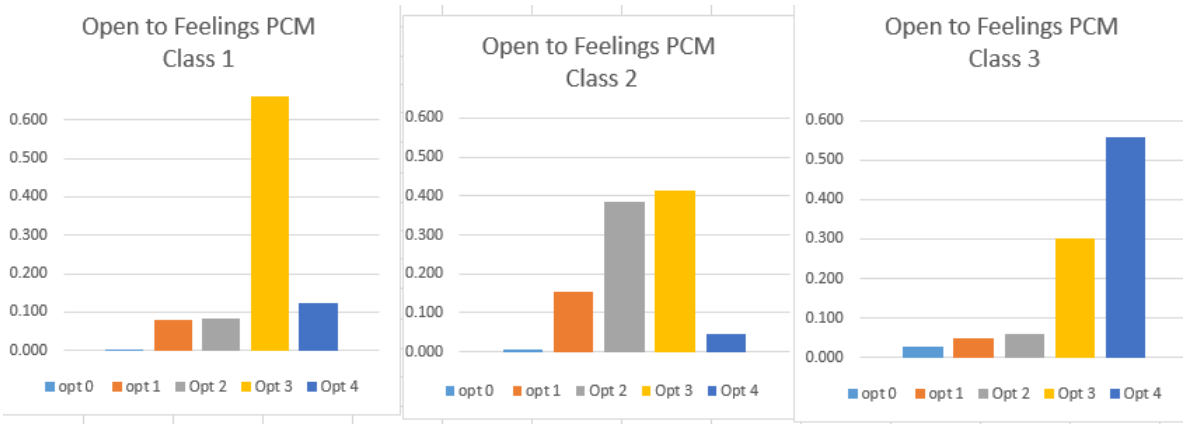
Note: Class sizes are $N_1 = 3947$ and $N_2 = 7460$. This represents 34.6% and 65.4% of the sample.

Figure 18: Openness to Experience Feelings (O3) Item Category Use for Three Class mixture GRM



Note: Sample sizes for the groups are $N_1 = 4321$, $N_2 = 3624$, and $N_3 = 3462$. This represents 37.9%, 31.8%, and 30.3% of the sample.

Figure 19: Openness to Experience Feelings (O3) Item Category Use for Three Class mixture PCM



Note: Sample sizes for the groups are $N_1 = 5428$, $N_2 = 2446$, and $N_3 = 3533$. This represents 47.6%, 21.4%, and 31.0% of the sample.

Category use for the three class mixture PCM for the Openness to Experience Feelings (O3) items can be seen in Figure 19. As can be seen in the figure, the first class preferred the Agree option (option three) more than the other two classes. The third class used the Strongly Agree option most frequently and also used the Agree option. The second class (also the smallest class) used the midpoint option more frequently than the other classes and also used the Agree option more than the midpoint.

More importantly, in the three class GRM and PCM mixtures for the Openness to Experience Feelings scale, none of the three classes preferred the midpoint. Each class preferred the Agree or Strongly Agree options over the other options. Thus, the Openness to Experience Feelings scale showed potentially low use of MRS.

4.1.1.3 Compliance

For the Compliance scale, traditional analyses expected the following response styles: a medium sized group with low MRS and a small group with moderate ERS. It was also expected that the correlation between the Compliance and MRS traits is negligible. The correlation between Openness to Experience Feelings and ERS was also expected to be negligible.

Table 21 shows the mixture model selection criteria for the Compliance (A4) facet. As with the first two scales, the information criteria (e.g., BIC, presented in the table) suggest a three class mixGRM solution while the classification quality measures suggest a two class mixPCM solution.

For the Compliance scale, the classification quality for the two and three class mixture solutions can be seen in Table 22. The classification quality of the two class mixGRM and mixPCM solutions are presented in first and second tables of Table 22. Since the Mean Class assignment Probabilities (MCAPs) for the two Class solutions are greater than or equal to .81, there is good classification quality for the two class mixGRM (2mixGRM) and mixPCM (2mixPCM) solutions, with the 2mixPCM having slighlty greater MCAPS than the 2mixGRM. In the three class solutions (in the third and fourth tables of Table 22), it is the extreme (E) class which has the largest diagonal MCAP of the three different classes for the two solutions. It is .77 for the 3mixGRM and .79 for the 3mixPCM. The diagonal MCAP is lower (.73 and .74 for the respective models) for the Midpoint class (M). The diagonal MCAP is even lower for the General (G) class. There is more classification uncertainty between the Midpoint and General classes compared to the classification uncertainty between the Extreme and General classes. Overall the three class classification quality for the Compliance scale is slightly better for the 3mixPCM than for the 3mixGRM.

The fourth column of Table 21 shows the amount of large absolute bivariate standardized Pearson residuals for the models. As with the first two scales, there are larger percentages for the standard models (28.4% and 35.1%) than the mixture models. There are fewer for the mixture GRM than the mixturePCM. The three class mixPCM has more (7.6%) than the three class mixGRM (1.1%).

Table 21: Mixture Model Selection Criteria for Compliance Facet

	Compliance Facet			
K (Model)	BIC	sEn	MMCAP	ABSPR > 3
1(GRM)	232,554	—	1	28.4%
1(PCM)	233,657	—	1	35.1%
2(mixGRM)	230,204	.53	.81	4.9%
2(mixPCM)	230,762	**.57**	**.82**	12.1%
3(mixGRM)	**229,888**	.40	.67	**1.1%**
3(mixPCM)	230,339	.41	.67	7.6%

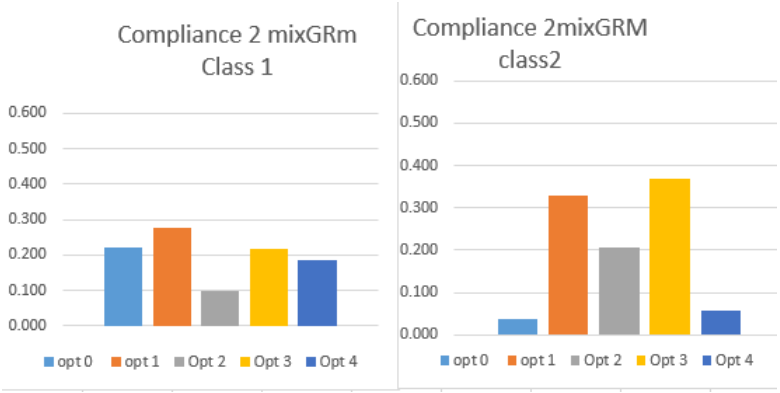
Note: K = Number of classes, ssBIC = sample size adjusted Bayesian Information Criterion, sEn = scaled entropy, MMCAP = Minimum Diagonal value of Mean Class Assignment Probabilities table. ASBPR = Percent of Absolute Bivariate Standardized Pearson Residuals that are large. 2(mixGRM) = two class constrained graded response model. 3(mixPCM) = three class constrained partial credit model.

Table 22: Mean Class Assignment Probability Tables for the Compliance scale

A4 2(mixGRM)			A4 2(mixPCM)			A4 3(mixGRM)				A4 3(mixPCM)			
	E	N		E	N		M	G	E		G	E	M
E	**.81**	.19	E	**.82**	.18	M	**.73**	.18	.09	G	**.68**	.10	.22
N	.12	**.88**	N	.11	**.89**	G	.23	**.67**	.11	E	.13	**.79**	.09
						E	.10	.13	**.77**	M	.19	.07	**.74**

Note: 2(mixGRM) = constrained two class graded response model. 3(mixPCM) = constrained three class partial credit model. E = Extreme class, N = Non-extreme class, M = Midpoint class, G = General class. Probabilities in bold bold indicate persons are classified with high probabilities in the respective class.

Figure 20: Compliance (A4) Item Category Use for Two class mixture GRM



Note: Sample sizes for the classes are $N_1 = 2969$ and $N_2 = 8438$. This represents 26.0% and 74.0% of the sample.

Category use by two different response style classes for the 2mixGRM for the Compliance (A4) items can be seen in Figure 20. The first class used the extreme options more than the second class. The first class also preferred extreme, Agree, and Diagree categories over the midpoint. The second class preferred non-extreme categories to extremes.

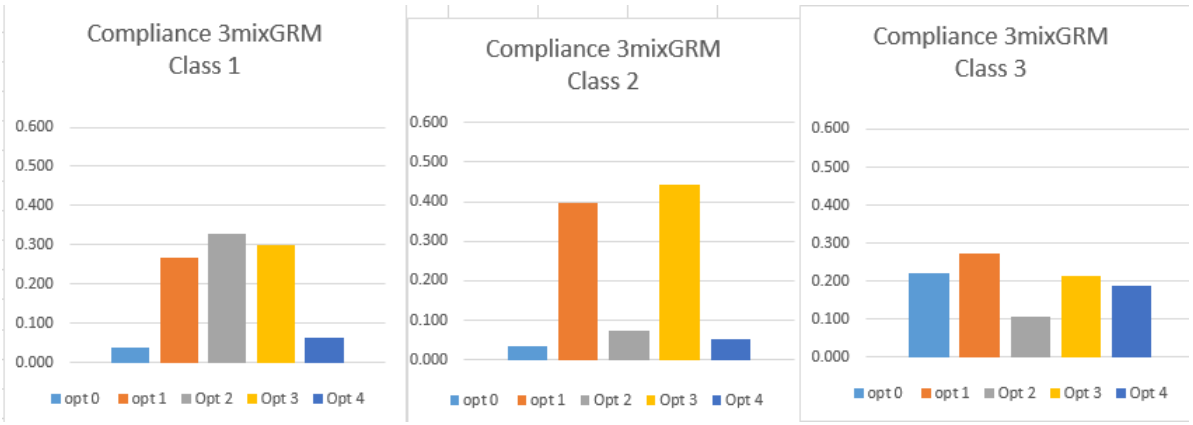
Category use by three different classes under the 3mixGRM for the Compliance (A4) items can be seen in Figure 21. The first class used the disagree option (one) more than the other two classes. The second class used the agree option more than the other classes. The third group used the midpoint more often than the other classes.

Category use by respondents under the 3mixPCM for the Compliance (A4) items can be seen in Figure 22. The first class tended to use the Agree or Disagree Categories most frequently. The second class used all categories; however the midpoint was used the least. The second class also had higher use of extremes than the other two classes. The third class preferred the midpoint over the other categories.

4.1.2 Summary of Mixture Model Results

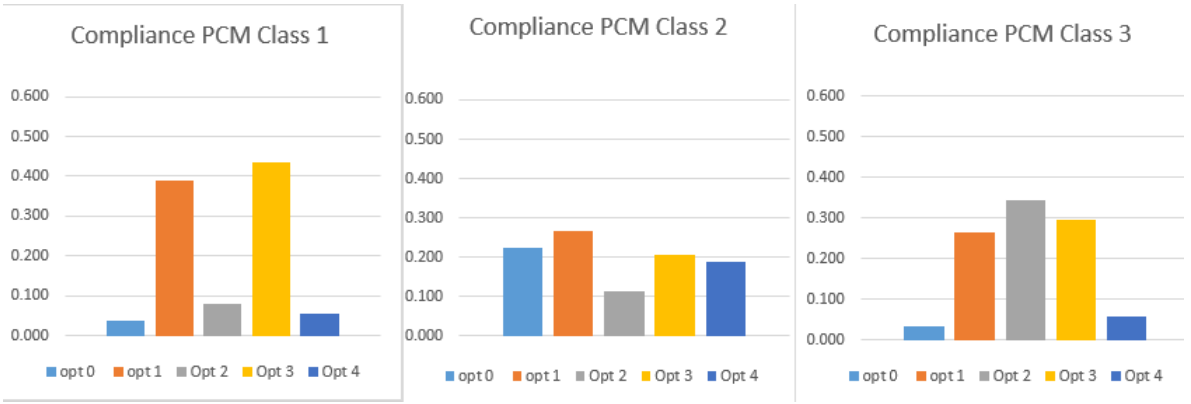
To summarize, the mixture GRM and mixture PCM show improved fit to the data over the standard IRT models. The classification quality statistics indicated that the two class

Figure 21: Compliance (A4) Item Category Use for Three class mixture GRM



Note: Sample sizes for the groups are $N_1 = 4327$, $N_2 = 4119$, and $N_3 = 2961$. This represents 37.9%, 36.1%, and 26.0% of the sample.

Figure 22: Compliance (A4) Item Category Use for Three class mixture PCM



Note: Sample sizes for the groups are $N_1 = 4521$, $N_2 = 2924$, and $N_3 = 3962$.

solutions had better classification quality than the three class solutions. For the Anxiety and Compliance Scales, the two class solutions had one class which preferred non-extreme options over extreme options and one class which preferred the extreme, Agree, and Disagree options over the midpoint. For the Openness to Experience Feelings scale, one class preferred the Strongly Agree and Agree options to the other options. A second class preferred the Agree and midpoint options to the other options.

Although the classification quality is not as high for the three class solution as it is for the two class solution, it is the three class solution for the Anxiety and Compliance scales which provides an estimate for a person to be in the class which prefers the midpoint over other categories. The three class solutions for the Anxiety and Compliance scales also provide an estimate for the person to be in a class which prefer the extreme, Agree, and Disagree options over the midpoint. With the Openness to Experience Feelings scale, in both two and three class solutions, there is always a class which prefers the Strongly Agree option and another that prefers the Agree option. In the three class solutions for this scale, none of the classes preferred the midpoint.

The mixture PCM has higher classification quality than the mixture GRM, yet the mixGRM fits better. The Model selection information criteria and the amount of large absolute bivariate standardized Pearson residuals (ABSPR) suggested using a three class solution over a one or two class solution for both the mixPCM and mixGRM. Additionally, the mixture GRM fits better than the mixture PCM since it had the least amount of ABSPR > 3 for all models for each scale. Thus, the three class mixture GRM solution is suggested for addressing ERS and MRS.

4.1.3 Multi-dimensional Model Results

The results of the three MIRT models are presented in this section. First, the findings from the Multi-dimensional Partial Credit Model analyses are presented.

4.1.3.1 Multi-dimensional Partial Credit Model Results

Table 23 shows the Bayesian information criteria (BIC) for the standard PCM and two and three dimensional

partial credit models. The two dimensional models involved the substantive trait and one of the response style traits (ERS or MRS). Recall from chapter three that the scoring function was used to define these latent response style traits. Each response style dimension had a common discrimination parameter. The three dimensional models involved all three traits.

The three dimensional model fit better than the two dimensional models for the Anxiety, Openness to Experience Feelings, and Compliance facets as indicated by the BIC. These findings are similar for the Anxiety facet for the models investigated by Wetzel and Carstensen (2015) who also studied two and three dimensional MPCM models. These researchers also found that three dimensional models fit better than two dimensional models for 25 of the other 29 facets; however, Wetzel and Carstensen (2015) did not specify for which particular four facets a two dimensional model fit better.

4.1.3.2 Multi-dimensional Nominal Response Model Results

Table 23 also shows the BIC for the multi-dimensional nominal response models examined for the Anxiety facets. The three dimensional models involved the substantive trait and both midpoint and extreme response style traits. The overall item category slopes for the substantive dimension were freely estimated in each of the models; however, the overall item category slopes for the MRS and ERS dimensions were constrained to be equal or freely estimated for the three dimensional models. For the two dimensional models involving one response style only (NRM-ERS, NRM-MRS), the overall item category slopes were freely estimated for both dimensions.

For three dimensional models, the BIC tended to be lowest for the MNRM model with freely estimated overall item items on all dimensions with all scales (FFF) except for a slight difference with the Anxiety scale. The FEF model with freely estimated overall item category slopes for the substantive and ERS dimensions and item slopes constrained to be equal for the MRS dimension had the lowest BIC. The table presents only the BIC for the FFF model. As can be seen in the table, all facets had the lowest BIC when the FFF model is compared with the standard and two dimensional models. For all three facets, the FFF model was selected to obtain the trait estimates since this model had the lowest information criteria.

4.1.3.3 Multi-Process Model Results

For the multi-process model analyses, recall that two process models measured level of agreement (direction) and either an indifference MRS or an intensity ERS trait level. Three process models measured trait levels for agreement, indifference, and intensity. Table 23 also shows the Bayesian Information criteria for the standard and multi-process models. The two process models are designated as Midpoint process (Mid-PM) and Extreme Process (Ext-PM). Multi-process Model 2 (M-PM2) is the three dimensional model for direction, indifference, and intensity processes. Both slope and difficulty parameters were estimated for all models. As can be seen from Table 23, the information criteria indicate that three process model fit better than the two process models. The M-PM2 fit also better than the standard models for the Anxiety, Openness to Experience Feelings, and Compliance scales.

4.1.4 Explained Variability in Responses

Related to the model information criteria is the amount of explained variability in responses. This is quantified using the general coefficient of determination R^2 (Nagelkerke, 1991) in this study. This R^2 enables researchers and practitioners to compare the usefulness of modeling additional dimensions with extra parameters. For the models examined, R^2 is not presented for the Multiprocess (M-PM) Models since the one and two process models are not nested within a three process model. The M-PM looks at different features of the data than the other MIRT and mixture models. The M-PMs use two-, three-, and four-point pseudoitems to model response processes. Dimensions are not added to a standard model as they are with the other models. For the M-PM, the R^2, the proportion of explained variation, would not be calculated.

Table 23: Bayesian Information Criteria and Explained Variability in Item Responses

		Anxiety		Openness to Experience Feelings		Compliance	
Model	p	BIC	R^2	BIC	R^2	BIC	R^2
PCM	33	232,563	—	203,391	—	233,657	—
PCM-ERS	35	229,328	0.248	198,226	0.382	230,867	0.218
PCM-MRS	35	231,028	0.127	202,203	0.119	232,866	0.069
MPCM	38	227,812	**0.343**	196,901	**0.436**	230,227	**0.263**
NRM	64	231,414	—	198,276	—	232,598	—
NRM-ERS	73	228,322	0.243	196,148	0.176	230,434	0.179
NRM-MRS	73	229,945	0.127	197,543	0.069	231,876	0.068
MNRM	83	**226,858**	0.340	**195,445**	0.232	**229,774**	0.231
Mid-PM	49	230,890	—	198,062	—	232,762	—
Ext-PM	41	229,876	—	197,108	—	232,027	—
M-PM2	51	228,813	—	196,414	—	231,723	—
PCM	33	232,563	—	203,391	—	233,657	—
2mixPCM	58	229,001	0.283	198,035	0.387	230,762	0.240
3mixPCM	83	228,065	0.353	197,002	0.452	230,339	0.282
GRM	40	230,152	—	198,526	—	232,260	—
2mixGRM	73	229,645	0.219	196,413	0.215	230,204	0.208
3mixGRM	106	227,003	0.281	196,041	0.261	229,888	0.250

Note: BIC = Bayesian Information Criterion, p = number of estimated parameters. R^2 = the general coefficient of determination (Nagelkerke, 1991). MRS = Midpoint Response Style, ERS = Extreme Response Style. PCM-ERS =Two dimensional Partial Credit model for trait and ERS dimensions, MPCM = multidimensional partial credit model for trait, MRS, and ERS dimensions. NRM-MRS = Two Dimensional NRM for substantive and MRS traits with freely estimated overall item category (FEOIC) slopes on both dimensions, NRM-ERS = Two Dimensional NRM for substantive and ERS traits with FEOIC on both dimensions, MNRM = Multidimensional Nominal Response Model with FEOIC slopes on substantive trait, MRS, and ERS dimensions. 2mixGRM = two class mixture GRM. 3mixPCM = three class mixture PCM. Mid PM = two process model measuring indifference (MRS) and agreement, Ext PM = 2 process model measuring intensity ERS and agreement. M-PM2 = Multi-Process model with varying item slopes for all binary pseudoitems for indifference, direction, and intensity. R^2 is not shown for the Multiprocess Models since one and two process models are not nested within a three process model.

For the models for the Anxiety (N1) scale, the amount of explained variability can be seen in Table 23. With the MNRM and MPCM, including an ERS dimension alone explains more of the variation in responses than including an MRS dimension alone. For example, the two dimensional models for the Anxiety scale show that PCM-ERS has $R^2 = .248$, while PCM-MRS has $R^2 = .127$. The two dimensional models for the MNRM show similar percentages (for NRM-ERS, $R^2 = .243$; NRM-MRS, $R^2 = .127$). ERS appears to be the more important response style than MRS.

For the Openness to Experience Feelings (O3) and Compliance (A4) scales, the amount of explained variability for the different models can be seen in Table 23. With these two scales, two dimensional models involving one response style also show that modeling ERS instead of MRS explains more of the variability. For example, using PCM-ERS instead of PCM-MRS, about 26.3% more of the response variability is explained for the Openness to Experience Feelings scale and about 14.9% more for the Compliance scale. Other research with the MPCM also found that ERS is the most important response style of ERS, MRS, acquiescence, and disacquiescence (Wetzel & Carstensen, 2015).

ERS is also seen to explain more response variation with the MNRM. For example with the Openness to Experience Feelings scale, the NRM-ERS $R^2 = .176$ while the NRM-MRS $R^2 = .069$. There are similar results for the Compliance scale ($R^2 = .179$ and 0.068, respectively).

With the mixture models, the ERS dimension is also seen to be the more important response style. Recall that with the mixture models, a two class mixture consisted of an extreme class and a non-extreme class. With the three class mixture models, the three classes were described as extreme, midpoint, and general, depending on relative category preferences. Thus, an extreme class always emerged.

However, adding the third class to the two class model does not increase R^2 as much as adding a second class to the standard (one class) IRT model. For example, with the Compliance scale (3mixPCM, $R^2 = .282$, 2mixPCM, $R^2 = .240$), adding the third class explains 4.2% more of the variability. This is larger than the 24.0% due to the two class model (i.e, separating an extreme from a non-extreme class). With the anxiety (N1) scale (3mixPCM, $R^2 = .353$, 2mixPCM, $R^2 = .283$), the third class explains 7.0% more of the variability. This is greater than the 28.3% with the two class (extreme-nonextreme) solution.

There is a similar result with the Openness to Experience Feelings scale. Therefore, with both mixture and MIRT models, using a third class or a third dimension to model MRS is useful in explaining additional variability in item responses but not as much as using the ERS dimension.

For the Openness to Experience Feelings scale, the two models with the largest amount of explained variability over the standard model are the three class mixPCM ($R^2 = .452$) and the Multidimensional PCM ($R^2 = .436$). The results for the Compliance scale are similar (3mixPCM, $R^2 = .282$, MPCM, $R^2 = .263$). The mixPCM explains about 2% more of the variability in responses than the MPCM. For the Anxiety scale, the 3mixPCM explains 1% more of the variation. Thus, the 3mixPCM and MPCM explain roughly the same about of response variation over the PCM. It takes more parameters with the 3mixPCM (83) than MPCM (38).

When the number of parameters in other models is examined, the MPCM has fewer parameters than the MRNM (38 vs. 83) yet explains more variance for the Openness to Experience Feelings scale ($R^2 = .436$ vs. .232) compared to the respective unidimensional model. The difference in variance explained is much smaller for the Compliance scale ($R^2 = .263$ vs. .231) and Anxiety scale ($R^2 = .343$ vs. .340).

Note also for Anxiety (3mixGRM, $R^2 = .281$, 2mixGRM, $R^2 = .219$), the third class explains 6.2% more of the variability than the two class which is similar to the increase from the 2mixPCM to 3mixPCM. Although the mixGRM fit better than the mixPCM for the respective number of classes, the mixPCM explains more of the variability than mixGRM over the respective standard model. A similar finding is true for the other facets.

4.1.5 Absolute and Relative Fit Results for Standard, Mixture, and Multidimensional Models

For the standard and multi-dimensional models, the absolute fit was assessed using the 95% confidence interval for the Root Mean Square Error of Approximation (RMSEA and the M_2 statistic (Cai & Hansen, 2013; Maydeu-Olivares, 2013; Maydeu-Olivares & Joe, 2006). The M_2 statistic is a limited information statistic based on the first and second moments. This statistic is recommended over the G^2 and Pearson's χ^2 statistics for situations when there are large and/or sparse contingency tables, as in this study. The M_2 statistic is approximately distributed as a χ^2 random variable.

All M_2 reflected a significant Chi-square statistic indicating that none of the models fit the data. But given the large sample size for this analysis and the fact that Chi-Square statistics are sensitive to sample size, this might be expected.

Table 24 shows the RMSEA for the PCM and GRM, M-PM, MPCM, and MNRM to assess absolute fit for these models. The RMSEA values indicate that all of the models **except** the three dimensional Multi-process model (M-PM2) and the two process model for direction and indifference tendencies (Mid-PM) fit the data. For example, the RMSEA values were greater than .10 (e.g.,.15 for the Anxiety scale) for the three dimensional Multi-process Model (M-PM) which indicates poor fit. The two dimensional Extreme-process models (models for Intensity and direction processes) did fit for all scales since the RMSEA values were less than .05.

The RMSEA for the three dimensional MNRM and MPCM were below .05 for all scales which suggests close approximate fit. For example, the MNRM for the Openness to Experience Feelings and Compliance scales had the lowest RMSEA value (.01). The two dimensional MNRM and MPCM also fit the data well for the scales. The two dimensional PCM for Anxiety and MRS did not fit as well as the two dimensional PCM for other traits and MRS. All 95% confidence intervals for the RMSEA values had a very narrow width due to the large sample size.

Absolute fit indices such as the M_2 and RMSEA statistics are not available for the mixture models in *Mplus*. Although Pearson χ^2 and Likelihood Ratio χ^2 values are given for mixture models in *Mplus*, the p values should not be trusted if there is a large discrepancy between the Pearson and Likelihood Ratio χ^2 values as this implied that at least one of the two statistics does not follow the theoretic χ^2 distribution (Geiser, 2013). For the scales in this study, there was a large difference in the Pearson and Likelihood Ratio χ^2 values; therefore the p values were not useful and not presented.

Recall that the amount of large absolute standard bivariate Pearson residuals was smallest for the three class mixture GRM compared with the three class mixture PCM and two class mixtures. This implies better fit of the three class mixture GRM.

Table 24 also presents the BIC for all standard, multidimensional, and mixture models so that relative fit can be assessed for all models. As can be seen from this table, the two models with the lowest BIC are the three class mixture GRM and the MNRM. The MPCM also has good fit. Therefore these three models are suggested as having the best fit in terms of the criteria presented here.

Table 24: Absolute and Relative Model Fit Criteria

	Anxiety		Openness to Experience Feelings		Compliance	
Model	BIC	RMSEA	BIC	RMSEA	BIC	RMSEA
PCM	232,563	.04	203,391	.05	233,657	.04
GRM	230,152	.03	198,870	.03	232,554	.031
NRM	231,414	.04	198,276	.04	232,598	.03
Ext-PM	229,876	.05	197,108	.04	232,027	.05
Mid-PM	230,890	.13	198,062	.19	232,762	.12
M-PM2	228,813	.15	196,414	.25	231,723	.10
PCM-ERS	229,328	.03	198,226	.03	230,867	.02
PCM-MRS	231,028	.06	202,203	.05	232,866	.04
MPCM	227,812	.03	196,901	.02	230,227	.02
NRM-ERS	228,322	.03	196,148	.02	230,434	.02
NRM-MRS	229,945	.03	197,543	.03	231,876	.03
MNRM	226,858	.02	195,445	.01	229,774	.01
2(mixGRM)	229,645	—	196,413	—	230,204	—
3(mixGRM)	227,003	—	196,041	—	229,888	—
2(mixPCM)	229,001	—	198,035	—	230,762	—
3(mixPCM)	228,065	—	197,002	—	230,339	—

Note: RMSEA = Root Mean Square Error Approximation, MNRM = Multidimensional Nominal Response Model with freely estimated overall item category slopes (FEOIC) on all three dimensions, NRM-ERS = Two dimensional Nominal Response Model for substantive and ERS traits with FEOIC slopes on both dimensions, NRM-MRS = Two Dimensional NRM for substantive and MRS traits with FEOIC slopes on both dimensions, MPCM = Multi-dimensional Partial Credit model for trait, ERS, and MRS dimensions, PCM-ERS =Two dimensional Partial Credit model for trait and ERS dimensions, M-PM2 = Multi-Process model with varying item slopes for all binary pseudoitems for indifference, direction, and intensity processes. Ext-PM = Two process model for Intensity and Direction, Mid-PM = Two process model for Indifference and Direction, Absolute fit statistics such as RMSEA are not available for the mixture models in *Mplus*. Bayesian Information criterion (BIC) is presented for all models for relative fit comparisons. 3(mixGRM) = three class mixture GRM.

4.1.6 Examining Correlations Between Trait estimates Within Scale Across Different Models

Another aspect to consider when determining which models are useful for modeling extreme and midpoint response styles is the correlation between substantive traits and between response style traits from the IRT models. The correlations between the model substantive trait estimates for the Anxiety (N1) scale are presented in Table 25. All of the correlations are greater than or equal to .914 when the PCM and GRM estimates are include. For the models which account for Response Styles, the correlations are greater than or equal to .959. These high correlations indicate that all of the models provide strong trait estimates while accounting for the hypothesized response style use.

The correlations between the model substantive trait estimates for the Openness To Experience Feelings (O3) and Compliance (A4) scales are presented in Table 25. All of the correlations are greater than or equal to .781 when the PCM and GRM estimates are included for the Openness to Experience Feelings scale. For the models which account for Response Styles, the correlations are greater than or equal to .862 for the Openness to Experience Feelings estimates. These high correlations indicate that all of the models provide marked to strong trait estimates while accounting for the hypothesized response style use in the Openness to Experience Feelings scale. The correlations between the trait estimates for the Compliance scale are even stronger. All of them are .938 or larger for the models which account for response styles.

The correlations between the model response style trait estimates are presented in Table 26. The correlations between ERS trait estimates are presented above the diagonal for the matrices in the table. The correlations between MRS trait estimates are below the diagonals for each matrix.

From the table, the correlations between model estimates for ERS were generally high for all model pairs and all scales. The minimum correlation between the extreme response estimates for two different models was r =.554 which occurred for the Openness to Experience

Table 25: Correlations between IRT Model Substantive Trait Estimates

	Anxiety (N1)						
	PCM	GRM	3mixPC	3mixGR	MPM	MPCM	MNRM
PCM	—	.986	.972	.982	.931	.974	.963
GRM		—	.946	.981	.914	.950	.956
3mixPC			—	.982	.971	.995	.982
3mixGR				—	.959	.983	.989
MPM					—	.979	.980
MPCM						—	.990
	Openness to Experience Feelings ╲ Compliance						
	PCM	GRM	3mixPC	3mixGR	MPM	MPCM	MNRM
PCM	—	.975	.988	.973	.933	.987	.975
GRM	.973	—	.948	.979	.926	.965	.975
3mixPC	.865	.781	—	.975	.938	.981	.970
3mixGR	.922	.933	.862	—	.935	.964	.979
MPM	.941	.922	.893	.951	—	.958	.960
MPCM	.910	.853	.947	.922	.960	—	.987
MNRM	.882	.877	.863	.969	.964	.947	—

Note: Correlations between model trait estimates for **Anxiety** are **above** the diagonal. Correlations between Model trait estimates for **Openness to Experience Feelings (O3)** are **below** the diagonal. Correlations between model trait estimates for **Compliance (A4)** are **above** the diagonal. 3mixPC =Three class mixture Partial credit model, 3mixGR = Three class mixture Graded Response model, MPM = Multi-Process Model, MPCM = Multidimensional Partial Credit Model, MNRM = Multidimensional Nominal Response Model.

Feelings scale between the MNRM and the multi-process model (MPM). It is only this correlation that is much lower than the others since the others are at least .800 for all model pairs for the three scales.

With respect to MRS trait estimates, the Anxiety and Compliance scales had high correlations for all model pairs. Most pairs for the Openness to Experience Feelings scale were also high since they were .809 or higher. However, two of the ten pairs were marked (e.g., r = .712 and .707); these two involved the mixture GRM with the mixPCM and M-PM. Thus, all of the model pairs tend to provide good estimates for MRS.

Finally, to examine how useful the two class mixture model substantive and response style trait estimates are, the correlations using two class mixture model estimates are given in appendix J. The size of the correlations indicate that the models are useful for obtaining substantive trait estimates and estimates for ERS. Since the correlations between MRS trait estimates are not strong, the two class models are not as useful as the three class models for estimating MRS.

4.1.7 Summary of Model Comparisons

To synthesize the results from the previous sections regarding the model comparisons, the following observations are made. The correlations between the model estimates for the substantive traits were high, so any of the two class mixture models or MIRT models could be used to account for extreme response style. The two dimensional multi-process model for Intensity ERS and direction also fit the data.

For all of the scales, the three dimensional M-PM did not fit as well as the three class mixture models. It also did not fit as well as the other MIRT models when the information criteria are compared and when the absolute fit statistics (the RMSEA) are examined in Table 24. For the Anxiety (N1), Openness to Experience Feelings (O3), and Compliance (A4) scales, the BIC was lowest for the MNRM with freely estimated overall item slopes for substantive trait and response style dimensions (FFF). After the MNRM, the 3mixGRM had the next lowest BIC and this is followed by the MPCM.

Table 26: Correlations between IRT Model Response Style Estimates

	Anxiety				
	3mixPC	3mixGR	MPM	MPCM	MNRM
3mixPC	—	.923	.870	.907	.905
3mixGR	.915	—	.812	.877	.876
MPM	.905	.809	—	.949	.930
MPCM	.905	.829	.930	—	.980
MNRM	.917	.844	.940	.992	—
	Openness to Experience Feelings				
	3mixPC	3mixGR	MPM	MPCM	MNRM
3mixPCM	—	.920	.878	.921	.925
3mixGRM	.712	—	.800	.904	.838
MPM	.850	.707	—	.932	.554
MPCM	.842	.809	.891	—	.795
MNRM	.854	.830	.912	.987	—
	Compliance				
	3mixPC	3mixGR	MPM	MPCM	MNRM
3mixPC	—	.979	.921	.899	.914
3mixGR	.981	—	.898	.897	.912
MPM	.906	.880	—	.950	.932
MPCM	.853	.856	.841	—	.951
MNRM	.904	.899	.903	.965	—

Note: Correlations between Model **Midpoint Response trait** estimates are **below** the diagonal. Correlations between model **Extreme Response trait** estimates are **above** the diagonal. 3mixPC = Three class mixture Partial credit model, 3mixGR = Three class mixture Graded Response model, MPM = Multi-Process Model, MPCM = Multidimensional Partial Credit Model, MNRM = Multidimensional Nominal Response Model.

Although the RMSEA and other absolute fit statistics are not available for the mixture models, the amount of large absolute standardized bivariate residuals could be found. The 3mixGRM had smaller amount than the 3mixPCM. Also, the three class models fit better than the two class models. The two class models had higher class assignment probabilities yet were not as useful for modeling midpoint response style as three class mixture models were.

In terms of the fit criteria examined, the three class mixture GRM is the preferred mixture model and the multi-dimensional partial credit model is the preferred MIRT model since these models account for both MRS and ERS. Although the MNRM fit slightly better than the MPCM, the MPCM has fewer estimated parameters than the MNRM (38 compared to 83). Also, the MPCM accounts for more variability in item responses than the MNRM over the respective standard IRT model.

4.2 EXAMINING RESPONSE STYLE USE FROM MODEL ESTIMATES

To demonstrate that the different response style models reflect groups or classes of similar respondents, groups of respondents were examined relative to their selection of midpoint and extreme options. This involved computing the mean use of MRS and ERS in each of the groups.

For the mixture models, the groups are formed from the classification probabilities (as part of the model output). For the three MIRT models, groups were formed by using the response style trait estimates and K-means clustering. The size of the midpoint and extreme groups differed by scale content.

To illustrate the relationships between substantive and response style traits, the estimates of latent correlations between different substantive and response style traits from the three multidimensional models are discussed. These correlations are not available for the mixture models which assume a discrete approach to response styles.

4.2.1 Examining Classes from Mixture Models

To examine response style use, first the sizes of the groups formed from the three class mixture models were examined. The mean proportions of midpoints and extreme options used by persons in the groups is used to characterize the groups using the language of correlation coefficients as with K-means clustering results in the previous chapter. The sizes of the groups formed by each mixture model were different depending on the scale content.

The different class sizes for the scales under the three class mixture PCM (3mixPCM) are presented in Table 27. For the Anxiety (N1) scale, there was a medium sized group (32.5%) of persons with moderate MRS. The persons in this class had a moderate mean proportion of midpoints used (M TMRS = .43). There was a small class (17.9%) which preferred the extreme options over the midpoint. These respondents had a moderate mean proportion of extremes used (M TERS =.45). Thus for the Anxiety scale, there tended to be low to moderate use of extreme response style (ERS). There tended to be medium use of midpoint response style (MRS) based on relative class size and mean proportions of midpoints used.

Of the three scales, the Openness to Experience Feelings (O3) scale had the largest extreme response class and the smallest midpoint response class for the mixture PCM (mixPCM). The Extreme size class was medium sized (31.0%) and was larger than the small Midpoint class (21.4%). Along with the low mean proportion of midpoints used in the Midpoint group (.39) and moderate mean proportion of extremes used in the Extreme group (.58), this implies that the Openness to Experience Feelings scale has medium level of ERS and low level of MRS.

The Compliance scale had two medium sized extreme and midpoint response style groups; however, the midpoint class is larger than the Extreme class (34.7% vs. 25.6%). There is a low mean proportion of midpoints used in the Midpoint class (.34) and a low mean proportion of extremes used in the Extreme group(.41). This implies that the compliance scale has low MRS use and low ERS use under the mixPCM.

Table 27: Mixture Model Class Sizes of Three Different Response Style Groups

	Three class mixture PCM		
Scale	Midpoint Size TMRS, TERS M(SEM)	Extreme Size TMRS, TERS M(SEM)	General Size TMRS, TERS M(SEM)
N1	32.5% **.43**(.002), .03(.001)	17.9% .11(.003), **.45**(.004)	49.6% .11(.001), .08(.001)
O3	21.4% **.39**(.003), .05(.002)	31.0% .06(.002), **.58**(.003)	47.6% .08(.001), .13(.002)
A4	34.7% **.34**(.002), .09(.002)	25.6% .11(.003), **.41**(.003)	39.6% .08(.001), .09(.001)
	Three class mixture GRM		
Scale	Midpoint Size TMRS, TERS M(SEM)	Extreme Size TMRS, TERS M(SEM)	General Size TMRS, TERS M(SEM)
N1	38.3% **.38**,(.003), .04(.001)	20.3% .12(.003), **.40**(.004)	41.5% .11(.002), .09(.002)
O3	30.3% **.32**(.003), .10(.002)	31.8% .07(.002), **.56**(.004)	37.9% .06(.001), .12(.002)
A4	37.9% **.33**(.002), .10(.002)	26.0% .11(.002), **.41**(.003)	36.1% .07(.001), .09(.001)

Note. The percentage of the sample (N = 11,407) assigned to the class designated as Midpoint (Midpoint preferred over Extremes, Extreme (Extremes preferred over Midpoint), or General (the Agree or Disagree options were preferred over the Midpoint and Extremes). TMRS = proportion of midpoints used used by persons in class, TERS = Mean proportion of extremes used used by persons in class, M = Mean of the proportion of midpoints(Extremes) used, SEM = Standard Error of the Mean. N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance.

For all of the scales, the General class which preferred agree and disagree categories over midpoints and extreme options was the largest of the three groups determined by the model. The General class always had negligible use of MRS and ERS as indicated by the mean proportions of midpoints and extremes. These mean proportions were all $\leq .11$.

The different class sizes for the scales under the three class mixture GRM (3mixGRM) are presented in Table 27. For the Anxiety (N1) scale, there was a small class which preferred the extreme options over midpoints (20.3%). This Extreme class had low use of ERS (M TERS = .40). There was a medium sized class which preferred the midpoint over extremes (38.3%). This class had low use of MRS (M TMRS = .38). Thus for the Anxiety scale, there tended to be low use of extreme response style (ERS) and low use of midpoint response style (MRS).

The Openness to Experience Feelings (O3) scale had a medium sized class of persons preferring the midpoint over extremes (30.%). This Midpoint response style class had low use of MRS (M TMRS = .32). The class which preferred extremes over midpoints was slightly larger than the Midpoint class (31.8%). This Extreme Group had a moderate use of ERS (M TERS = .56). These results suggest that the Openness to Experience Feelings scale has moderate use of ERS and low use of MRS.

The Compliance (A4) scale had a medium sized Midpoint response style class (37.9%) which had a low mean proportion of midpoints used (M TMRS = .33). The Extreme response style class was smaller (26.0%) and had a low mean proportion of extremes used (M TERS = .41). Thus, the Compliance scale tended to have low use of ERS and low use of MRS under the mixGRM.

In summary, the use of midpoint and extreme options for both the 3mixPCM and 3mix-GRM validated the interpretation of the groups. For example, the Midpoint group use midpoint options more than other groups and extreme options less than other groups.

4.2.2 Examining Groups from Multidimensional Model Estimates

To demonstrate that the response style estimates from the MIRT models also produce three different response style groups, K means clustering analyses were done using model based MRS and ERS estimates as the clustering variables. The different class sizes from K means clustering using the M-PM (K-MPM) response style trait estimates are presented in Table 28.

From Table 28, the Anxiety scale showed moderate use of indifference MRS (M TMRS = .42) in a medium sized K-MPM Midpoint group (31.4%). There was also moderate use of intensity ERS (M TERS = .42) in a small K-MPM Extreme group (21.4%). The Open to Experience Feelings scale had a medium sized K-MPM Midpoint group (28.6%) with low use of indifference MRS (M TMRS = .33). There was a medium sized K-MPM Extreme group (27.8%) with moderate use of intensity ERS (M TERS = .61). The Compliance scale had a medium sized K-MPM Midpoint group (30.5%) with low indifference MRS (M TMRS = .33) and a medium sized K-MPM Extreme group (22.5%) with moderate use of intensity ERS (M TERS = .43). The general groups for all three scales showed negligible indifference MRS and intensity ERS use with mean proportions of midpoints and extremes used which were less than or equal to .17.

The different class sizes from K means clustering using the MPCM response style trait estimates (K-MPCM) are presented in Table 28. The Anxiety scale had a medium sized K-MPCM Midpoint group (29.9%) with moderate MRS use (M TMRS = .44) and a small K-MPCM Extreme group (23.2%) with low use of ERS (M TERS = .40). The Open to Experience Feelings scale had a medium sized K-M Midpoint group (25.8%) with low use of MRS (M TMRS = .38) and a larger medium sized K-MPCM Extreme group (35.7%) with moderate use of ERS (M TERS = .56). The A4 scale had a medium sized K-MPCM Midpoint group (34.9%) with low use of MRS (M TMRS = .35) and a smaller medium sized K-MPCM Extreme group (31.5%) with low use of ERS (M TERS = .38). The general K-MPCM groups for all three scales showed negligible MRS and ERS use with mean proportions of midpoints and extremes used which were less than or equal to .13.

Table 28: K means groups from Multi-dimensional Model Response Style Trait Estimates

Scale	Midpoint Size M TMRS, TERS (SEM)	Extreme Size M TMRS, TERS (SEM)	General Size M TMRS, TERS (SEM)
	Multi-process Model Groups		
N1	31.4% **.42**(.003), .01(.001)	21.4% .07(.002), **.42(.004)**	47.2% .14(.002), .08(.001)
O3	28.6% **.33**(.003), .03(.001)	27.8% .03(.001), **.61(.003)**	43.6% .09(.001), .17(.002)
A4	30.5% **.33**(.002), .06(.001)	22.5% .10(.002), **.43(.003)**	46.9% .12(.002), .13(.001)
	Multi-dimensional Partial Credit Model Groups		
N1	29.9% **.44**(.002), .03(.001)	23.2% .09(.002), .**40**(.004)	46.9% .13(.001), .06(.001)
O3	25.8% **.38**(.003), .05(.002)	35.7% .05(.001), **.56**(.003)	38.5% .06(.001), .10(.001)
A4	34.9% **.35**(.002), .08(.001)	31.5% .11(.002),**.38**(.002)	33.7% .07(.001), .08(.001)
	Multi-dimensional Nominal Response Model Groups		
N1	29.7% **.44**(.002), .02(.001)	25.6% .10(.002),**.38**(.004)	44.7% .13(.001), .06(.001)
O3	24.4% **.39**(.003), .07(.002)	34.2% .06(.001),**.51**(.004)	42.8% .06(.001), .15(.002)
A4	29.3% **.37**(.002), .09(.002)	29.4% .10(.002), **.39**(.003)	41.2% .10(.001), .08(.001)

Note. Percentage of the sample (N = 11,407) assigned to group designated as Midpoint (Midpoint preferred to Extremes), Extreme (Extremes preferred to Midpoint), or General (Agree or Disagree options preferred). M TMRS = Mean proportion of midpoints used by persons in group, M TERS = Mean proportion of extremes used by persons in group. SEM = standard error of the mean, N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance.

The different class sizes from K means (K-MNRM) clustering using the MNRM response style trait estimates are presented in Table 28. The N1 scale had a medium sized K-MNRM Midpoint group (29.7%) with moderate MRS use (M TMRS = .44) and a medium sized K-MNRM Extreme group (25.6%) with low ERS use (M TERS = .38).

The Openness to Experience Feelings scale had a small K-MNRM Midpoint group (24.4%) with low MRS use and a larger medium sized Extreme group (34.2%) with moderate ERS use (M TERS = .51). The Compliance scale had a medium sized K-MNRM Midpoint (29.3%) group with low MRS use (M TMRS = .37) and a medium sized Extreme group (29.4%)with low ERS use (M TERS = .39). Though medium in size, the general K-MNRM groups were larger than the extreme and midpoint groups for all scales. They had negligible use of MRS and ERS with mean proportions of midpoints and extremes used which were all $\leq .15$.

The K means groups from the MPCM and MNRM showed a few differences from the M-PM K means groups. For the Anxiety scale the M-PM Extreme groups showed moderate use of intensity ERS while the MPCM and MNRM groups showed low use of ERS. For the Compliance scale, the M-PM Extreme group showed moderate use of intensity ERS while the other MIRT models showed low ERS use.

Finally, the mixture model results and the MIRT model with K means analyses results were used to form revised statements about groups and response style effects. The revised statements regarding possible response style effects are presented in Table 29. The differences from the response style effects originally described in the previous chapter (Table 16) are shown in bold font. Most of the statements indicate a change concerning what would be expected for ERS use. The Extreme group was small for mixture models and medium sized for MIRT models for the Anxiety scale instead of small for all models. The use of ERS for the Anxiety scale Extreme group varied from low (e.g., M TERS = .38) to moderate (e.g., M TERS = .45). The Openness to Experience Feelings scale had medium sized groups for all models instead of small groups. Also, the Openness to Experience Feelings scale Extreme group had Moderate use of ERS instead of Marked use of ERS. The Compliance scale had medium groups with low to moderate use of ERS instead of a small group with Moderate ERS.

Table 29: Revised Statements regarding Response Style Groups and Personality Traits

Scale	Statements about Groups and Response Style Effects
N1	Medium sized group w/ **Low to** Moderate MRS
	Small **to Medium** group w/ **Low** to Moderate ERS
O3	Medium size group w/ Low MRS
	Medium size group w/ **Moderate** ERS
A4	Medium size group w/ Low MRS
	Small to **Medium group w/ Low to Moderate ERS**

Note: The statements were made based upon examining the mean proportions of midpoints and extreme options used in each of the mixture model classes and in groups formed using the MIRT model response style estimates. MIRT model estimate were used with K means clustering analysis. N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance, MRS = Midpoint Response Style

Regarding MRS, the Anxiety scale had a medium sized Midpoint response group with moderate MRS use for all models except the mixture GRM whose Midpoint group showed low MRS use. This differs slightly from the expected Medium sized group with Moderate MRS use. For the Openness to Experience Feelings and Compliance scales, there were medium sized Midpoint groups with low MRS use which is what was expected from the traditional analyses in the previous chapter.

4.2.3 Multidimensional Model Estimated Latent Correlations between Facet and Response Style Traits

The MIRT models take a dimensional view of response styles. Because the MIRT models in this study assumed that the substantive and response style traits were correlated, it is also important to examine the sign and size of the latent correlations between these traits. The latent correlation is estimated as part of the model output. Traditional analyses expected negligible correlations between Anxiety and MRS and between Anxiety and ERS. There were similar expectations for the Compliance trait and MRS and ERS. For Openness to Experience

Table 30: Model Estimated Latent Correlations between Traits

	N1	O3	A4
M-PM2			
MRS (m)	**-.21**	**-.68**	**.27**
ERS (e)	.13	**.57**	**-.28**
MPCM			
MRS	-.13	**-.27**	.13
ERS	.03	.07	**-.30**
MNRM			
MRS	-.12	**-.30**	.09
ERS	-.11	**-.44**	-.05

Note: N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance, MRS (m) = indifference process Midpoint Response Style tendency, MRS = Midpoint Response Style trait ERS (e) = intensity process Extreme Response Style tendency. ERS = Extreme Response Style trait, M-PM2 = Three process model for indifference, direction, and intensity, MPCM = Multidimensional Partial Credit Model for ERS and MRS, MNRM = Multidimensional Nominal Response Model for ERS and MRS. **Nontrivial correlations** are given in bold.

Feelings, however, traditional analyses expected a moderate, negative correlations between the substantive and MRS traits. A marked, positive correlation between the Openness to Experience Feelings and ERS trait was also expected.

To examine relationships between substantive and response style traits, the model estimates for the latent correlations between the substantive trait and the midpoint and extreme response traits are not available for mixture models, yet they are examined for MIRT models. These model estimated latent correlations are presented in Table 30. The nontrivial correlations are given in bold font.

From the table, the negligible latent correlations for Anxiety (N1) for the MPCM and MNRM indicate no relationship between MRS and ERS use and Anxiety. There is also a negligible correlation between the Anxiety scale and the intensity ERS process for the multiprocess model (M-PM). There is a low, negative correlation for Anxiety with the indifference MRS process (r = -.21) for the M-PM. This implies that anxiety has no relationship with the intensity process (ERS) and that as anxiety increases, there is a tendency to use the indifference MRS process to a somewhat lower extent. The other MIRT models detected no relationship between anxiety and Extreme or Midpoint response style tendencies. The difference between the M-PM and other MIRT model correlations is attributed to the M-PM modeling response processes while the other MIRT models model response style tendencies.

There is further support that the M-PM differs from the other MIRT models in what it explains when the other scales are examined. For all three MIRT models, there are negative latent correlations between Openness to Experience Feelings (O3) and MRS yet the M-PM correlation is different. The correlations from the MPCM (r = -.27) and MNRM (r = -.30) are similar in size and low while the one for the M-PM (r = -.68) is marked and more than twice as large. Thus for the Openness to Experience Feelings scale, there is some support for a low, negative relationship with MRS . The correlations indicate that as Openness to Experience Feelings increases, the less likely persons are to use MRS. That the latent correlation from the M-PM is larger is interpreted as the M-PM modeling a different MRS trait (the indifference MRS process) from the other MIRT models. The MPCM and MNRM model MRS tendencies.

The latent correlations also indicate differences in the relationship between ERS and the Openness to Experience Feelings scale. There is moderate, positive correlation (r = .57) for the M-PM which implies a moderate relationship between use of the intensity process (ERS) and O3. For the MPCM, there is a negligible correlation (r = .07). For the MNRM, there is a low negative relationship (r = -.44) between Openness to Experience Feelings and ERS, which seems unusual, given the other two correlations. Thus, the models imply three distinct relationships between Openness to Experience Feelings and ERS. That the MNRM and M-PM have different correlations has been found in previous research comparing the MNRM with a Process Model for Extreme Response style (Leventhal, 2017; Thissen-Roe & Thissen, 2013). This indicates that the two models estimate different traits (an extreme response style tendency vs. an intensity ERS process).

For the Compliance (A4) scale, there are also some differences in what the three MIRT models indicated. The M-PM and MPCM revealed a low, negative relationship between Compliance and ERS. The two respective correlations were -.28 and -.30. This would imply that the more compliant person may be somewhat less likely to use the intensity ERS process and to have ERS tendencies since the relationship is not strong nor moderate. The MNRM showed a negligible correlation between Compliance and ERS (r = -.05) which suggests no relationship between Compliance and ERS tendency.

The MNRM also showed a negligible relationship between Compliance and MRS (r = .09) as did the MPCM (r = .13). The M-PM had a low correlation between indifference MRS and Compliance (r = .27). This would imply, at best, a low relationship between use of the indifference MRS process with the Compliance scale; however, the negligible correlations from the other models would imply no relationship between Compliance and MRS tendency.

The above results for the MPCM are comparable to those from Wetzel and Carstensen (2015). Two dimensional models involving the desired trait and a response style trait revealed small latent correlations between Compliance (A4) and ERS (r = -.31) and Openness to Experience Feelings (O3) and MRS (r = -.21). The two dimensional models showed very little correlation between Anxiety and ERS or MRS. The three dimensional models involving the substantive and ERS and MRS traits fit better than the two dimensional models and produced similar correlations.

To further explore the relationships between substantive and response style traits, the correlations between the substantive and response style trait estimates were examined. These are presented in Table 31. The nontrivial correlations are given in bold font. From the table, the Anxiety subscale had negligible correlations between substantive and ERS trait estimates for all models. Also, there were negligible correlations between anxiety and MRS estimates for the three class mixture models. There was a low, negative correlation between the anxiety and MRS indifference process trait estimates (r = -.283) and a low, negative correlation between the anxiety and MRS trait estimates for the MPCM (r = -.218) and for the MNRM (r = -.200).

For the Openness to Experience Feelings scale, there were low, positive correlations between the substantive and ERS trait estimates for all modes except the M-PM. The M-PM had a marked, positive correlations between substantive and ERS instensity trait estimates (r = .763). For the mixture models, there were low, negative correlations between the substantive and MRS trait estimates. There were marked, negative correlations between the substantive and MRS trait estimates for the MPCM and MNRM. For the M-PM, there was a high, negative correlation between the substantive and MRS indifference trait estimates (r = -.861).

For the Compliance scale, there were negligible correlations between the substantive and response style trait estimates for the two mixture models and the MNRM. For the M-PM, there was a moderate, positive correlation between indifference MRS and compliance (r = .434) and a negative , moderate correlation between intensity ERS and compliance (r = -.406). For the MPCM, there was a moderate, negative correlation between compliance and ERS trait estimates (r = -.407) and a low, positive correlation between compliance and MRS trait estimates (r = .218).

Finally, the above findings are summarized with revised statements concerning the relationships between the different personality traits and response styles. These are given in Table 32. The bold font distinguishes changes from the original statements made in the previous chapter (Table 16). Most of the changes are due to the M-PM. These statements are presented first for each scale and are followed by statements for the MIRT models. This

Table 31: Correlations between Substantive and Response Style Trait Estimates

	N1	O3	A4
3mixPCM			
MRS	-.078	**-.349**	.076
ERS	.016	**.240**	-.123
3mixGRM			
MRS	-.043	**-.292**	.074
ERS	.017	**.287**	-.117
M-PM2			
MRS (m)	**-.283**	**-.861**	**.434**
ERS (e)	.183	**.763**	**-.406**
MPCM			
MRS	**-.218**	**-.605**	**.218**
ERS	.061	**.351**	**-.407**
MNRM			
MRS	**-.200**	**-.615**	.168
ERS	-.111	**-.235**	-.138

Note: N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance, MRS (m) = indifference process tendency, MRS = Midpoint Response Style trait ERS (e) = intensity process tendency. ERS = Extreme Response Style trait, M-PM2 = Three process model for indifference, direction, and intensity, MPCM = Multidimensional Partial Credit Model for ERS and MRS, MNRM = Multidimensional Nominal Response Model for ERS and MRS. **Nontrivial correlations** are given in bold.

highlights the important difference between the noncompensatory (or partial compensatory) M-PM and the compensatory MPCM and MNRM. The M-PM estimates response process traits while the compensatory models estimate response style tendency traits.

Table 32: Statements regarding Relationships between Response Style and Personality Traits

Statements about Relationships with Response Style Effects	
Anxiety (N1)	
Low negative r between N1 and **indifference** MRS	negligible r between N1 and intensity ERS
negligible r between N1 and MRS	negligible r between N1 and ERS
Openness to Experience Feelings (O3)	
Marked negative r between O3 and **indifference** MRS	**Moderate** positive r between O3 and **intensity** ERS
Low negative r between O3 and MRS	**negligible to Low** r between O3 and ERS
Compliance (A4)	
Low positive r between A4 and **indifference** MRS	Low negative r between A4 and **intensity** ERS
negligible r between A4 and MRS	negligible to **Low negative** r between A4 and ERS

Note: The statements were made based upon examining the model estimated latent correlation between the response style and personality trait. The statements concerning indifference MRS and intensity ERS are for the Multi-process Model. The other statements are for the other MIRT models. N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance, MRS = Midpoint Response Style, ERS = Extreme Response Style. r = model estimated **latent** correlation.

5.0 DISCUSSION

In this chapter, a review of the study's purpose and methods and a summary of the major findings are provided. Some recommendations are given for selecting the best model to account for extreme (ERS) and midpoint (MRS) response styles. This is followed by a discussion of the limitations of the study. Finally, a discussion of future research is presented.

5.1 REVIEW OF THE STUDY'S PURPOSE AND METHODS

The use of instruments whose items contain an ordered response format is widespread in personality assessment. When response style use is suspected, the sum score should not be used to provide trait estimates for persons since it may be biased. Instead, estimates based on IRT models which account for response styles are suggested. Many studies have illustrated use of the Multi-dimensional Nominal Response Model (MNRM), the Multi-dimensional Partial Credit Model (MPCM), and the mixture partial credit (mixPCM) model to address use of response styles (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011; Falk & Cai, 2015; Rost, 1991; Wetzel & Carstensen, 2015; Wetzel, Carstensen, & Böhnke, 2013).

The purpose of this study is to contribute to the limited research with the Multi-process Model (M-PM) and the mixture Graded Response model(mixGRM) by comparing them with the other above models which account for Extreme and Midpoint Response style use. Three personality subscales from the German version of the *NEO-PI-R* personality instrument (Ostendorf & Angleitner, 2004) were used to illustrate how the five models account for response style use. The Anxiety (N1), Openness to Experience Feelings (O3), and Compliance (A4) subscales were chosen since each appeared to invoke use of MRS or ERS differently.

First Exploratory Factor analyses and correlational analyses using traditional response style measures (mean proportions of extremes and midpoints used) were conducted to demonstrate use of ERS and MRS in the selected scales. The Anxiety scale was proposed to exhibit moderate use of MRS and low use of ERS. The Openness to Experience Feelings scale appeared to exhibit low use of MRS and marked use of ERS. The Compliance scale appeared to exhibit low use of MRS and low use of ERS.

K-means clustering results based on traditional response style measures were used to show that three groups of persons in the sample generated the item responses to the scales. For each scale, the groups differed in size and level of response style use as indicated by the mean proportions of midpoints and extremes used. The data were then analyzed with the standard IRT and five IRT models which account for response styles.

The models were compared using fit criteria and statistics available from software output, response style group characteristics, and correlations between model estimates for substantive and response style traits. The estimated latent correlations between substantive and response style traits for the MIRT models were examined for size and sign. When possible, the increase in explained variability of a complex model over a standard IRT model was also examined.

5.2 SUMMARY OF MAJOR FINDINGS

Research Question A asks if modeling response styles with mixture and multidimensional models improves model-data fit over standard partial credit (PC) and graded response (GR) models. Three research questions support the answer to **Question A** since each of the five models which account for response styles was compared to a standard IRT model.

The IRT models viewed the response style variables as either discrete or continuous. With a discrete approach, the mixture model categorizes each person as having a particular response style or not based on the class assignment probabilities. With the continuous view, the MIRT model provides each person with a response style trait estimate for each dimension of the MIRT model. The results of these two approaches are discussed in turn and summarize answers to research questions A1, A2, A3, and B1.

5.2.1 Summary of Mixture Model Findings

Research question A1 asked if the mixture models fit better than the standard IRT models and how the two mixture models compared with each other. For the mixture and standard models, the amount of absolute standardized bivariate Pearson residuals (ASBPR) greater than three could be determined using software output. Models with less amounts of ASBPR fit the data better than those with larger amounts.

The information criteria and the small amount of large ASBPR indicated that the mixture models fit better than the standard models. The fit statistics also indicated that the mixture graded response model (mixGRM) fit better than the mixture partial credit model (mixPCM) for all three scales. Additionally, these measures also indicated that for both mixGRM and mixPCM, the three class models (3mixGRM, 3mixPCM) fit better than the corresponding two class models. Thus, three different response style groups were identified.

The entropy value and Mean Latent Class Probabilities for Most Likely Latent Class Membership (Class Assignment) indicated that the two class model would provide better classification than the three class PCM and GRM mixtures; however, the three class model explained more variation in item responses than a two class mixture model. It was also the three class mixtures which provided better MRS trait estimates than two class mixtures when correlations with the MIRT model estimates were examined. Additionally, the three class mixture PCM had better classification quality than the three class mixture GRM. The three class mixture GRM fit better than the three class mixture PCM.

The mixture models identified medium sized Midpoint response style groups which had low to moderate use of MRS for the Openness to Experience Feelings and Compliance scales and moderate MRS use for the Anxiety scale. For the three class mixPCM and mixGRM, the mean proportion of midpoints used in the Midpoint class was slightly larger for the Anxiety scale than the mean proportion of midpoints used in the Midpoint classes the other scales. This indicates that there may be slightly more use of MRS with the Anxiety scale than with the Openness to Experience Feelings and Compliance scales.

The mixture models identified small Extreme response groups for the Anxiety scale. The use of ERS in these groups was low to moderate. The mixture models identified medium sized Extreme response style groups for the Openness to Experience Feelings and Compliance scales. There was moderate use of ERS for the Openness to Experience Feelings scale and low use of ERS for the Compliance scale. Thus, there may be more use of ERS with the Openness to Experience Feelings scale than with the Anxiety and Compliance scales.

The mixture models tended to have larger Midpoint and Extreme classes than the groups formed from K-means clusters. Recall that the mixture models used item response patterns while the K-means clustering algorithm used traditional response style measures (mean proportions of extremes or midpoints used). With the traditional measures, the substantive and response style trait effects are not separated and are therefore biased.

The general classes from the mixture models were smaller than the corresponding general groups from the K-means algorithm. This implied that the mixture models assigned more persons to the midpoint and extreme groups than the K-means clustering algorithm.

The midpoint, extreme, and general response style groups from the mixture models had different sizes for different scales. There were also different levels of response style use as indicated by the mean proportion of midpoints and extremes used. This is interpreted as use of "scale specific" response styles across the different scales. It does not indicate anything about general response style use without further analyses.

5.2.2 Summary of MIRT Model Findings

Research question A2 asks if the M-PM fit better than standard IRT models. The information criteria indicated that the relative fit was better for the M-PM for all three scales. For the absolute fit though, the RMSEA indicated that the three process M-PM did not fit the data well for any of the scales. The two process model for indifference and direction also did not fit the data well. However, the two process model for intensity and direction did fit the data well. This better fit of the two process model for ERS over standard models is similar to findings with other scales (Thissen-Roe & Thissen, 2013).

Research question A3 asked if the MPCM and MNRM fit better than the standard IRT models. The information criteria indicated that these three dimensional models for trait, ERS, and MRS fit better than standard IRT models. Additionally, the RMSEA indicated that the MNRM and MPCM had close approximate fit to the data while the M-PM did not.

The information criteria for two dimensional models for ERS indicated that the MNRM-ERS fit better than the MPCM-ERS and Ext-PM (two process model). In studying models accounting for ERS, Leventhal (2017) also found the MNRM was preferred to the IRTtree model for ERS (a two-process model) as it often had better fit in empirical and simulation studies. Leventhal (2017) assumed that the substantive and ERS traits were independent and called for research in which these traits were correlated. The MIRT models in this study contribute to such research.

Research question B1 asked how the model estimated latent correlations between substantive and response style traits compare with each other for the MIRT models. There were some scale-dependent differences for the substantive trait to ERS trait latent correlations. For all three of the three dimensional MIRT models, the trait to ERS latent correlations were negligible for the Anxiety scale. The Openness to Experience Feelings to intensity ERS latent correlation was moderate for the M-PM while the analogous correlation between Openness to Experience Feelings and ERS traits was negligible for the MPCM and was negative and moderate for the MNRM. With the Compliance scale, for the model estimated trait-ERS latent correlations, there was a different pattern from the first two scales. The M-PM and MPCM had low negative correlations, while the MNRM had a negligible latent correlation.

The model estimated latent correlations between the substantive and MRS traits were similar in size for the MPCM and MNRM and these differed from the analogous latent correlations for the M-PM between indifference MRS and substantive traits. For the Anxiety scale, the trait to MRS latent correlation was negligible for the MNRM and for the MPCM while there was a low negative correlation between Anxiety and indifference MRS for the M-PM. For the Openness to Experience Feelings scale, there was also a low, negative trait to MRS latent correlation for the MNRM and for the MPCM while there was a marked negative latent correlation between Openness to Experience Feelings and indifference MRS for the M-PM. For the Compliance scale, the correlation between trait and MRS were negligible for

the MNRM and MPCM while there was a low positive correlation between Compliance and indifference MRS for the M-PM. The difference in the size of the correlations is attributed to the compensatory MIRT models (MPCM and MNRM) providing response style tendency trait estimates for ERS and MRS while the noncompensatory M-PM provides response process trait estimates for indifference MRS, intensity ERS, and direction (agreement or not). Previous research with IRT tree models for ERS like the M-PM has also indicated that these models estimate a unique construct-irrelevant variance factor (i.e., a different ERS trait) from the ERS trait of the MNRM (Leventhal, 2017; Thissen-Roe & Thissen, 2013).

In the current study, the three dimensional MIRT model response style trait estimates were also used to form three different response style groups using K-means clustering. The sizes of these groups differed across the three scales as the mixture model classes did. This provides additional support that each personality trait invoked different use of "scale-specific" ERS and MRS.

5.2.3 Findings Comparing Mixture and Multdimensional Models

General Research question B2 asked how the correlations between trait estimates from the different models compare. The correlations between substantive trait estimates for all models which accounted for responses styles were high for all scales. The lowest was the correlation between the three class mixGRM and three class mixPCM Openness to Experience Feelings scale (.862). This indicates that any of the models could be used to provide a substantive trait estimate that has been corrected for possible MRS and ERS use.

When the correlations between estimates for the response style traits are examined, all of the model pairs tended to have high correlations for ERS tendencies (.800 or larger) except for the moderate correlation for MRS estimates between the MRNM and the M-PM for the the Openness to Experience Feelings scale (.554). Thus, all of the models could be used for measuring ERS tendency in persons.

The correlations between MRS tendency estimates were also high for all scales. However, with the Openness to Experience Feelings scale, there were two pairs which were marked (e.g., .707). These were between the three class mixGRM and class mixPCM and between the three class mixGRM and M-PM. Thus, the models could be used to provide MRS tendency measures for persons.

Research question A4 asked how the models compared in terms of explained variability. The M-PM could not be compared with other models for this question since the unidimensional models are not nested within the M-PM. Examining two dimensional compensatory MIRT model results, the MPCM and MNRM with one response style and one substantive trait indicated that modeling the ERS dimension explained more of the response variation than modeling the MRS dimension. This may indicate more impact of ERS than MRS in this dataset. This finding agrees with the results of Wetzel and Carstensen (2015) who also studied the MPCM. This finding also provides support to research which suggests that ERS is more important than MRS in explaining item responses.

The general coefficient of determination R^2 was larger for the three class models and three dimensional models than it was for two class and two dimensional models. Thus, the three class or three dimensional models would be preferred based on these results. To summarize, the three dimensional MNRM and three class mixture GRM tend to fit better than the other models. However, it is the three dimensional MPCM and three class PCM which explain more item response variability.

5.3 RECOMMENDATIONS FOR CHOOSING A MODEL

In this study, the purpose was to compare how five different models account for extreme and midpoint response styles in a real dataset. The five models account for extreme and midpoint response styles yet do so in different ways. With constant item discrimination parameters, the mixture PCM, the MPCM with common item discrimination parameters, and the MNRM with fixed order of the categories, assume the substantive trait affects item responses in the same way, while the M-PM and mixture GRM assume the substantive trait

affects item responses differentially. The mixture models take an exploratory approach since the response styles are not specified before the analyses (Böckenholt & Meiser, 2017). The characteristics of the classes, such as mean proportions of midpoints and extreme options used, must be examined carefully to determine what type of response styles exist in each class (e.g, Acquiescence, ERS, etc.). By specifying midpoint and extreme response dimensions, the MIRT models in this study can be seen as taking a confirmatory approach. The MIRT models test whether response-style processes or response style tendencies are plausible explanations of the observed data.

General Research question C asked which model is the best for addressing response styles. When psychometricians choose a model, there are many criteria to consider. These include the nature of the response styles modeled, assumptions about the attitudinal judgment process, and pragmatic concerns with implementation and estimation (Böckenholt & Meiser, 2017).

With the first criterion, the practitioner or researcher chooses to view response styles as discrete or continuous latent variables. When a discrete view is chosen, the mixture models are implemented and the models for different numbers of classes must be estimated. Besides information and other fit criteria, the classification quality must be examined. The classification quality is important since the trait estimates are adjusted based on class assignment. Wetzel, Böhnke, and Rose (2016) caution that if the person is assigned to the wrong class, the substantive trait may be adjusted incorrectly.

For the mixture models in this study, the class assignment probabilities were higher for the two class models than for the three class models. The two class models accounted for ERS and provided good estimates for ERS traits, yet did not provide good trait estimates for MRS. The three class models, however, provided better trait estimates for MRS than the two class models (See Table 26 in Chapter 4 and Table 34 in Appendix J). Thus, to obtain an estimate to be in the MRS class, the practitioner or researcher must accept a classification quality that is not as good as the two class model. This could possibly compromise the substantive trait estimate.

If a continuous view of response styles is chosen, then a MIRT model must be chosen. The choice of a MIRT model in this study involves choosing a model which specifies response style tendencies that compensate for the effects of the substantive trait (MNRM or MPCM) or a model which specifies noncompensatory response process traits which are part of a sequential judgment process (M-PM). These response process traits are different from response style tendency traits as indicated by the differences in the model estimated latent correlations between substantive and response style traits presented in Table 30 in Chapter 4.

A second factor that researchers may consider in choice of model concerns the latent judgment process. The MPCM, MNRM, and mixture models in this study assume an ordinal assessment where the response options indicate gradual degrees of the substantive latent trait (Böckenholt & Meiser, 2017; Rost, 1991). The two and three dimensional Multi-Process Models, however, assume that agreement or disagreement with item content can be decomposed into a sequence of decision processes. The processes may be three binary processes (indifference, direction, and intensity) or a combination of a binary process (e.g., indifference) and another process to explain the item responses. The Multi-process models suggest the polytomous item response is the result of the respondent's answers to a series of mental queries about ambivalence toward item content and intensity of an attitudinal position (Böckenholt, 2017).

Although, the three process M-PM did not fit the data well as other models in this study did, the two process model for intensity ERS and direction did fit the data well. Only the two process model can be recommended for the scales in this study. If psychometricians are interested in modeling judgment as a sequence of decisions based on response process traits such as indifference MRS or intensity ERS, then the M-PM should be considered since it is the only model that provides such estimates.

If researchers believe modeling response style tendencies instead of response style processes is a better reflection of the latent judgement process, then the M-PM is not considered further. If a researcher or practitioner wants to account for ERS, then the two class mixture PCM and two dimensional MPCM-ERS are suggested, since these models explain more of the response variability than the two dimensional MNRM. The two class PCM has better classification quality than the two class GRM. If the researcher's goal is to account for both

MRS and ERS, then the MPCM is suggested over the mixture models and other MIRT models since it has only slightly larger BIC than that of the MNRM. Both the MNRM and MPCM fit the data well according for the RMSEA, however the MPCM has the larger values of explained variability over the unidimensional models (R^2) for each subscale compared to the R^2 values for the MNRM. Additionally, the MPCM has a smaller number of estimated parameters than the MNRM (38 vs. 83).

A third factor that psychometricians may consider is what is required to implement and estimate a model. A potential drawback of the mixture models is the time consuming estimation that their implementation requires (Böckenholt & Meiser, 2017). McCrea (2013) suggests that researchers also need time to examine response profiles of persons in the classes to interpret the classes clearly and realistically. The MIRT models, on the other hand, have straightforward implementation which usually requires less estimation time than the mixture models. Psychometricians using MIRT models like the MPCM and MRNM do not have to concern themselves with the classification quality as they would when using a mixture model. Thus, for the scales and models examined in this study, the MPCM is suggested overall, since this model appears to fit the data well, it provides both ERS and MRS trait estimates, and the implementation is straightforward. The MPCM also explains more item response variability over the standard model than the MNRM.

5.4 LIMITATIONS

As with any research project, the research presented here has some limitations. The first limitation is the study uses an available real data set and the presence of response styles is inferred from the results of the analyses. In using real data, the item response generating mechanism is unknown. It is possible that some respondents who use high (low) extreme categories may have truly high (low) levels of the trait. Respondents who select the midpoint may truly have a medium trait level. Therefore, the recommendations in section 5.3 are possibly limited to the real dataset that was used.

A second limitation is that some respondents may select categories due to insufficient effort responding or social desirability responding (SDR). These types of responding cannot be examined for the subjects since the data collection did not involve the necessary items to measure the degree of such processes. Methods using an infrequency scale or social desirability scale exist to examine these types of responding (Fischer & Fick, 1993; Huang, Bowling, Liu, & Li, 2015). These methods require using additional items and are beyond the scope of this study since those scales were not included in the questionnaire.

This second limitation may not really be a main concern for the following reason. Although some professionals may argue that assessing for social desirability or inconsistent, random responding is important, McCrae and Costa (2010) argue that these scales may not be necessary for the following reasons: (1) In a clinical or volunteer context, most respondents do not bias their responses. (2) Scales designed for SDR may hinder accurate assessment and do not work. (3) What appears to be inconsistent responding to researchers is not useful in detecting actual random responding.

The third limitation is that a small number of facets is examined. There are three practical reasons for this. The first reason is previous research revealed a limited number of facets (16) for which the constrained mixture PCM (mixPCM) fit better than the unconstrained mixPCM (Wetzel et al., 2013). Use of the constrained mixture model is desired since with this model, the same trait is measured in each class and class differences in item responses are attributed to response scale use only. When the unconstrained model fits better, class differences exist due to differences in the measured trait in each class and in use of response options. The second reason is that previous findings revealed that many of the personality trait facets did not exhibit large correlations with extreme nor midpoint response style (Wetzel & Carstensen, 2015). Only five of the 30 lower order facet subscales showed non-negligible correlations with ERS or MRS. Of these five, only two (Openness to Experience Feelings and Compliance) were facets in which a constrained mixture PCM fit better than an unconstrained mixPCM or there were convergence/estimation issues. Thirdly: the goal of the study is to use a small number of facets to illustrate how the mixture graded response

model and multi-process model address extreme and midpoint response style use compared to the mixPCM and MPCM. Analyzing other facets is speculated to show similar results and not provide anything very different from these analyses.

An important fourth limitation is that there are a relatively small number of items in each scale (eight). The Multidimensional Nominal Response Model (MNRM) for response styles can be sensitive to any anomalies in the data. Previous work with this model involved items with at least 10 items per scale (Bolt & Newton, 2011; Falk & Cai, 2015). There is no guarantee the solutions are fully converged and stable. To check that the solutions are fully converged and stable would require changing the start values and estimating the model again. Unfortunately, there is no direct mechanism to do this in *flexMIRT* with MIRT models as there is in *Mplus* with the mixture models. It would have to be done manually which requires more software programming and estimation time. The possibility of non-convergent or unstable solutions from the MNRM could affect the interpretation of the results.

Related to the items is the number of response options. Items with seven response options may be able to isolate a midpoint response tendency more than five point items. The present study found more impact due to extreme response than midpoint response tendencies. This may be due to the relatively small number of options in the scale.

The sixth limitation of the study is that only two response styles were examined and there are others. But note that the multi-process model is designed only to account for midpoint and extreme response styles. Acquiescence and disacquiescence response styles are examined with other structural equation models that could involve method factors or random intercepts for the positively and negatively worded questionnaire items (Plieninger & Meiser, 2014) or other multi-dimensional models with trait and response style factors (e.g., Wetzel & Carstensen, 2015).

A seventh limitation concerns the use of MIRT model estimates with K means clustering to form groups. The response style trait estimates have error which is not incorporated into the analyses. The K means algorithm is not sensitive to varying amounts of error in the point estimates that are used. Thus, this could affect how persons are assigned to the response style groups.

An eighth limitation may be that the size of the groups exhibiting the response tendencies (MRS, ERS) were small to relatively small compared to the general or non-response tendency group. Reducing the size of this general group by randomly deleting subjects may increase the impact of the response tendency groups and thereby increase the ability of the models to capture response tendency.

Lastly, the study is also limited in that the IRT models examined do not address a general measure of ERS or MRS tendency. The estimates for ERS or MRS in this study are useful only for capturing scale-specific response styles since the models use homogeneous items from each scale. With only three scales examined, general response style tendencies cannot be described without additional analyses. Examining additional scales and performing additional analyses (e.g., a second order latent class analysis) could reveal more about general response style tendencies.

Another way to measure and control for general extreme or midpoint response style tendencies for each person involves use of more complicated models using heterogeneous items from other scales. Such models are not part of this study since the trait estimates from those models would be different due to the general response style tendencies modeled. Such trait estimates would not be comparable to estimates from the models which adjust for "scale specific" response style tendencies in this study (Wetzel & Carstensen, 2015).

5.5 FUTURE RESEARCH

To continue the research begun in this study, there are several possibilities. The model output seemed reasonable in term of trait estimates, parameter estimates, and standard errors. Support regarding *flexMIRT* software raised the possibility that some of the solutions for the MNRM might not be stable, nor convergent (Li Cai, personal communication, July 27, 2017). Due to this possible limitation, the data could be re-analyzed with different starting values to check the convergence and stability of the solutions. Future research with

MIRT models could involve manually changing the random starting values and re-estimating the model. The likelihoods from the models can be examined to determine the maximum one to have some assurance that a fully converged and stable solution has been found.

A second possibility is to examine how adding covariates to the models helps to explain more variability in the responses. In this study, the largest practical difference between number of extremes and midpoints used occurred between males and females for the Openness to Experience Feelings scale (See Table 8 and Table 9). Thus, including gender in the mixture, multi-process, and other MIRT models may help to provide further insight into response style use.

A third possibility is to randomly split the data into two parts. McCrea (2013) suggests that when the models include a large number of parameters and the modeling approach is purposely exploratory, there is a risk of overfitting the models to the data. Estimating the models on each randomly chosen half of the data provides a check for whether the model replicates or not. If the model replicates, then overfitting has probably not occurred.

The fourth possibility involves models to examine or control for general response style tendencies. For example, with the continuous view of response styles, Falk and Cai (2015) used the MNRM to extend the work of Bolt and Newton (2011) so that the models included more than one response style and more than two substantive scales. The models they studied incorporated six substantive and two response style traits. Wetzel and Carstensen (2015) studied two and three dimensional models for one substantive trait and two response style traits using homogeneous and additional heterogeneous items with the MPCM. Additional research with the MPCM with more than one substantive trait may be possible.

Another possibility for extensions to this study is to examine a complex model involving more facets from the same content domain. It would be interesting to determine how the MNRM performs with traits which are expected to correlate since they measure aspects of personality from the same domain, yet different constructs. For example, all 48 items from the six scales of the Conscientiousness or other personality domain could be examined for response style use.

Scales from the same domain would be expected to correlate. The correlations between sum score trait estimates for such scales could be compared with correlations between IRT model trait estimates to determine if the correlations from models which account for response styles are lower. This would imply that the model has adjusted the trait estimates for response style use.

Another way to extend this study is to use the mixture model output to examine consistency of general response style use. The class assignment variables which indicate a person's class for each subscale could be used in a second order latent class analysis to see how many respondents are consistent in their use of response styles across different trait scales.

Lastly, Wetzel et al. (2016) conducted a simulation study with mixture and multidimensional models accounting for ERS use in scales with four-point items. A simulation study involving five point scale items could be conducted to examine models for both ERS and MRS.

To summarize, the current study contributes to research regarding five models accounting for "scale-specific" ERS and MRS. All of the models fit better than standard IRT models when the information criteria are compared. The Multidimensional Nominal Response Model has appeared to fit better than the other models (Multidimensional PCM, mixture PCM, mixture GRM, and Multi-process Model). The MNRM has a much larger number of parameters than the MPCM. Furthermore, the MPCM explained more of the variability in item responses over the standard IRT models than the MNRM. For the scales analyzed here, since classification quality is only satisfactory for the three class models, the three mixture models cannot be recommended for addressing MRS and ERS. Thus, the Multi-dimensional Partial Credit Model can be. Researchers and practitioners are encouraged to consider various factors (e.g., straight-forwardness of implementation, nature of response styles, judgment process, amount of explained variability) when selecting a model.

Only additional research with other scales can provide further insight regarding how these models can be used to account for response styles with a given scale. Further research can also indicate how well extensions to these models can assess general response style tendencies.

APPENDIX A

TWO CLASS CONSTRAINED MIXGRM *MPLUS* CODE

```
INPUT INSTRUCTIONS$
 title: Read Anxiety    GRM  in Mplus$
 data:file = "C:\Users\mLucci\Documents\1A Anxietynew\AnxUnd61.dat";
 variable: names = IDN anx1-anx8 study sex
 age NEON1 zero one two
  three four N1tmrs N1ters;
 missing = all(-99);
 missing = all(6-9);
 auxiliary is IDN;
 usevariables =  Anx1-anx8;! NEON1 'Anxiety'
 categorical = Anx1-anx8 ;
 classes = c(2);
 !COMPUTE NEON1=MEAN.6 N091,N061,N031,N001,N151,N211,N121,N181)*8.
 analysis:
    type = mixture;
    estimator = MLR;
       algorithm = integration;
```

```
   integration = standard(15);
   adaptive = on;
   cholesky = on;
   link = logit;
miterations = 300;
starts = 500 200;
stiterations = 20;
   processors = 4 (starts);
         Model:
 %overall%
      anxiety by anx1* (1)
      anx2(2)
       anx3(3)
       anx4(4)
        anx5 (5)
       anx6 (6)
       anx7 (7)
       anx8(8) ;
       [anxiety@0];
        anxiety@1;
  %c#1%
  anxiety by anx1* (1)
      anx2(2)
       anx3(3)
       anx4(4)
        anx5 (5)
       anx6 (6)
       anx7 (7)
       anx8(8) ;
       [anxiety@0];
```

```
        anxiety@1;
        [anx1$1-anx8$4];
%c#2%
  anxiety by anx1* (1)
      anx2(2)
       anx3(3)
       anx4(4)
        anx5 (5)
       anx6 (6)
       anx7 (7)
       anx8(8) ;
       [anxiety@0];
        anxiety@1;
         [anx1$1-anx8$4];
        output: tech10 ;
```

APPENDIX B

TWO CLASS CONSTRAINED MIXPCM *MPLUS* CODE

```
INPUT INSTRUCTIONS
 title: Read Anxiety
 in Mplus  using version 7.4 :)
  Try constrained mixed PCM
 data:file = "C:\Users\mLucci\Documents\1A Anxietynew\AnxUnd61.dat";
  variable: names = IDN  anx1-anx8 ;
 missing = all(-99);
 missing = all(6-9);
 auxiliary is IDN;
 usevariables = anx1-anx8;! NEON1 'Anxiety'
 categorical = anx1-anx8 (pcm);
  !COMPUTE NEON1=MEAN.6 (NO91,NO61,NO31,NO01,N151,N211,N121,N181)*8.
  !comment to show scale items from NEO-PI-R.
 classes = C(2);
 analysis:
  type = mixture;
    estimator = MLR;
      algorithm = integration;
```

```
   integration = standard(15);
   adaptive = on;
   cholesky = on;
 !  link = logit;
miterations = 500;
starts = 500 100;
stiterations = 20;
   processors = 4 (starts);
       Model:
       %overall%
   anx by anx1* (1)
          anx2 (1)
          anx3 (1)
          anx4 (1)
          anx5 (1)
          anx6 (1)
          anx7 (1)
          anx8 (1);
         [anx@0];
        anx@1;
     %c#1%
      [anx1$1-anx8$4] (t1 - t32) ;
       %c#2%
         [anx1$1-anx8$4] (t33 -t64) ;
  model constraint:
new(sum11 sum12 sum13 sum14 sum15 sum16 sum17 sum18);
sum11 = t1 + t2 +t3 + t4 ;
sum12 = t5 + t6 +t7 +t8 ;
sum13 = t9 + t10 + t11 +t12 ;
sum14 = t13 + t14 + t15 + t16 ;
```

```
sum15 = t17 + t18 +t19 + t20 ;
sum16 = t21 +t22 +t23 +t24 ;
sum17 = t25 +t26 +t27 + t28 ;
sum18 = t29 +t30 +t31 + t32 ;

new(sum21 sum22 sum23 sum24 sum25 sum26 sum27 sum28);
 sum21 = t33 +t34 + t35 +t36 ;
 sum22 = t37 + t38 + t39 + t40 ;
 sum23 = t41 + t42 + t43 +t44 ;
 sum24 = t45 + t46 + t47 + t48 ;
 sum25 = t49 + t50 + t51 + t52 ;
 sum26 = t53 +t54 +t55 +t56 ;
 sum27 = t57 +t58 +t59 + t60 ;
 sum28 = t61 +t62 +t63 + t64;
 ! sum28 = [anx8$1]+[anx8$2]+[anx8$3]+[anx8$4];
 0 = sum11 - sum21 ;
     0 = sum12 - sum22 ;
     0 = sum13 - sum23 ;
     0 = sum14 - sum24 ;
      0 = sum15 - sum25 ;
     0 = sum16 - sum26 ;
     0 = sum17 - sum27 ;
     0 = sum18 - sum28 ;
 output:  ;
```

APPENDIX C

FLEXMIRT CODE FOR MULTI-PROCESS MODEL

```
<Project>
  Title = "Feelins MPM varyL gof M2 calib Score Jun 1";
  Description = "Experience Feelings MPM calibSco ";
<Options>
Mode = Calibration;
  TechOut=yes;
  NumDec = 2;
  SaveCOV = Yes;
  SavePRM = Yes;
  SaveSCO = Yes;
  Score = EAP;
   GOF = Extended;
   M2 = Full;
  SaveDBG = Yes;
NewThreadModel = Yes;
processors = 4;
    <Groups>
  %G1%
   File = "FeelsUnd61BPIs.dat";
  Missing =-99;
```

```
  Varnames = IDN, traitf1-traitf8,  sex ,age,
   neoo3, MBP1-MBP8, DBP1-DBP8 , EBP1-EBP8;
  Select =  MBP1-MBP8, DBP1-DBP8 , EBP1-EBP8;
  N =11407;
Dimensions = 3;
  Ncats(MBP1-MBP8, DBP1-DBP8 , EBP1-EBP8) = 2;
  Model(MBP1-MBP8, DBP1-DBP8 , EBP1-EBP8) = Graded(2);
<Constraints>
Fix (MBP1-MBP8, DBP1-DBP8,EBP1-EBP8),Slope;
Free (MBP1-MBP8),Slope(1);
Free (DBP1-DBP8),Slope(2);
Free Cov(2,1);
Free (EBP1-EBP8),Slope(3);
Free Cov(3,1);
Free Cov(2,3);
```

APPENDIX D

MPLUS CODE FOR MULTI-PROCESS MODEL

```
INPUT INSTRUCTIONS
title:Anxiety new MPM Unconstrained binary pseduo item discriminations
  miss -99;
  data:file =
   "C:\Users\mLucci\Documents\1A Anxiety new\AnxUnd61BPIS.dat";
Variable: names = IDN trait1-trait8 study sex age
 NEON1 zero one two three four MBP1-MBP8 DBP1-DBP8
  EBP1-EBP8 ;
  missing = all(-99);
  auxiliary is IDN;
    usevariables =  MBP1-MBP8 DBP1-DBP8  EBP1-EBP8;
   categorical = MBP1-MBP8 DBP1-DBP8  EBP1-EBP8 ;
      analysis:
     type = general;
     estimator = MLR;
      algorithm = integration;
     integration = standard(15);
     adaptive = on;
     cholesky = on;
     link = logit;
```

```
 miterations = 500;
 !starts = 100;
 processors = 8;
      Model:
 mf1 by MBP1-MBP8*;
 df2 by DBP1-DBP8*;
 ef3 by EBP1-EBP8*;
[mf1@0 df2@0 ef3@0];
 mf1@1 df2@1 ef3@1;
     output:  ;
    plot:  type=plot1 plot2 plot3;
    savedata:  save = fscores;
         file is anxMPMscores.dat;
```

APPENDIX E

MPCM CONSTRAINED SLOPES *FLEXMIRT* CODE

```
Anxiety real MPCM Apr 6
<Project>
  Title = "Anxiety PCM const MRS ERS ";
  Description = "Try Anxiety MPCM model contrain MR ER slopes";
<Options>
  Mode = Calibration;
Processors = 2;
<Groups>
  %G1%
  File = "AnxUnd61.dat";
  Missing =9;
  Varnames = id, trait1-trait8,study, sex,age, NEON1,
zero, one, two, three,four , N1tmrs,N1ters ;
  Select = trait1-trait8;
  N =11407;
Dimensions = 3;
  Ncats(trait1-trait8) = 5;
  Model(trait1-trait8) = GPC(5);
Ta(trait1-trait8)=(
0 0 0 0 0 0 0 0 0 0 0 0,
```

```
1 0 0 0 0 0 0 0 0 0 0 0,
2 0 0 0 0 0 0 0 0 0 0 0,
3 0 0 0 0 0 0 0 0 0 0 0,
4 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 1 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 1 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 1 0 0 0);
<Constraints>
Fix (trait1-trait8),ScoringFN;
Fix (trait1-trait8),Slope;
Free (trait1-trait8),Slope(1);
Equal (trait1-trait8),Slope(1);
Free (trait1-trait8),Slope(2);
Equal (trait1-trait8),Slope(2);
Free Cov(2,1);
Free (trait1-trait8),Slope(3);
Equal (trait1-trait8),Slope(3);
Free Cov(3,1);
Free Cov(2,3);
```

APPENDIX F

MNRM ESTIMATED CATEGORY SLOPES *FLEXMIRT* CODE

```
<Project>
Title ="Experience Feelings F F F MRS ERS real nominal Calib Score JUly 28";
Description="Experience Feelings FFF MRs ERS nominal slope
calibration and score";
<Options>
  Mode = Calibration;
  TechOut=yes;
  NumDec = 2;
  SaveCOV = Yes;
  SavePRM = Yes;
  SaveSCO = Yes;
  SE = SEM;
  Score = EAP;
  SaveDBG = Yes;
processors = 4;
<Groups>
  %G1%
  File = "FeelsUnd61sm.dat";
  Missing =9;
  Varnames = id, traitf1-traitf8,study, sex,age ;
```

```
  Select = traitf1-traitf8;
  N =11407;
Dimensions = 3;
  Ncats(traitf1-traitf8) = 5;
  Model(traitf1-traitf8) = nominal(5);
Ta(traitf1-traitf8)=(
0 0 0 0 0 0 0 0 0 0 0 0,
1 0.7071 1 0.7071  0 0 0 0 0 0 0 0,
2 1 0 -1 0 0 0 0 0 0 0 0,
3 0.7071 -1 0.7071 0 0 0 0 0 0 0 0,
4 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 1 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 1 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 0 0 0 0,
0 0 0 0 0 0 0 0 1 0 0 0);
<Constraints>
Fix (traitf1-traitf8),ScoringFN;
Fix (traitf1-traitf8),Slope;
Free (traitf1-traitf8),ScoringFN(2,3,4);
Free (traitf1-traitf8),Slope(1);
Free (traitf1-traitf8),Slope(2);
Free Cov(2,1);
Free (traitf1-traitf8),Slope(3);
Free Cov(3,1);
```

```
Free Cov(2,3);
```

APPENDIX G

TWO K-MEANS RESPONSE STYLE GROUPS

The different sizes for the analyses with two groups for the chosen scales are presented in Table 33. The Openness to Experience Feelings scale had the largest number of persons in the extreme response group. The Anxiety scale had the smallest number of persons in the non-extreme response group. The sizes of the extreme groups were larger than the non-extreme groups for the Anxiety (N1) scale. The size for the non-extreme groups were larger for the Open to Experience Feelings (O3)and Compliance (A4) scales. Category use for the scales for the two K-means solutions can be seen in appendix H.

Table 33: K-means Cluster Results for Two Different Response Style Groups

Scale	Extreme Size TMRS, TERS M(SD)	Non-extreme Size TMRS, TERS M(SD)
N1 Anxiety	56.0% .08(.08), .20(.22)	44.0% .38(.14), .04(.08)
O3 Open to Experience Feelings	32.1% .04(.07), .59(.19)	67.9% .19(.18), .09(.11)
A4 Compliance	31.3% .09(.10), .39(.15)	68.7% .22(.16), .08(.08)

Note. The percentage of the sample (N = 11,407) assigned to the group in which the Extreme options were preferred to the Midpoint (Extreme size) and where the midpoint was preferred to the Extreme options (Non-extreme size). TMRS = Mean proportion of midpoints used in group, TERS = Mean proportion of extremes used in group, M = Mean of the proportion of midpoints(Extremes) used, SD = Standard deviation.

APPENDIX H

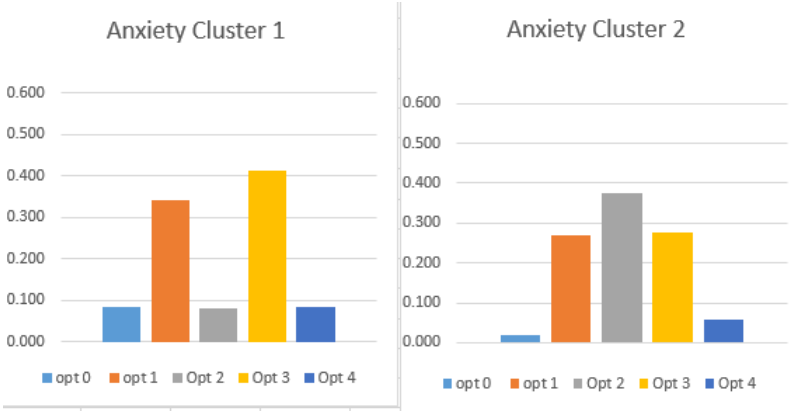
TWO K-MEANS CATEGORY USE

Category Use for the Anxiety Items from the best Two K-means analysis appears in Figure 23. Class 1 uses extreme options more than Class 2. The second class uses the midpoint more than the first class.

Category Use for the Openness to Experience Feelings Items from the best Two K-means analysis appears in Figure 24. Class 2 uses extreme options more than the midpoint. The first class uses the midpoint more than the second class.

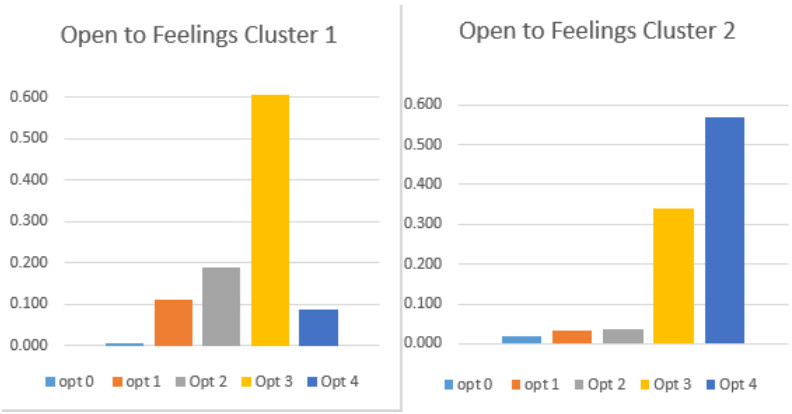
Category Use for the Compliance Items from the best Two K-means analysis appears in Figure 25. Class 2 uses extreme options more than the midpoint. The first class uses the midpoint more than the second class.

Figure 23: Anxiety (N1) Item Category Use for Two K-means solution



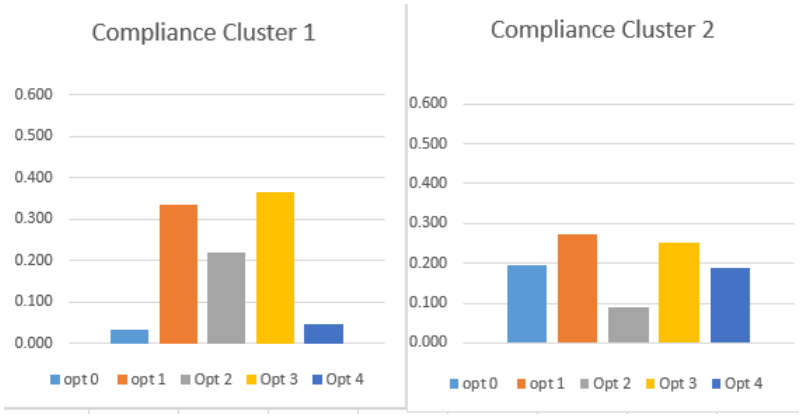
Note: Sample sizes for the classes are $N_1 = 6383$ (extreme) and $N_2 = 5024$ (non-extreme).

Figure 24: Openness to Experience Feelings (O3) Item Category Use for Two K-means solution



Note: Sample sizes for the classes are $N_1 = 7,746$ (non-extreme) and $N_2 = 3,661$ (extreme).

Figure 25: Compliance (A4) Item Category Use for Two K-means solution



Note: Sample sizes for the classes are $N_1 = 7,831$ (non-extreme.) and $N_2 = 3,576$ (extreme).

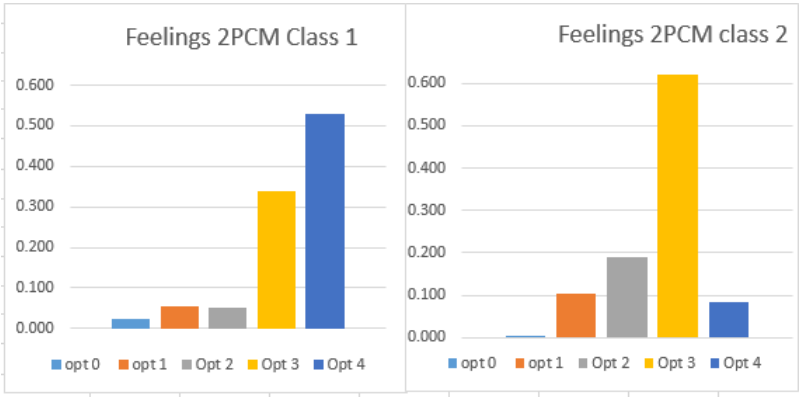
APPENDIX I

TWO CLASS PCM CATEGORY USE

Category Use for the Openness to Experience Feelings Items under the Two class mixture Partial Credit Model for appears in Figure 26. Class 1 uses extreme options more than Class 2. The second class uses non-extreme categories more than extreme categories.

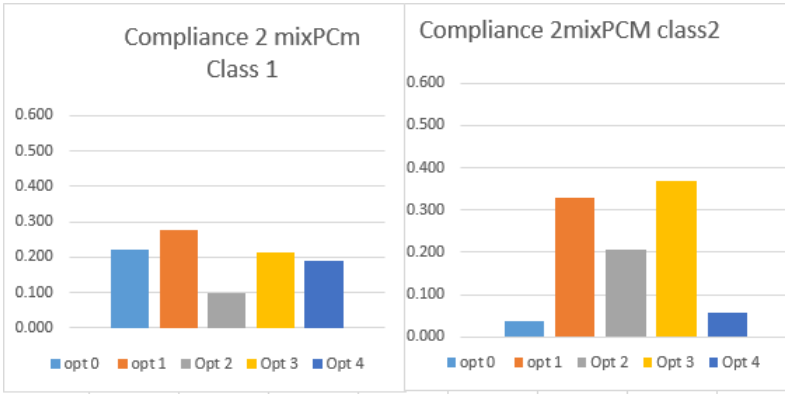
Category Use for the Compliance Items under the Two class mixture Partial Credit Model for appears in Figure 27. Class 1 uses extreme options more than Class 2. The second class uses non-extreme categories more than extreme categories.

Figure 26: Openness to Experience Feelings(O3) Item Category Use for Two class mixture PCM



Note: Sample sizes for the classes are $N_1 = 4015$ (extreme) and $N_2 = 7392$ (non-extreme). This represents 35.2% and 64.8% of the sample.

Figure 27: Compliance (A4) Item Category Use for Two class mixture PCM



Note: Sample sizes for the classes are $N_1 = 3142$ (extreme) and $N_2 = 8265$ (non-extreme). This represents 27.5% and 72.5% of the sample.

APPENDIX J

TRAIT ESTIMATE CORRELATIONS USING TWO CLASS MIXTURE MODELS

Table 34 shows the correlations between response style estimates when the two class mixture models are used instead of the three class mixture models. The correlations which are at least 0.05 lower than the corresponding correlations using three class mixture models are given in bold font.

The correlations between ERS model estimates using two class mixture models are similar in strength to the correlations between ERS estimates using three class mixture models. Only the correlation between the two class mixture PCM and MNRM ERS estimates for the Openness to Experience Feelings scale decreased more than .05. It became .709 which is marked instead of the strong .925.

When the correlations between MRS are examined, for all scales, there are six pairs with a correlation that is less than .05 from corresponding correlation using the three class mixture model. These are the six correlations between MRS trait estimates from a mixture model with a MIRT model.

The correlations for the Anxiety scale are moderate to marked (instead of strong as with the three class mixtures). With the two class mixtures, the correlations for the Openness to Experience Feelings Scale are also moderate to marked instead of marked to strong. Lastly, for the Compliance scale, the correlations are low to marked instead of strong.

This indicates that the two class mixture models are **not** as useful for describing MRS use as the three class mixture models are. The two class mixtures are useful for describing a person as having an extreme or a **non-extreme** response style. With a non-extreme response style, the extreme options are preferred much less than the non-extreme options (agree, disagree, and neither agree nor disagree). The agree and disagree options tended to be preferred over the midpoint in the scales analyzed in this study.

Using the two class mixture models, correlations were also found between the substantive trait estimates. The correlations between model substantive trait estimates for the Anxiety (N1) scale are presented in Table 35. All of the correlations are greater than or equal to .949 when the PCM and GRM estimates are include. For the models which account for Response Styles, the correlations are greater than or equal to .914. These high correlations indicate that all of the models provide strong trait estimates while accounting for the hypothesized response style use.

Using two class mixtures, the correlations between the model substantive trait estimates for the Openness To Experience Feelings (O3) and Compliance (A4) scales are presented in Table 35. All of the correlations are greater than or equal to .815 when the PCM and GRM estimates are include for the Openness To Experience Feelings scale. For the models which account for Response Styles, the correlations are greater than or equal to .815 for the Openness To Experience Feelings estimates. These high correlations indicate that all of the models provide marked to strong trait estimates while accounting for the hypothesized response style use in the Openness To Experience Feelings scale.

The correlations between the trait estimates for the Compliance scale are even stronger. All of them are .934 or larger for the models which account for response styles. There is not much difference in correlations between **substantive** trait estimates in using the two class mixture models instead of three class mixture models.

Table 34: Correlations Between IRT Response Style Estimates using Two Class Mixtures

	Anxiety				
	2mixPC	2mixGR	MPM	MPCM	MNRM
2mixPC	—	.954	.898	.922	.920
2mixGR	.954	—	.855	.910	.908
MPM	**.706**	**.678**	—	.949	.930
MPCM	**.550**	**.565**	.931	—	.980
MNRM	**.570**	**.589**	.940	.992	—
	Openness to Experience Feelings				
	2mixPC	2mixGR	MPM	MPCM	MNRM
2mixPC	—	.907	.887	.924	**.709**
2mixGR	.907	—	.800	.906	.850
MPM	**.715**	**.583**	—	.932	.554
MPCM	**.541**	**.461**	.891	—	.795
MNRM	**.580**	**.497**	.912	.987	—
	Compliance				
	2mixPC	2mixGR	MPM	MPCM	MNRM
2mixPC	—	.980	.923	.895	.913
2mixGR	.980	—	.861	.895	.912
MPM	**.682**	**.664**	—	.950	.932
MPCM	**.375**	**.381**	.841	—	.951
MNRM	**.446**	**.446**	.903	.965	—

Note: Correlations between Model **Midpoint Response trait** estimates are **below** the diagonal. The 'MRS' estimates from the two class mixture models are better described as **non-extreme** response style estimates since persons in the non-extreme class preferred agree or disagree options over the midpoint. Correlations between model **Extreme Response trait** estimates are **above** the diagonal. The correlations which are at least 0.05 lower than the corresponding correlations using three class mixture models are given in bold font. mix2PC = Two class mixture Partial credit model, mix2GR = Two class mixture Graded Response model, MPM = Multi-Process Model, MPCM = Multidimensional Partial Credit Model, MNRM = Multidimensional Nominal Response Model.

Table 35: Correlations between IRT Model Substantive Trait Estimates using Two Class Mixtures

	Anxiety (N1)						
	PCM	GRM	2mixPC	2mixGR	MPM	MPCM	MNRM
PCM	—	.986	.974	.981	.931	.974	.963
GRM		—	.949	.982	.914	.950	.956
2mixPC			—	.983	.971	.995	.982
2mixGR				—	.956	.982	.989
MPM					—	.979	.980
MPCM						—	.990
	Openness to Experience Feelings ╲ Compliance						
	PCM	GRM	2mixPC	2mixGR	MPM	MPCM	MNRM
PCM	—	.975	.988	.973	.933	.987	.975
GRM	.973	—	.948	.979	.926	.965	.975
2mixPC	.881	.815	—	.975	.938	.981	.970
2mixGR	.930	.941	.916	—	.934	.963	.978
MPM	.941	.922	.932	.959	—	.958	.960
MPCM	.910	.853	.978	.932	.960	—	.987
MNRM	.882	.877	.912	.970	.964	.947	—

Note: Correlations between model trait estimates for **Anxiety** are **above** the diagonal. Correlations between Model trait estimates for **Openness to Experience Feelings (O3)** are **below** the diagonal. Correlations between model trait estimates for **Compliance (A4)** are **above** the diagonal. 2mixPC = Two class mixture Partial credit model, 2mixGR = Two class mixture Graded Response model, MPM = Multi-Process Model, MPCM = Multidimensional Partial Credit Model, MNRM = Multidimensional Nominal Response Model.

BIBLIOGRAPHY

Agresti, A. (1996). *Categorical data analysis*. New York, NY: John Wiley and Sons.

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken, NJ: John Wiley and Sons.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of irt models for graded responses. *Applied Psychological Measurement*, *19*(1), 101-119. doi: 0146-6216/95/010101-19\$2.20

Asparouhov, T., & Muthén, B. (2012). *Using mplus tech11 and tech14 to test the number of latent classes* . Retrieved from www.statmodel.com/examples/webnotes/webnote14.pdf

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed rasch modelling of responses to neo-ffi items. *Personality and Individual Differences*, *40(6)*, 1235-1245.

Ayidiya, S. A., & McClendon, M. J. (1990). Response effects in mail surveys. *Public Opinion Quarterly*, *54*(2), 229-247.

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.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*(4), 608-628.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrik*, *37*(1), 29–51.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. doi: 10.1037/a0028111

Böckenholt, U. (2017). Measuring response styles in likert items. *Psychological Methods*, *22*(1), 69–83. doi: doi:http://dx.doi.org/10.1037/met0000106

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multiprocess irt models: A review and tutorial. *British journal of mathematical and statistical psychology*, *70*(1), 159–181. doi: DOI:10.1111/bmsp.12086

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*, 335-352. doi: 10.1177/0146621608329891

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*(5), 814-833.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodal inference: a practical information-theoretic approach*. New York, N. Y.: Springer Science & Business Media.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota multiphasic personality inventory-2: Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*(2), 195-212.

Chamorro-Premuzic, T., & Furnham, A. (2010). *The psychology of personnel selection*. Cambridge, UK: Cambridge University Press.

Chen, C., Lee, S. Y., & 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.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*(2), 187-212.

Cho, Y. (2013). *The mixture distribution polytomous rasch model used to account for response styles on rating scales: A simulation study of parameter recovery and classification accuracy* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Clark, S. L., Muthén, B., J. Kaprio, J., D'Onofrio, B., Viken, R., & Rose, R. J. (2013). Models and strategies for factor mixture analysis: An example concerning the structure underlying psychological disorders. *Structural Equation Modeling: a Multidisciplinary Journal*, *20*(4), 681-703.

Costa, P. T., & McCrae, R. R. (1992). *Professional manual: revised neo personality inventory (neo-pi-r) and neo five-factor inventory (neo-ffi)*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (2014). The neo inventories. In R. P. Archer & S. R. Smith (Eds.), *Personality assessment* (pp. 229–260). New York, NY: Routledge.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*, 3-31.

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*(2), 155-170. doi: 10.1177/014662169401800205

De Beuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. *Quality & Quantity*, *44*(4), 761-775. doi: 10.1007/s11135-009-9225-z

De Jong, M. G., Steenkamp, J. B. E., 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.

DeVillis, R. F. (1991). *Scale development*. Newbury Park, CA: Sage Publications.

Diamantopoulos, A., Reynolds, N. L., & Simintiras, A. C. (2006). The impact of response styles on the stability of cross-national comparisons. *Journal of Business Research*, *59*(8), 925-935. doi: 10.1016/j.jbusres.2006.03.001

Dolnicar, S., & Grün, B. (2007). Cross-cultural differences in survey response patterns. *International Marketing Review*, *24*(2), 127–143. doi: 10.1108/02651330710741785

Egberink, I. J., Meijer, R. R., & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture irt to investigate scalability and predictive validity. *Journal of Research in Personality*, *44*(2), 232–244.

Eid, M., & Raubner, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*(1), 20-30.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Falk, C. F., & Cai, L. (2015). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328–347. doi: 10.1037/met0000059

Ferrando, P. J., Morales-Vives, F., & Lorenzo-Seva, U. (2016). Assessing and controlling acquiescent responding when acquiescence and content are related: A comprehensive factor-analytic approach. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-13. doi: 10.1080/10705511.2016.1185723

Fischer, D. G., & Fick, C. (1993). Measuring social desirability: Short forms of the marlowe-crowne social desirability scale. *Educational and Psychological Measurement*, *53*(2), 417-424.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines*. Boston, MA: Pearson Education, Inc.

Franzblau, A. N. (1958). *A primer of statistics for non-statisticians*. New York, NY: Harcourt, Brace, and World.

Geiser, C. (2013). *Data analysis with mplus*. New York, NY: Guilford Press.

Gnambs, T., & Hanfstingl, B. (2014). A differential item functioning analysis of the german academic self-regulation questionnaire for adolescents. *European Journal of Psychological Assessment*, *30*(4), 251260. doi: 10.1027/1015-5759/a000185

González-Romá, V., & Espejo, B. (2003). Testing the middle response categories "not sure","in between" and "?" in polytomous items. *Psicothema*, *15*(2), 278-284.

Greenleaf, E. (1992a). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, *29*(2), 176-188.

Greenleaf, E. (1992b). Measuring extreme response style. *Public Opinion Quarterly*, *56*(3), 328-352.

Grove, R., Baillie, A., Allison, C., Baron-Cohen, S., & Hoekstra, R. A. (2015). Exploring the quantitative nature of empathy, systemising and autistic traits using factor mixture modelling. *The British Journal of Psychiatry*, *207*(5), 400–406.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, N. J.: John Wiley & Sons.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Newbury Park, CA: Sage Publications, Inc.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin*, *69*(3), 192-203.

Han, K. C. T., & Paek, I. (2014). A review of commercial software packages for multidimensional irt modeling. *Applied Psychological Measurement*, *38*(6), 1-13. doi: 10.117/0146621614536770

Harzing, A. (2006). Response styles in cross-national survey research a 26-country study. *International Journal of Cross Cultural Management*, *6*(2), 243-266.

Harzing, A. W., Brown, M., Köster, K., & Zhao, S. (2012). Response style differences in cross-national research. *Management International Review*, *52*(3), 341-363.

Helmes, E., Holden, R. R., Carstensen, C. H., & Ziegler, M. (2014). Response bias, malingering, and impression management. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 16–46). Boston, MA: Elsevier, Inc.

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(2), 202-226.

Hoffmann, S., Mai, R., & Cristescu, A. (2013). Do culture-dependent response styles distort substantial relationships? *International Business Review*, *22*(5), 814-827.

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Houts, C. R., & Cai, L. (2015). *flexmirt version 3: Flexible multilevel multidimensional item analysis and test scoring*. Seattle, WA: Vector Psychometric Group.

Huang, J. L., Bowling, N. A., Liu, M., & Li, U. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*(2), 299-311.

Hui, C., & Triandis, H. (1989). Effects of culture and response format on extreme response style. *Public Opinion Quarterly*, *49*(2), 253-260.

IBM Inc. (2015). *IBM SPSS Statistics for Windows Version 23* . Armonk, NY: IBM Corp.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, *55*, 243-252.

Jeffries, N. O. (2003). A note on "testing the number of components in a normal mixture.". *Biometrika*, *90*, 991–994.

Jin, K. Y., & Wang, W. C. (2014). Generalized irt models for extreme response style. educational and psychological measurement. *Educational and Psychological Measurement*, *74*(1), 116-138.

Johnson, T. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, *68*(4), 563-583.

Johnson, T., & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, *35*(1), 92-911.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles evidence from 19 countries. *Journal of Cross-cultural psychology*, *36*(2), 264-277.

Jordan, L. A., Marcus, A. C., & Reeder, L. G. (1980). Response styles in telephone and household interviewing: A field experiment. *Public Opinion Quarterly*, *44*(2), 210-222.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*(1), 136-153. doi: 10.1080/10705510701758406

Kelderman, H. (1996). Multidimensional rasch models for partial-credit scoring. *Applied psychological measurement*, *20*(2), 155-168. doi: 10.1177/014662169602000205

Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the big five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, *49*(2), 161–177. doi: 10.1080/00273171.2013.866536

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical cfa and irt. *Structural Equation Modeling*, *18*(2), 212–228. doi: 10.1080/10705511.2011.557337

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213-236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537-567.

Leite, W. L., & Cooper, L. A. (2010). Detecting social desirability bias using factor mixture models. *Multivariate Behavioral Research*, *45*(2), 271-293. doi: 10.1080/00273171003680245

Leventhal, B. (2017). *Extreme response style: Which model is best?* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1-55.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767-778.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Lubke, G., & Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, *43*(4), 592–620.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39.

Magidson, J., & Vermunt, J. K. (2002). Latent class models for clustering: A comparison with k-means. *Canadian Journal of Marketing Research*, *20*(1), 36-43.

Martin, J. (1964). Acquiescencemeasurement and theory. *British Journal of Social and Clinical Psychology*, *3*(3), 216–225.

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

MathWorks. (2016). *Matlab [computer software].* Natick, MA: Mathworks, Inc. Retrieved from https://www.mathworks.com/

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). Defining, measuring, and scaling affective constructs. In *Instrument development in the affective domain* (pp. 48–90). New York, NY: Springer.

McCrae, R. R., & Costa, P. T. (2010). *Professional manual: Neo inventories for the neo personality inventory-3 (new-pi-3), neo five-factor inventory (neo-ffi-3), and neo personality inventory-revised (neo-pi-r).* Odessa, FL: Psychological Assessment Resources.

McCrea, R. L. (2013). *Rethinking the nature of mental disorder: a latent structure to data from three national psychiatric morbidity surveys* (Unpublished doctoral dissertation). University College London, London, UK.

McDonald, R. P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Lawrence Earlbaum.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley & Sons.

Morren, M., Gelissen, J. P. T. M., & Vermunt, J. K. (2011). Dealing with extreme response style in crosscultural research: A restricted latent class factor analysis approach. *Sociological Methodology*, *41*(1), 13-47.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*(1), 73-90.

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., & Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics*, *3*(4), 459-475.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (Seventh ed.). Los Angeles, CA: Muthén and Muthén.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide*. Muthén and Muthén.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263-280.

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Personality and Social Psychology*, *65*(1), 113–131.

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

Ostendorf, F., & Angleitner, A. (2004). *Neo-pi-r: Neo-persnlichkeitsinventar nach costa und mccrae*. Hogrefe.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models (no. 144)*. Thousand Oaks, CA: Sage Publications, Inc.

Paulhus, D. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes: Measures of social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York, NY: Guilford.

Peer, E., & Gamliel, E. (2011). Too reliable to be true? response bias as a potential source of inflation in paper-and-pencil questionnaire reliability. *Practical Assessment, Research & Evaluation*, *16*(9), 1-8.

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, *65*(1), 70-89. doi: 10.1177/0013164404268672

Peterson, R. A., Rhi-Perez, P., & Albaum, G. (2014). A cross-national comparison of extreme response style measures. *International Journal of Market Research*, *56*(1), 89-110.

Peugh, J., & Fan, X. (2013). Modeling unobserved heterogeneity using latent profile analysis: a monte carlo simulation. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4), 616-639.

Plieninger, H. (2016). Mountain or molehill? a simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*(1), 1-19. doi: 0013164416636655

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess irt models for separating content and response styles. *Educational and Psychological Measurement*, *74*(5), 875-899. doi: 0013164413514998

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879.

Pozzebon, J. A., Ashton, M. C., & Visser, B. A. (2014). Major changes personality, ability, and congruence in the prediction of academic outcomes. *Journal of Career Assessment*, *22*(1), 75–88. doi: 10.1177/1069072713487858

Ramaswamy, V., DeSarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with pims data. *Marketing Science*, *12*(1), 103-124.

Reckase, M. (2009). *Multidimensional item response theory (vol. 150).* New York, NY: Springer.

Reynolds, N. L., & Smith, A. (2010). Assessing the impact of response styles on cross-cultural service quality evaluation: a simplified approach to eliminating the problem. *Journal of Service Research*, *13*(2), 230-243. doi: 10.1177/1094670509360408

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A bayesian hierarchical approach. *Journal of the American Statistical Association*, *96*(453), 20-31.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *44*(1), 75-92.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*(1), 75-92.

Rost, J. (1997). Logistic mixture models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449–463). New York, NY: Springer.

Rost, J. (2004). *Lehrbuch testtheorie - testkonstruktion [textbook test theory - test construction]* (2nd ed.). Bern: Huber.

Rost, J., Carstensen, C. H., & von Davier, M. (1997). Applying the mixed rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). Münster, Germany: Waxmann Verlag.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, *16*(2), 155-180. doi: 10.1016/j.hrmr.2006.03.004

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2).

Samejima, F. (1979). Constant information model on the dichotomous response level. In D. J. Weiss (Ed.), *The 1979 computerized adaptive testing conference* (pp. 145–165). Minneapolis, MN: University of Minnesota Press.

Sarstedt, M., & Mooi, E. (2014). Cluster analysis. In *A concise guide to market research* (pp. 273–324). Berlin: Springer-Verlag. Retrieved from 10.1007/978-3-642-53965-7_9

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*(5), 407–424.

Sawatzky, R., Ratner, P. A., Kopec, J. A., & Zumbo, B. D. (2012). Latent variable mixture models: a promising approach for the validation of patient reported outcomes. *Quality of Life Research*, *21*(4), 637–650. doi: 10.1007/s11136-011-9976-6

Schimmack, U., Böckenholt, U., & Reisenzein, R. (2002). Response styles in affect ratings: Making a mountain out of a molehill. *Journal of personality assessment*, *78*(3), 461483.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333-343.

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*, *15*(3), 463–487. doi: 10.1177/1094428112444611

Steinley, D. (2003). Local optima in k-means clustering: What you don't know may hurt you. *Psychological Methods*, *8*(3), 294–304. Retrieved from http://dx.doi.org/10.1037/1082-989X.8.3.294

Steinley, D., & Brusco, M. (2011). Choosing the number of clusters in k-means clustering. *Psychological Methods*, *16*(3), 285–297. Retrieved from http://dx.doi.org/10.1037/a0023346

Sterba, S. K. (2013). Understanding linkages among mixture models. *Multivariate Behavioral Research*, *48*(6), 775–815.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26 psychometrics* (pp. 683–718). New York, NY: Sage Publications, Inc.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393-408.

Thissen, D., & Cai, L. (2016). Nominal categories models. In W. van der Linden (Ed.), *Handbook of item response theory* (Vol. One, pp. 49–73). Boca Raton, FL: Chapman and Hall/CRC Press.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories reponse model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Development and applications* (p. 43-75). New York, NY: Taylor and Francis. doi: 10.4324/9780203861264.ch3

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to likert-type items. *Journal of Educational and Behavioral Statistics*, *38*(5), 522-547.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York, NY: Cambridge University Press.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. 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.

Van Rosmalen, J., Van Herk, H., & Groenen, P. J. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, *47*(1), 157-172.

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. doi: 10.1093/ijpor/edso21

von Davier, M., & Khorramdel, L. (2013). Differentiating response styles and construct-related responses: A new irt approach using bifactor and second-order models. In *New developments in quantitative psychology* (pp. 463–487). New York, NY: Springer.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypot. *Econometrica: Journal of the Econometric Society*, *57*(2), 307-333.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescent and extreme response styles. *Applied Psychological Measurement*, *34*(2), 105-121. doi: 10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods*, *15*(1), 96-110.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*(3), 409-422. doi: 10.1007/s11747-007-0077-6

Wetzel, E. (2013). *Investigating response styles and item homogeneity using item response models* (Unpublished doctoral dissertation). Otto-Friedrich-Universität Bamberg, Bamberg, GE.

Wetzel, E., Böhnke, J. R., & Rose, N. (2016). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement*, *76*(2), 304–324. doi: 10.1177/0013164415591848

Wetzel, E., & Carstensen, C. H. (2015). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*. doi: 10.1027/1015-5759/a000291

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*(2), 178-189.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2015). The stability of extreme response style and acquiescence over eight years. *Assessment*, *23*(3), 1-13. doi: 10.1177/1073191115583714

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods*, *12*(1), 58-79. doi: :10.1037/1082-989X.12.1.58

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and markov chain monte carlo estimation. *British journal of mathematical and statistical psychology*, *26*(3), 339–352.

Wu, P. C., & Huang, T. W. (2010). Person heterogeneity of the bdi-ii-c and its effects on dimensionality and construct validity: using mixture item response models. measurement and evaluation in counseling and development. *Measurement and Evaluation in Counseling and Development*, *43*(3), 155-167.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (pp. 111–153). Westport, CT: American Council on Education and Praeger Publishers.

Zettler, I., Lang, J. W., Hülsheger, U. R., & Hilbig, B. E. (2015). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self- and observer reports. *Journal of personality*. doi: 10.1111/jopy.12172

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, *20*(1), 71-87. doi: 10.1177/014662169602000107